T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

LİTERATÜRDEN BİLGİ ÇIKARIMI; BİR GERÇEK ZAMANLI WEB TABANLI METİN MADENCİLİĞİ UYGULAMASI

Başak OĞUZ YOLCULAR

DOKTORA TEZİ

2016-ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK VE TIBBİ BİLİŞİM ANABİLİM DALI

LİTERATÜRDEN BİLGİ ÇIKARIMI; BİR GERÇEK ZAMANLI WEB TABANLI METİN MADENCİLİĞİ UYGULAMASI

Başak OĞUZ YOLCULAR

DOKTORA TEZİ

DANIŞMAN
Yrd. Doç. Dr. Neşe ZAYİM

2016-ANTALYA

Sağlık Bilimleri Enstitüsü Müdürlüğüne;

Bu çalışma jürimiz tarafından Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı Tıp Bilişimi Programında Doktora tezi olarak kabul edilmiştir. ...../...../.....

İmza

Tez Danışmanı : Yrd. Doç. Dr. Neşe ZAYİM
Akdeniz Üniversitesi

Üye : Prof. Dr. Sadi ÖZDEM
Akdeniz Üniversitesi

Üye : Doç. Dr. Selçuk ÇÖMLEKÇİ
Süleyman Demirel Üniversitesi

Üye : Doç. Dr. Adil ALPKOÇAK
Dokuz Eylül Üniversitesi

Üye : Yrd. Doç. Dr. Uğur BİLGE
Akdeniz Üniversitesi

Bu tez, Enstitü Yönetim Kurulunca belirlenen yukarıdaki jüri üyeleri tarafından uygun görülmüş ve Enstitü Yönetim Kurulu'nun ...../...../..... tarih ve ...../..... sayılı kararıyla kabul edilmiştir.

Prof. Dr. Narin DERİN

Enstitü Müdürü

ETİK BEYAN

Bu tez çalışmasının kendi çalışmam olduğunu, tezin planlanmasından yazımına kadar bütün safhalarda etik dışı davranışımın olmadığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara kaynak gösterdiğimi ve bu kaynakları da kaynaklar listesine aldığımı beyan ederim.

Başak OĞUZ YOLCULAR
İmza

Tez Danışmanı
Yrd. Doç. Dr. Neşe ZAYİM
İmza

TEŞEKKÜR

Bu tezin hazırlanmasında bana rehberlik eden danışmanım Yrd. Doç. Dr. Neşe ZAYİM'e,

Çalışmalarım sırasında bana yol gösteren Anabilim Dalımızdaki değerli hocalarıma, moral ve yardımlarını esirgemeyen mesai arkadaşlarıma,

Tez çalışmam boyunca üzerimden ilgisini eksik etmeyen ve her zaman verdiği fikirlerle yol gösterici olan babam Prof. Dr. Nurettin OĞUZ'a,

Lisansüstü eğitimim sürecinde verdikleri desteklerden dolayı Sağlık Bilimleri Enstitüsü'nün değerli çalışanlarına,

Artık aynı yerde çalışmasak da her zaman her konuda bana destek olan arkadaşım Mehmet Kemal SAMUR'a,

Bana her zaman destek oldukları ve sağladıkları tüm imkanlar için öncelikle eşim Nihat Ozan YOLCULAR, oğlum Güney YOLCULAR'a ve ailemin diğer fertlerine sonsuz teşekkürlerimi sunarım.

ÖZET

Amaç: Bu çalışmanın amacı, Pubmed literatür veri tabanında bulunan makale özetleri kullanılarak, sağlık bakım profesyonellerine hasta bakımında veya klinik araştırmalarda ihtiyaç duydukları bilgiye (kanıta) erişimlerinde ve bilgiyi değerlendirmelerinde yardımcı olacak web tabanlı bir sistem geliştirmektir.

Yöntem: Sistem geliştirme süreci, Pubmed literatür veri tabanından makale özetlerine erişim, metin madenciliği teknikleri kullanılarak metnin ön işlenmesi, medikal varlıkların etiketlenmesi, özetlerden amaç ve istatistiksel terimlerin çıkarımı ve web ara yüzü aracılığı ile gösterimini kapsamaktadır. Özetlere erişim için Biopython Kütüphanesi, medikal varlıkları etiketlemek için Becas Annotator web servisi, istatistiksel terimler için ise NCBO Annotator ve terimleri içeren bir liste kullanılmıştır. Özetlerdeki amaç cümleleri sözlük tabanlı olarak geliştirilen yeni bir algoritma ile çıkartılmaktadır. Etiketlenen varlıklar arasındaki ilişki örüntülerinin bulunması amacıyla birlikte bulunma frekansları hesaplanmaktadır.

Bulgular: Özetler içerisinde etiketlenen varlıklar farklı renklerle vurgulanarak Pubmed benzeri bir ara yüzle kullanıcıya sunulmaktadır. Sistem erişilen makalenin amacını, çalışmada kullanılan istatistiksel terimleri otomatik olarak belirlemekte ve makaleye ait bazı özellikler ve etiketlenen medikal varlıklar ile birlikte tablo biçiminde kullanıcıya sunmaktadır. Farklı sınıflara ait kavramların birlikte bulunma frekansları tablo biçiminde ve grafiksel olarak sunulmaktadır. Amaç çıkarma modülünün kesinlik, hassasiyet ve f-ölçütü değerleri sırasıyla %95, %83,5, %90, istatistiksel terimleri çıkarma modülünün kısmi eşleşme değerlendirme sonuçları %95,4 kesinlik, %88,3 hassasiyet ve %91,7 f-ölçüt, tam eşleşme değerlendirme sonuçları sırasıyla %94,1, %67,8 ve %78,8 şeklindedir.

Sonuç: Sistem Pubmed'te yer alan özetleri analiz ederek medikal bilgiye hızlı erişimi web tabanlı olarak sunmaktadır. Ayrıca literatürdeki diğer sistemlerle karşılaştırıldığında; (i) geniş çaptaki sınıflara ait varlıkları çıkartması (ii) farklı ara yüzlerle kullanıcıya daha hızlı gözden geçirme imkanı sunması ve (iii) ikiden fazla sınıfa ait varlıklar arasındaki ilişki örüntülerini çıkarması ile ayrıcalıklı olduğu görülmektedir.

Anahtar Kelimeler: pubmed, metin madenciliği, literatür madenciliği, bilgi çıkarımı, python

ABSTRACT

Objective: The aim of this study is to develop a web based literature mining system which retrieves Pubmed abstracts to provide tools for information search and evaluation needs of healthcare professionals and researchers in their research and clinical routines.

Method: System development process includes retrieving abstracts from Pubmed literature database, text preprocessing by using text mining techniques, annotating and extracting medical entities, aim sentences and statistical methods of studies, and presenting the results through the web interfaces. In order to retrieve abstracts from Pubmed, a library called BioPython has been used. Becas annotator has been prefered to annotate the medical entities like disease, gene and protein, drug etc. A new algorithm based on dictionary-based method was developed to extract aim sentence of studies. Frequency distribution has been calculated to discover relationship between the tagged entities.

Results: The system tags entities in different color in accordance with their classes and presents the results in a similar interface with Pubmed. It automatically extracts aim of a study and statistical terms used in a study and it demonstrates the results in a different interface with tabular format along with several features of article and the tagged medical entities. Based on the selected entity class by user, co-occurrence frequency of entities are calculated and presented in a table format and visualized with a bar chart. The aim extraction module achieved 83.5% recall, 95% precision and 90% f-measure and statistical term extraction module achieved 95.4% precision, 88.3% recall ve 91.7% f-measure in partial evaluation, 94.1% precision, 67.8% recall and 78.8% f-measure in exact evaluation.

Conclusion: The system provides a web-based platform for mining medical information from Pubmed and it is unique in that it (i) extracts a wide range of entity classes; (ii) allows users to rapid review the results with different interfaces; and (iii) extracts not only binary relation but also relation between more than two entity types with multiple selection choices.

Key words: pubmed, text mining, literature mining, information extraction, python

İÇİNDEKİLER

ÖZET I

ABSTRACT II

İÇİNDEKİLER III

TABLOLAR DİZİNİ V

ŞEKİLLER DİZİNİ VI

SİMGELER VE KISALTMALAR VII

1. GİRİŞ 1

2. GENEL BİLGİLER 3

2.1. Kanıta Dayalı Tıp ve Önemi 3

2.2. Metin Madenciliği 3

2.2.1. Doğal Dil İşleme 4

2.2.2. Bilgi Erişimi 5

2.2.3. Bilgi Çıkarımı 7

2.2.4. Biyomedikal Metin Madenciliği 9

2.3. Literatürden Bilgi Erişimi: Literatür Madenciliği 13

2.3.1. Literatür Madenciliği ile İlgili Geliştirilen Sistemler 14

2.3.2. Var Olan Sistemlerdeki Eksiklikler 16

3. GEREÇ VE YÖNTEM 18

3.1. Sistem tasarımı ve geliştirme süreci 18

3.1.1. Programlama Dili: Python 18

3.1.2. Özetlere Erişim 20

3.1.3. Metin Ön İşleme 22

3.1.4. Biyomedikal Varlıkların ve İstatistiksel Terimlerin Etiketlenmesi 24

3.1.5. Özetlerden Makale Amacının Çıkarılması 29

3.1.6. Web Ara yüzü ve Sonuçların Sunumu 31

3.2. Sistem Performansının Değerlendirilmesi 33

3.2.1. Amaç Çıkarma Modülünün Değerlendirilmesi 33

3.2.2. İstatistiksel Terimleri Çıkarma Modülünün Değerlendirilmesi 34

3.2.3. Değerlendirme Aşamasında Kullanılan Performans Ölçütleri 35

4. BULGULAR 37

4.1. Bilgi Çıkarım Süreci 37

4.1.1. Kullanıcı Sorgularının Sunucuya Gönderilmesi 37

4.1.2. Özetlere Erişim 38
4.1.3. Medikal Varlıkları Etiketleme 38
4.1.4. İstatistiksel Terimleri Çıkarma 41
4.1.5. Makalenin Amacını Belirleme 42
4.1.6. Birlikte Bulunma Frekansları ve Grafiksel Gösterim 43
4.2. Web Ara yüzleri 45
4.3. Sistem Değerlendirme Sonuçları 51
4.3.1. Amaç Çıkartma Modülünün Değerlendirme Sonuçları 51
4.3.2. İstatistiksel Terimleri Çıkartma Modülünün Değerlendirme Sonuçları 52
5. TARTIŞMA 54
KAYNAKLAR 64
EKLER 80
EK-1. Amaçlarda Sık Kullanılan Kelimeler
ÖZGEÇMİŞ 85

TABLOLAR DİZİNİ

Tablo **Sayfa**

2.1. Literatürden bilgi çıkarımı ile ilgili geliştirilen web tabanlı sistemler **15**
3.1. BeCAS içerisinde yer alan varlık tipleri ve veri kaynakları **26**
3.2. Değerlendirme matrisi **35**
4.1. Amaç çıkarma modülü değerlendirme sonuçları **51**
4.2. Amaç çıkarma modülü performans yüzdeleri **51**
4.3. İstatistiksel terimleri çıkarma modülü tam eşleşme değerlendirme sonuçları **52**
4.4. İstatistiksel terimleri çıkarma modülü kısmi eşleşme değerlendirme sonuçları **52**
4.5. İstatistiksel terimleri çıkarma modülü performans yüzdeleri **53**

ŞEKİLLER DİZİNİ

Şekil **Sayfa**

2.1. 1986'dan 2015 yılına kadar Pubmed'de yayınlanan makale sayısındaki artış **14**

3.1. Ardışık düzen **18**

3.2. PubMed özetlerine erişim sağlayan kod bloğu **21**

3.3. POS etiketleme örneği **23**

3.4. Chunking örneği **24**

3.5. BeCAS erişim ve etiketleme fonksiyonu **27**

3.6. NCBO Annotator erişim ve etiketleme fonksiyonu **28**

3.7. Anahtar kelime listesi kullanılarak istatistiksel terim etiketleme fonksiyonu **29**

3.8. Amaç çıkarma modülü **31**

3.9. CGI mimarisi **33**

4.1. Jquery ve Ajax Komutları **37**

4.2. Erişilen özet formatı örneği **39**

4.3. Becas Annotator ile etiketlenen özet örneği **40**

4.4. Vurgulama fonksiyonu **41**

4.5. NCBO Annotator ile elde edilen çıktı örneği **42**

4.6. Varlık listelerini birleştiren kod bloğu **44**

4.7. Frekans dağılımlarını hesaplayan kod bloğu **45**

4.8. Frekans dağılım grafiğini oluşturan kod bloğu **45**

4.9. Giriş ara yüzü **46**

4.10. Pubmed benzeri sonuç ara yüzü **47**

4.11. Tablo formatında sonuç ara yüzü **49**

4.12. Frekans tabanlı ilişki örüntüleri **50**

SİMGELER VE KISALTMALAR

KDT	**:**	Kanıta Dayalı Tıp
DDİ	**:**	Doğal Dil İşleme
İVT	**:**	İsimlendirilmiş Varlık Tanıma
BÇ	**:**	Bilgi Çıkarımı
BE	**:**	Bilgi Erişimi
UMLS	**:**	Unified Medical Language System
NLM	**:**	National Library of Medicine
GO	**:**	Gene Ontology
NLTK	**:**	Natural Language Toolkit
POS	**:**	Part-of-speech
FDA	**:**	Food and Drug Administration
BECAS	**:**	The Biomedical Concept Annotation System
NCBO	**:**	The National Center for Biomedical Ontology
URL	**:**	Uniform Resource Locator
CGI	**:**	Common Gateway Interface

1. GİRİŞ

Sağlık bakımı sadece bir hizmet değil aynı zamanda bir hayat kurtarma mekanizmasıdır. Bu mekanizmada verilen klinik hizmetin kalitesini arttırmak sağlık kurum veya kuruluşlarının en önemli görevlerinden biridir (Hung ve ark., 2012). Sağlık bakım kurumları bu görevi yerine getirmek üzere bilgi sistemleri, karar destek sistemleri gibi bilgi teknolojilerinden faydalanmaktadır. Kanıta dayalı tıp (KDT) da sağlık bakım kalitesinin korunmasında ya da iyileştirilmesinde hayati önem taşıyan bir süreçtir (Drolet ve Lorenzi, 2012). Rosenberg ve Donald (1995) KDT'yi, "klinik karar vermenin temeli olarak güncel araştırma bulgularının sistematik olarak bulunması, değerlendirilmesi ve kullanılması süreci" olarak tanımlamışlardır. Aynı zamanda KDT, klinik uzman bilgisi ile mevcut en iyi kanıtların entegre edilmesi sürecidir (Sackett ve ark., 2007). KDT'nin en önemli adımlarından biri ise en iyi kanıtı bulmaktır. İnternetin gelişmesi ile birlikte çoğu sağlık bakım uzmanı güncel kanıtlara erişmek için geleneksel kanıt toplama yöntemleri (kitap, dergi, meslektaş vb.) yerine çevrim içi medikal veri tabanlarını veya arama motorlarını kullanmaktadır. Böylelikle karar vericiler, güncel bilgilere yer ve zamandan bağımsız olarak hızlı ve etkin bir şekilde erişebilmektedirler.

Günümüzde metin arama motorları, yürütülen araştırma faaliyetlerinde araştırmacılara yardımcı olan önemli bir araç haline gelmiştir. Sağlık alanında, MEDLINE gibi veri tabanları, bilgi keşfi için kullanılabilecek yüksek sayıda metin koleksiyonları sağlamaktadır. Fakat geniş veri koleksiyonlarından istenilen bilgiye erişim hem iş gücü hem de aşırı zaman kaybına yol açmaktadır. Bu sebeple, internetteki geniş veri kaynaklarının işlenmesi için gerekli araç ve tekniklere olan ihtiyaç giderek artmaktadır (Petric ve ark., 2009). Örneğin bir hastalık için ilaç bulmaya çalışan bir araştırmacının, kendisinden önce yapılmış tüm çalışmaları olabildiğince hızlı bir şekilde incelemesi ve bu inceleme sürecinde belgelerin içeriğine, konusuna, içinde geçen kavramlara ve bu kavramların diğer belgelerde geçen farklı kavramlarla ilişkisine ulaşması gerekir (Güven, 2007).

Bu tez çalışmasının amacı, sağlık bakım profesyonellerine hasta bakımında veya klinik araştırmalarda ihtiyaç duydukları bilgiye (kanıta) erişimlerinde ve değerlendirmelerinde yardımcı olacak web tabanlı bir sistem geliştirmek ve

literatürden gerçek zamanlı olarak elde edilen güncel bilgiler ile kanıt temelli hasta bakım ve klinik araştırma sürecine katkı sağlamaktır. Çalışmada, kullanıcı sorgusuna göre Pubmed veri tabanındaki makalelere otomatik olarak erişen, erişilen makalelerin içeriklerini metin madenciliği yöntemleri ile analiz ederek biyomedikal varlıklar arasındaki ilişkileri ortaya çıkaran ve geniş kapsamlı terminoloji kullanımı ile makaleleri en iyi şekilde özetleyecek özelliklerin çıkartılmasını ve kullanıcıya sunulmasını sağlayan bir sistem geliştirilmesi amaçlanmıştır.

Bu çalışmanın beklenilen en önemli katkıları;

1. Sağlık bakım uzmanlarının ve araştırmacıların medikal literatürde istedikleri bilgiye daha hızlı ve kolay bir şekilde ulaşmalarını sağlamak,
2. Web tabanlı altyapı ile daha geniş bir kitleye ulaşmak,
3. Sağlık bakım uzmanlarının güncel bilgilere erişimini kolaylaştırarak kanıta dayalı tıbbi bakımın verilmesinde katkı sağlamak,
4. Kullanıcı sorgusu sonucunda elde edilen makaleleri en iyi şekilde özetleyen özellikleri kullanıcıya sunarak kullanıcıların daha az zamanda ve emekle yüzlerce makaleyi gözden geçirebilmesini sağlamak,
5. Çok çeşitli terminolojiler kullanılarak sınıf bakımından (hastalık, ilaç vb.) daha geniş ve kapsamlı sonuçların elde edilmesini sağlamak,
6. Araştırmalarda kullanılacak medikal veri seti oluşumuna katkı sağlamak,

olarak sıralanabilir.

2. GENEL BİLGİLER

2.1. Kanıta Dayalı Tıp ve Önemi

Kanıta Dayalı Tıp (KDT) hasta bakımı ile ilgili alınan kararlarda mevcut en iyi kanıtların dikkatli, şeffaf ve akılcı kullanımıdır (Sackett ve ark., 1996). Klinik tecrübe, sistematik araştırma ile elde edilen mevcut en iyi kanıtlarla, hasta değer ve beklentilerinin entegrasyonudur. Hasta koşulları ve tercihleri ile mevcut en iyi kanıtların birleşmesi, klinisyen kararlarının kalitesini geliştirmek için uygulanır (Gambrill, 1999). Tarihi temelleri İskoç epidemiyolojist Archibald Cochrane'a dayanır. KDT'ye hizmet veren uluslararası bir organizasyona (Cochrane Collaboration) da ismi verilmiştir. KDT kavramı, 1992 yılında Journal of the American Medical Association da yayınlanan bir makale ile ön plana çıkmıştır (E-BMW Group, 1992).

Yeni bilgilere gereksinim, geleneksel bilgi kaynaklarının yetersizliği, tıp dergilerini okumak için zamanın kısıtlı olması gibi sorunlar ve bilgiye ulaşma araç ve yöntemlerindeki gelişim, toplumun sağlık alanındaki farkındalığının ve bilgiye ulaşımının artması KDT'ye olan ilginin artmasını sağlamıştır. KDT'nin en önemli adımlarından biri en iyi kanıtı bulmaktır. Gelişen bilgi ağlarıyla birlikte araştırmacılar güncel kanıtlara erişmek için geleneksel kanıt toplama yöntemleri (kitap, dergi, meslektaş vb.) yerine çevrim içi medikal veri tabanlarını veya arama motorlarını kullanmaktadır. Böylelikle karar vericiler, güncel bilgilere yer ve zamandan bağımsız olarak hızlı ve etkin bir şekilde erişebilmektedirler. Sağlık alanında özellikle Pubmed araştırmacılar ve klinisyenler tarafından güncel bilgilere veya kanıta ulaşmak için sıklıkla kullanılan çevrimiçi veri tabanıdır.

2.2. Metin Madenciliği

Bilgi teknolojilerindeki yeniliklerle birlikte birçok alanda veriler veri tabanlarında saklanmaya başlanmış ve yapılandırılmış formatta yüksek hacimlerde veriler üretilmiştir. Yapılandırılmış formatta bulunan bu verilerden önceden bilinmeyen, geçerli ve uygulanabilir bilgiyi çıkarmaya yönelik olarak veri madenciliği yöntemleri geliştirilmiştir. Fakat tüm bu gelişmelere rağmen pek çok araştırma alanında ve günlük hayatın içinde üretilen bilgiler ağırlıklı olarak yapılandırılmamış metin dokümanlar şeklinde oluşturulmaya veya saklanmaya devam edilmiştir. Özellikle

internet kullanımının hızla artması ile birlikte giderek artan bu doküman yığınları içinde önemli bilgilere erişmek için farklı yöntemlerin geliştirilmesine yönelik ihtiyaç ortaya çıkarmıştır. Metin madenciliği bu ihtiyaçtan dolayı ortaya çıkan, yapılandırılmamış verileri kullanarak içerisindeki bilgileri gün ışığına çıkaran ve özellikle 2000'li yıllardan sonra ilginin giderek arttığı önemli bir alandır (Konchady, 2006). Metin madenciliği, belirli bir formatta olmayan, serbest metin formatındaki veriler içerisinde gizli olan nitelikli bilginin çıkarılması, düzensiz haldeki verinin formatlanması sürecidir. Metin Madenciliği, Metin Veri Madenciliği (İng. Text Data Mining) ve Metin Veri tabanlarından Bilgi Keşfi (İng. Knowledge Discovery from Textual Databases) olarak da adlandırılır (Hotho ve ark., 2005).

Metin madenciliği tekniklerinin tıpta kullanımı son yıllarda büyük oranda artmıştır. Tıptaki verilerin genel olarak serbest metin formatında bulunması, hasta ile ilgili önemli bilgilerin gözden kaçmasına, bilgiye erişimin zorlaşmasına sebep olmaktadır. Yapılan klinik çalışmalar, araştırma raporları, hastane kayıtları, doktor notları, hasta formları ve faturalar tıptaki en önemli veri kaynaklarıdır. Bu verilerin çoğu serbest metin formatında bulunmaktadır (Konchady, 2006). Özellikle elektronik sağlık kayıtları, sağlık bilgi yönetiminin son yıllarda en önemli hedeflerinden birisiyken, böyle bir sistemin başarısının, klinik dokümantasyonun serbest metin formatında yapılmasından dolayı sınırlanmış durumda olması bu tür yöntemlere olan ihtiyacı ortaya çıkarmıştır.

Bu alanda yapılan çalışmalar incelendiğinde metin madenciliğinin, doğal dil işleme (İng. natural language processing, DDİ), bilgi çıkarımı (İng. Information Extraction, BÇ), isimlendirilmiş varlık tanıma (İng. Named Entitiy Recognition, İVT) ve bilgi erişimi (İng. Information Retrieval, BE) çalışmaları ile çoğu zaman iç içe kullanıldığı görülmüştür.

2.2.1. Doğal Dil İşleme

Dil yeteneği, insan beyninin nasıl çalıştığına ışık tutan insan türüne özgü tek özellik olduğu için dilbilim, bilişsel bilimlerde önemli bir yer tutar. Dilin bilgisayar ortamında modeli oluşturulabilirse iletişim için oldukça yararlı bir araç elde edilmiş olur. DDİ, ana işlevi bir doğal dili çözümleme, anlama, yorumlama ve yeniden üretme olan bilgisayar sistemlerinin tasarımını ve gerçekleştirilmesini konu alan bir mühendislik alanıdır. DDİ, yapay zeka (bilgi gösterimi, planlama, akıl yürütme, vb.),

biçimsel diller kuramı (dil çözümleme), kuramsal dilbilim ve bilgisayar destekli dilbilim, bilişsel psikoloji gibi çok değişik alanlarda geliştirilmiş kuram, yöntem ve teknolojileri bir araya getirir (Erhardt ve ark., 2006). 1950 ve 1960'larda yapay zekanın küçük bir alt alanı olarak görülen bu konu, araştırmacıların ve gerçekleştirilen uygulamaların elde ettiği başarılar sonunda artık bilgisayar bilimlerinin temel bir disiplini olarak kabul edilmektedir. Örneğin çoğumuzun kullandığı kelime işlemcilerde bulunan hatalı yazılmış sözcüğün bulunması ve düzeltilmesi özelliği bu tip uygulamaların en basitlerinden biridir (Oğuz, 2009).

DDİ çalışmalarında çözümleme yapabilmek için sözdizimsel (sentaktik) ve anlambilimsel (semantik) olmak üzere iki yaklaşım ortaya çıkmıştır. Sözdizimsel analiz çalışmalarında genel olarak cümlenin yapısını anlamaya ve cümlede bulunan öğelerin (özne, yüklem, nesne vb.) belirlenmesine yönelik algoritmalar geliştirilmiştir. Anlambilimsel analiz çalışmalarında ise sözdizimini oluşturan morfolojik ögelerin ayrılması yani, sözdizimsel analiz ile anlam taşıyan kelimelerin sınıflandırılması işleminden sonra gelen anlamlandırma veya anlama sürecine odaklanılmıştır. Bu süreçte anlam taşıyan kelimelerin, ekler ve cümle hiyerarşisi içindeki konumlarının saptanması sayesinde birbirleri ile ilişkileri kurulabilir. Bu ilişkiler anlam çıkarma, fikir yürütme gibi ileri seviye bilişsel fonksiyonların oluşturulmasında ham bilgi olarak kullanılabilmektedir.

Günümüzde tıp literatüründe DDİ ile ilgili olarak yapılan birçok çalışma bulunmaktadır. PubMed'de "natural language processing" anahtar kelimeleri ile arama yapıldığında 3514 adet sonuç olduğu görülmüştür. Sosyal medya verilerinin analizi (Myslin ve ark., 2013; Alvaro ve ark., 2015; Sarker ve ark., 2016), influenza ile ilgili Google arama sonuçlarının sınıflandırılması (Maki ve ark., 2015), klinik raporlardaki biyomedikal varlıkların ilişkilerinin belirlenmesi ve sınıflandırılması (Alicante ve ark., 2016; Doan ve ark., 2016; Yadav ve ark., 2016), konu başlıklarının otomatik olarak analiz edilmesi (Cui ve ark., 2011; Lu ve ark., 2013; Han ve ark., 2015) vb. gibi birçok konuda DDİ metotları kullanılmaktadır.

2.2.2. Bilgi Erişimi

İnternet, çeşitli konularda metin, ses, video ve diğer tip dokümanlara erişebileceğimiz bir bilgi ağıdır ve bu ağ her geçen gün katlanarak genişlemekte ve doküman sayısı gitgide artmaktadır. Bireyler kendi bilgi ihtiyaçları doğrultusunda

ilgili bilgiye erişmeye ihtiyaç duyarlar. Bilgi erişimi (BE), kişinin istediği bilgiye ulaşmak için bilgiyi toplama, işleme ve analiz etme adımlarını içeren süreç olarak tanımlanabilir (Chowdhury, 2010). Bu amaç için geliştirilmiş sistemlere de Bilgi Erişim Sistemleri denir. Bir BE sisteminin temel bileşenlerinin aşağıda belirtilen üç ana parçadan oluştuğu söylenebilir;

1. Bir doküman koleksiyonu ya da bu dokümanları temsil eden dizin terimlerini içeren tutanaklar
2. Kullanıcı sorguları
3. Kullanıcıların sorgularında yer alan terimler ile doküman koleksiyonunda yer alan dokümanlara atanan terimleri karşılaştırıp ilgili çakışan dokümanları sunan erişim kuralları (Onur, 2007)

Günlük hayatımızda Google ve benzeri arama motorları herhangi bir konu hakkında bilgi edinmek istediğimizde sıklıkla başvurduğumuz web tabanlı BE sistemleridir. Genel olarak BE sistemlerinin işlevi doküman yığınındaki ilgili dokümanın tümüne erişmek ve ilgili olmayanları elemektir.

Tıp literatürü incelendiğinde çeşitli alanlarda farklı metin veya multimedya koleksiyonları için birçok BE sistemi geliştirildiği görülmüştür. Wu ve arkadaşları (2015) hastalıkla ilişkili makalelere Pubmed'den erişmek için bir sistem geliştirmişlerdir. Bui ve arkadaşları (2015) ise sorgu genişletme ve makaleleri derecelendiren yeni bir metotla Pubmed sonuçlarının iyileştirilmesini sağlamışlardır. Başka bir çalışmada ise farklı alanlardaki sağlıkla ilgili çeşitli web sitelerine erişim sağlanması için içerik tabanlı arama algoritması geliştirilmiştir (Merabti ve ark., 2015). Metin koleksiyonlarının yanı sıra resim, ses ve video formatındaki bilgilere erişim için de önemli çalışmalar yapılmıştır. Zhao ve arkadaşları (2015) biyomedikal alanda 25000 Youtube videosunu toplamış ve bu videoları işleyip dizinleyerek yeni bir ara yüzle bu videolara erişim sağlamışlardır. Giannakopoulos (2015), ses dosyalarından özellik çıkarma, ses sinyallerinin sınıflandırılması ve içerik analizi ve görselleştirme işlemlerinin yapılabileceği bir Python kütüphanesi geliştirmiştir. Bellafqira ve arkadaşları (2015) doktorların veri tabanından ilgili görüntüye daha hızlı erişebilmelerini sağlamak için içerik tabanlı görüntü erişimi sistemi geliştirmişlerdir. Her ne kadar sistemlerin başarı oranları yeni yöntemler

geliştirilerek arttırılsa da henüz yüzde yüz doğruluğu sağlayan sistemler tasarlanamamıştır.

2.2.3. Bilgi Çıkarımı

Bilgi Çıkarımı (BÇ), doğal dille yazılmış metinler içerisinden önceden belirlenmiş sınıflara göre olayların, varlıkların ya da varlıklar arası ilişkilerin ve bunlara ait ilişkili özelliklerin çıkartılması sürecidir (Cowie ve Lehnert, 1996). Bu sistemler doğal dille yazılmış serbest formatta bulunan dokümanlardaki belirli veri parçalarıyla ilgilenmektedir. Yani yapılandırılmamış dokümanlardan yapılandırılmış bilgiyi çıkarmaya çalışmaktadırlar.

Sistemler genellikle yapılandırılmamış metinleri veri tabanı tablosuna aktarılabilecek bir formata dönüştürmektedirler (Cowie ve Lehnert, 1996; Feldman ve ark., 2002; Feldman ve ark., 2008). Metinlerdeki kişi, yer veya organizasyon isimleri gibi faydalı bilgiler metinleri derin bir şekilde anlamaya çalışmadan çıkartılmaktadır (Konchady, 2006). BÇ sistemleri isimlendirilmiş varlık tanıma (İVT), referans çözümleme, ilişki çıkarma ve olay çıkarma olmak üzere dört kategoride incelenebilir (Jakub Piskorski, 2013).

İVT yöntemi, BÇ'nin önkoşulu olarak görülmektedir (Erhardt ve ark., 2006). İVT, önceden belirlenmiş kategorilere göre (kişi isimleri, organizasyonlar, yerler vb.) metin içerisindeki kelimeleri bulmayı ve sınıflandırmayı amaçlayan sistemlerdir (Erhardt ve ark., 2006). İVT sistemlerinin geliştirilmesinde önceden tanımlanmış kavramlar, varlıklar, ontolojiler ve terminolojiler önemli rol almaktadır. UMLS (Unified Medical Language System) yaygın olarak kullanılan, varlıklar arasındaki hiyerarşik ve anlamsal ilişkilerin tanımlandığı en önemli kavram dizinlerinden biridir. NLM (National Library of Medicine) tarafından geliştirilmektedir. Temel klinik kodlama ve referans sistemlerinin terminoloji, semantik ve formatları arasında bağlantılar kuran bir "metathesaurus" içeren bir sistemdir. Uzman bir "lexicon", bir "semantic" ağ ve bir enformasyon kaynakları haritalaması içeren UMLS, atmıştan fazla biyomedikal sözlüğü entegre ederek 900.000 kavrama ait iki milyon ismi dizinlemektedir (Bodenreider, 2003). Genetik alanında ise genler ve proteinler arasındaki ilişkileri belirleyen GO (Gene Ontology) (Smith ve ark., 2003) yazılım geliştirmede sağladığı faydalarla yapılan çalışmalarda sıklıkla kullanılmaktadır. Biyomedikal alanda İVT ile ilgili yapılan çalışmalarda serbest metinlerdeki gen ve

protein isimlerini otomatik olarak tanımaya odaklanılmıştır (Cohen ve Hersh, 2005). Hanisch çalışmasında gen ve protein isimlerinin yer aldığı ve kelimelerin anlamsal olarak sınıflandırıldığı geniş bir sözlük kullanmıştır (Cohen ve Hersh, 2005). He ve arkadaşları (2009) PubMed'deki makale özetlerinden insan proteinleri arasındaki ilişkileri, birlikte bulunma durumlarına ve etkileşimde olduğu kelimelere bakarak çıkartan, web tabanlı bir araç geliştirmişlerdir. Biyolojik varlıkların tüm tiplerinin tam olarak belirtildiği bir sözlük bulunmaması, isimlerinin çok kelimeli olabilmesi, aynı varlığın birden fazla isim alabilmesi vb. yaşanan problemler arasındadır (Cohen ve Hersh, 2005).

Referans çözümleme, aynı nesne veya bulgunun bir metinde farklı cümlelerde tekrar refere edilme durumlarında nesnelerin veya bulguların birbirinden farklı mı aynı mı olma durumunun tespit edilmesini sağlayan bir yöntemdir. Örneğin; "Ali 2 haftadır kol ağrısı çekmektedir. Doktoru, rahatsızlığının sebebini öğrenmek için röntgen çektirmesini önermiştir.". Referans çözümlemede "belirtilen (İng. antecedent)" ve "belirten (İng. anaphor)" olmak üzere iki öğe bulunmaktadır (Zheng ve ark., 2011). Birinci cümledeki kol ağrısı belirtilen ve rahatsızlığı da belirtendir. İkinci cümledeki "rahatsızlığı" kelimesi ilk cümledeki "kol ağrısı" varlığını refere etmektedir. Referans çözümleme bu tarz durumlarda belirtilen ve belirten varlıkların tespit edilmesini sağlayan ve son yıllarda birçok çalışmada başvurulan bir yöntem olmuştur (Crowley ve ark., 2005; Dai ve ark., 2011; Chen ve ark., 2013; Chowdhury ve Zweigenbaum, 2013; Dubey ve ark., 2013; Griffon ve ark., 2014; Lavergne ve ark., 2015; Spasic ve ark., 2015).

İlişki ve olay çıkarma, metin içerisindeki varlıklar arasındaki ilişkilerin ve olayların belirlenmesi ve sınıflandırılması sürecidir. Örneğin; "MAJEZİK çeşitli ağrıları ortadan kaldırmak ya da hafifletmek amacı ile kullanılan steroid olmayan antiinflamatuvar ilaçlar (NSAİ) olarak bilinen grupta yer alır." cümlesinden, "majezik-amacı_ile_kullanılan-çeşitli ağrıları ortadan kaldırmak ya da hafifletmek" veya "majezik-yer_alan-steroid olmayan antiinflamatuvar ilaçlar (NSAİ)" gibi iki ilişki çıkartılabilir. Çalışmalara bakıldığında protein-protein (Huang ve ark., 2006; Miwa ve ark., 2009; Krallinger ve ark., 2011; Liu ve ark., 2016a), hastalık-gen (Dai ve ark., 2013; Guo ve ark., 2016; Zhao ve ark., 2016), hastalık-tedavi (B. Rosario, 2002; O. Frunza, 2010), ilaç-ilaç etkileşimi (Segura-Bedmar ve ark., 2014; Segura-Bedmar ve ark., 2015) vb. gibi ikili (İng. binary) ilişkiler üzerinde odaklanıldığı

görülmektedir. Olay ve ilişki çıkarma, BÇ'nin en zor alanlarından biri olarak görülmektedir.

2.2.4. Biyomedikal Metin Madenciliği

Biyomedikal alanda metin madenciliği ile ilgili yapılan çalışmalar incelendiğinde isimlendirilmiş varlık tanıma, metin sınıflandırma, hipotez üretme, eş anlamlı kelimeleri veya kısaltmaları çıkarma ve ilişki çıkarma olmak üzere çalışmaların beş kategoride toplandığı görülmüştür (Cohen ve Hersh, 2005). Bu bölümde bu konu başlıklarında ile ilgili yapılan çalışmalara, geliştirilen sistemlere ve kullanılan metotlara yer verilecektir.

İsimlendirmiş Varlık Tanıma

Literatürde İVT ile ilgili yapılan çalışmalara bakıldığında geliştirilen sistemlerin sözlük tabanlı, kural tabanlı, istatistik tabanlı (denetimli-denetimsiz öğrenme yöntemleri) veya bu yöntemlerin birleşimi ile elde edilen metotları içeren yapıda olduğu görülmektedir. İVT çalışmaları genetik (gen, protein ve ilişkili biyolojik ve genetik terimleri bulma) ve medikal (hastalık, ilaç isimleri ve diğer medikal terimleri bulma) olmak üzere iki alanda incelenebilir.

İVT sistemleri başlangıçta kural tabanlı ve sözlük tabanlı olarak geliştirilmekteydi (Friedman ve ark., 1994; Fukuda ve ark., 1998; Rindflesch ve ark., 2000; Tanabe ve Wilbur, 2002). MedLEE, sözlük tabanlı olarak geliştirilen sistemlere örnek verilebilir (Friedman ve ark., 1994). Bu sistem, kontrollü sözlük kullanarak klinik metinlerdeki terimleri etiketleyen bir doğal dil işleyicisidir. Sistemin performansını değerlendirmek için 230 radyoloji raporu kullanılmış ve hassasiyet ve kesinlik ölçütleri sırasıyla %70 ve %87 olarak bulunmuştur. Benzer olarak EDGAR isimli sistem de Medline literatür veri tabanındaki kanser ile ilgili özetlerde bulunan gen ve ilaçlar ile ilgili bilgiyi UMLS terminolojisini kullanarak çıkarmaktadır (Rindflesch ve ark., 2000). AbGene biyomedikal literatürde bulunan gen ve protein isimlerinin etiketlenmesi için geliştirilmiş en başarılı İVT sistemlerinden biridir (Tanabe ve Wilbur, 2002). Protein ve gen isimlerinin morfolojik, bağlamsal ve gramer yapısı temel alınarak hem kural tabanlı metin işleme yöntemlerini hem de sözlük tabanlı bir yaklaşımı benimsemektedir. Savova ve arkadaşları tarafından geliştirilen cTAKES, elektronik sağlık kayıtlarında bulunan serbest metin formatındaki klinik raporlar için geliştirilmiş açık kaynak kodlu DDİ sistemidir (Savova ve ark., 2010b). cTAKES

kural tabanlı yöntemleri ve sözlük tabanlı metotları birleştirerek bilgi çıkarım sürecini sağlamaktadır.

Son yıllarda etiketlenmiş (annotated) metin koleksiyonlarındaki artış ile birlikte kural tabanlı veya sözlük tabanlı yaklaşımdan denetimli (Hidden Markov Models, Conditional Random Fields) öğrenme yöntemlerine geçiş başlamıştır (Zhang ve Elhadad, 2013). Biyomedikal alanda çalışan doğal dil işleyiciler için geliştirilen GENIA metin koleksiyonu, destek vektör makineleri (Mitsumori ve ark., 2005), Hidden Markov Models (Zhang ve ark., 2004), Conditional Random Fields (McDonald ve Pereira, 2005; He ve Kayaalp, 2008) gibi çeşitli denetimli öğretim yöntemlerinin kullanıldığı sistemlerin geliştirilmesini sağlamıştır.

Metin Sınıflandırma

Metin sınıflandırma, bir dokümanın veya parçasının verilen bir konuyu veya belirli bir bilgiyi içermesi gibi karakteristiklere sahip olup olmadığını belirlemeyi hedeflemektedir (Cohen ve Hersh, 2005). Metin sınıflandırma, ham metinleri önceden belirlenen bir veya birden fazla kategoriye atayan metin madenciliğinin anahtar teknolojilerinden biridir (Dai ve Liu, 2014). Metin sınıflandırma çalışmalarında yaygın olarak makine öğrenmesi teknikleri kullanılmaktadır (Sebastiani, 2002). Son yıllarda, bilgi teknolojilerindeki ve DDİ alanındaki ilerlemelerle birlikte çeşitli sınıflandırma konuları için artan sayıda denetimli sınıflandırma yaklaşımları geliştirilmiştir (Dai ve Liu, 2014). Bu yaklaşımlar Karar ağaçları (Murthy, 1998; De Comité ve ark., 2003), yapay sinir ağları (Ruiz ve Srinivasan, 2002; Yu ve ark., 2008), naive Bayes (Lee ve ark., 2012), destek vektör makineleri (Sun ve ark., 2009; Wang ve Chiang, 2009; Kumar ve Gopal, 2010), ve k en yakın komşu (Jiang ve ark., 2012) gibi yöntemleri içermektedir. Bu sınıflandırıcılar içinde en kapsamlı performansa sahip olanlar destek vektör makineleri, k en yakın komşu ve naive bayes yöntemidir (Su ve ark., 2006). Bui ve Zeng-Treitler (2014) çalışmalarında Regular Expression Discovery (RED) algoritmasını geliştirerek SMOKE veri setinde bulunan sigara içme durumu ile ilgili bilgileri içeren metin kesitlerini ve PAIN veri setini kullanarak acı durumları ile ilgili metin kesitlerini sınıflandırmışlardır. RED sınıflandırıcının her iki veri setinde %80,9-%83,0 doğruluk başarısına ulaştığı belirtilmiştir. Sarker ve Gonzalez çalışmalarında (2015), makine öğrenmesi algoritmalarını kullanarak klinik ve sosyal medya metinlerinden ilaç yan etkileri ile ilgili bilgiyi çıkartmış, bu bilgiye göre

metinleri otomatik olarak sınıflandırmış ve sistemin f-ölçüt değerini %81,2 olarak hesaplamışlardır. Başka bir çalışmada Laplacian destek vektör makinesi algoritması kullanılarak 820 karın CT, MRI ve ultrason raporları kanserli karaciğer lezyonlarına göre sınıflandırılmış ve sistemin makro-F1 skoru %74,1 olarak bulunmuştur (Garla ve ark., 2013).

Hipotez üretme

Hipotez üretme, daha önceden bilinmeyen ilişkileri çıkarmayı hedeflemektedir. Biyomedikal varlıklar arasındaki daha önceden bilinmeyen ilişkilerin çıkartılması ilk olarak Swanson tarafından sunulmuştur (Swanson, 1986, 1990). Bilimsel literatürdeki bulgularla bağlantılı olarak hipotezlerin keşfi için ABC modelini tasarlamıştır. ABC modelinde birbirinden farklı iki kayıt seti (A ve C) arasındaki bilinmeyen ilişkiler, A ve C ile sıkça rastlanan ortak bir B öğesi kullanılarak tespit edilmeye çalışmaktadır. Swanson'a göre AB ilişkisi ve BC ilişkisi önceden bilimsel literatürde ayrı ayrı yayınlanmış, fakat birlikte düşünülmemiştir (Petric ve ark., 2009). Swanson, modelini manuel olarak uygularken, günümüzde bu süreci otomatikleştiren yöntemler geliştirilmeye başlanmıştır (Cohen ve Hersh, 2005). Weeber ve arkadaşların Swanson'ın modelini otomatikleştirerek Pubmed literatür veri tabanı için bir kavram tabanlı DDİ sistemi olan DAD sistemini geliştirmişlerdir (Weeber ve ark., 2000). Raynaud's hastalığı ile ilgili Pubmed'de yer alan 385 literatür gözden geçirme makalesini kullanarak sistemi tasarlamışlardır. Srinivasan ve Libbus (2004) Open Discovery algoritmasını kullanarak bir besin maddesi olan zerdeçalın ilişkili olduğu kavramları araştırmışlar ve özellikle retina ile ilgili hastalıklarda, Crohn hastalığında ve omurilik ile ilgili rahatsızlıklarda fayda sağladığını tespit etmişlerdir. Bunun yanı sıra zerdeçalın çeşitli genlerle de ilişkisi olduğunu göstermişlerdir. Weeber'in (2007) yaptığı başka bir çalışmada ise literatürden UMLS terimlerini kullanarak yeni hipotezler üreten Literaby isimli bir sistem geliştirilmiştir. Lindsay ve Gordan (1999) kelime sayısı gibi sözcükler ile ilgili istatistikleri kullanarak literatürdeki gizli ilişkileri keşfetmeyi hedeflemişlerdir.

İlişki çıkarma

İlişki çıkarma literatürde en sık rastlanılan metin madenciliği alanlarından biridir. Önceden belirlenen ilişki tiplerine göre varlıklar arasındaki ilişki örüntüleri tespit edilmeye çalışılmaktadır. Şu anda ilişki çıkarma çalışmalarında genel olarak denetimli ilişki çıkarma, yarı denetimli ilişki çıkarma ve denetimsiz ilişki çıkarma

olmak üzere üç temel yöntem kullanılmaktadır (Li ve ark., 2015). Yarı denetimli ilişki çıkarma yöntemi ile yapılan çalışmalarda (Rozenfeld ve Feldman, 2008; Nguyen ve ark., 2015; Zhang ve ark., 2015; Zhou ve Zhong, 2015; Liu ve ark., 2016b) ilişki kategorileri etiketlenmiş metinleri içeren koleksiyonlar kullanılarak besleme yapılmakta ve ilişki örüntüleri öğrenilmektedir. Daha sonra bu örüntüler daha önceden etiketlenmemiş metin koleksiyonlarına uygulanarak yeni setler elde edilmektedir. Örüntüleri elde etmeye devam etmek için üretilen yeni setler besleme setine eklenmekte ve böylelikle ilişki çıkarma etiketlenmemiş koleksiyonlar üzerinde devam etmektedir. He ve arkadaşları (2006) aynı cümle içinde geçen isimlendirilmiş varlıkları bularak birbirlerine olan uzaklıklarını hesaplamış, vektörlere dönüştürmüşler ve varlık örüntülerinden bazılarını seçerek ilk aşamada ilişki besleme seti olarak kullanmışlardır. Cui ve arkadaşları (2009) proteinler arasındaki ilişki örüntülerini çıkarmak için destek vektör makinelerini ve aktif öğrenme stratejilerini kullanmışlardır. Denetimli ilişki çıkarma yöntemlerinin kullanıldığı çalışmalarda (Rink ve ark., 2011; Zheng ve Blake, 2015) ise önceden ilişki formatları etiketlenmiş metin setleri üzerinde yöntemler uygulanmaktadır. Roberts ve arkadaşları (2008) hasta raporlarından varlıklar arasındaki ilişkileri çıkarmak için destek vektör makinelerini kullanmışlar ve %72 F1 skoru elde etmişlerdir. Nguyen ve arkadaşları (2015) sözdizimsel örüntüleri kullanarak biyomedikal literatürden ikili ilişkileri çıkartan PASMED isimli bir sistem geliştirmişlerdir. Denetimsiz ilişki çıkarmada genel yaklaşım etiketlenmemiş metin koleksiyonundaki benzer ilişki örüntülerini bularak onları kümelemektir (Eichler ve ark., 2008; Mohanty ve ark., 2014). Quan ve arkadaşları (2014) Polinominal Kernel yöntemini kullanarak biyomedikal literatürdeki protein-protein ve gen-intihar arasındaki ilişkileri etiketlenmemiş veri setlerinden çıkarmaya çalışmışlar ve %55 f skoru elde etmişlerdir. Alicante ve arkadaşları (2016) İtalyanca klinik raporlarından standart DDİ araçları ve özellik vektörlerini kullanarak ilişki örüntülerini keşfetmişlerdir.

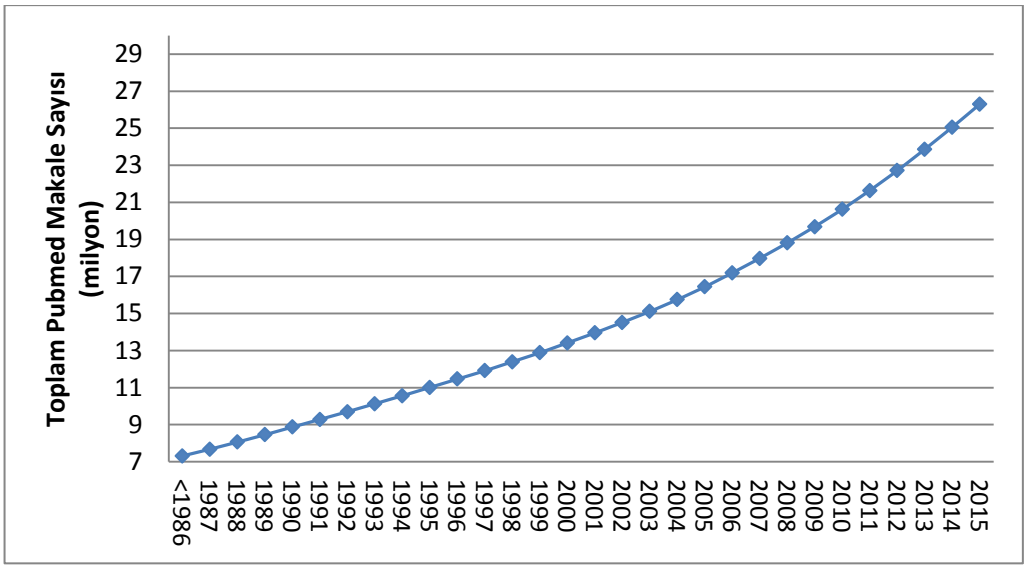
Eş anlamlı sözcükleri ve kısaltmaları çıkarma

Birçok biyomedikal varlık çeşitli isimlere ve kısaltmalara sahiptir. Bu sebeple kısaltmaların ve eş anlamlı sözcüklerin tek bir varlık tanımı ile eşleştirilmesi önem kazanmıştır. Bu alanda yapılan çalışmalarda genel olarak gen isimlerinin eş anlamlıları ve biyomedikal terimlerin kısaltmalarını çıkartma konusuna odaklanılmıştır (Cohen ve Hersh, 2005). Henriksson ve arkadaşları (2014) rastgele

dizinleme ve permütasyon yöntemlerini kullanarak medikal makalelerdeki ve klinik dokümanlardaki kısaltmalar ve eş anlamlı terimler için aday terimler üretmişlerdir. Yu ve arkadaşları (2002) Medline veri tabanındaki makalelerden ve dergi makalelerinden gen ve protein isimlerinin eş anlamlılarını bulan ve aday terimler üreten bir yazılım geliştirmişlerdir. Yazılımın %71 kesinlik ölçütü skoru elde ettiğini belirtmişlerdir. Ao ve Takagi (2005) biyomedikal literatürdeki kısaltmaları çıkartan ALICE isimli bir sistem geliştirmişlerdir. Sistem, sezgisel örüntü eşleştirme kurallarını kullanarak %95 hassasiyet, %97 kesinlik skorlarına ulaşmıştır. Cohen (2004) varlıkların birlikte bulunma ağlarını kullanarak Medline veri tabanındaki özetlerden protein ve gen isimlerinin eş anlamlılarını çıkartmış ve herhangi bir İVT aracı kullanmadan %22 F skoru elde etmiştir.

2.3. Literatürden Bilgi Erişimi: Literatür Madenciliği

Bilimsel literatür biyomedikal kavramlar arasındaki ilişkileri sunan zengin bir veri kaynağıdır (Frijters ve ark., 2010). Son 20 yılda teknolojideki gelişmeler ve araştırma kapasitesindeki hızlı artış ile birlikte sağlık literatüründe katlanarak artan bir büyüme yaşanmıştır. Biyomedikal alanda, Medline gibi veri tabanları, bilgi keşfi için kullanılabilecek yüksek sayıda metin koleksiyonları sağlamaktadır. Şekil 2.1'de 1986'dan 2015 yılına kadar Pubmed'de yayınlanan makale sayılarındaki artış gösterilmektedir. Akademik anlamdaki bu birikim, araştırmacıların bilimsel keşif yapmalarında ve sağlık bakım profesyonellerinin sağlıkla alakalı meselelerin yönetiminde avantaj sağlamıştır. Fakat bu geniş ve hızla büyüyen makale yığınından istenilen bilgiyi çıkarmak giderek zorlaşmıştır. Genel olarak, bilimsel literatür kaynaklarına erişimde metin arama motorları kullanılmakta ve arama sonuçları kullanıcılara web sayfa listeleri veya makale listeleri vb. görünümlerde sunulmaktadır. Fakat var olan haliyle metin formatındaki makale yığınlarından veya web sayfalarından istenilen bilgiye erişmek zor ve zaman kaybına yol açmaktadır. Bu sebeple, internetteki geniş veri kaynaklarının işlenmesi için gerekli araç ve tekniklere olan ihtiyaç giderek artmaktadır. Metin madenciliğindeki ilerlemelerle birlikte araştırmacılara hızlı ve etkili bir şekilde arama ve ilgili makaleye erişmelerinde yardımcı olan web tabanlı sistemler geliştirilmeye başlanmıştır.



Şekil 2.1. 1986'dan 2015 yılına kadar Pubmed'de yayınlanan makale sayısındaki artış

2.3.1. Literatür Madenciliği ile İlgili Geliştirilen Sistemler

Literatürden bilgi çıkarımı ile ilgili çeşitli alanlarda farklı veri tabanları için geliştirilmiş birçok sistem bulunmaktadır. Bu sistemler arama sonuçlarını yeniden sıralayan, konu başlıklarına göre sonuçları kümeleyen, sonuçları anlamsal ve görsel olarak zenginleştiren ve arama ara yüzünü ve erişim deneyimini iyileştiren olmak üzere dört kategoride toplanabilir (Kim ve Rebholz-Schuhmann, 2008).

Literatürden yapılan taramalar sonucu 21 adet web tabanlı sisteme ait özellikler Tablo 2.1'de verilmiştir. Genel olarak sistemlerin Pubmed veya Medline veri tabanları için geliştirildiği görülmektedir. Bu yüzden daha çok isimlerinde "Pub" veya "Med" hecesini içermektedir. Sistemlerin çoğu (RefMed, MedlineRanker, iPubMed vb.) akademik kökenli araştırmacılar tarafından geliştirilse de devlet (EBIMED, askMEDLINE, PICO vb.) veya özel sektör (Quertle, PubFocus, GoPubMed vb.) tarafından geliştirilen sistemler olduğu da görülmektedir. Yıla göre bir analiz yapıldığında geliştirilen sistemlerde 2008-2010 yılları arasında bir artış olduğu gözlenmektedir. Birçok arama motorunda olduğu gibi çoğu sistem sonuçları liste şeklinde kullanıcıya sunmakta, bazıları ise semantik ilişkileri çıkartan ve gösteren grafik veya tablo formatını kullanmaktadır. Sistemlerde tam metin makaleler üzerinde çalışılmaya başlanılsa da zaman ve maliyet faktörleri nedeniyle çoğunlukla makalelerin özetleri üzerinde yoğunlaşılmaktadır. Ayrıca geliştirilen sistemlerde daha çok sonuçların yeniden sıralanması ve ara yüzün iyileştirilmesi yönünde odaklanıldığı görülmektedir.

Tablo 2.1. Literatürden bilgi çıkarımı ile ilgili geliştirilen web tabanlı sistemler

Sistem Adı	Yıl	Veri tabanı	Amaç
MedlineRanker(Fontaine ve ark., 2009)	2009	Medline	Önemli kelimeleri belirlemek, bu kelimeleri kullanarak tekrar arama yapmak ve sonuçları sıralamak
LitInspecter(Frisch ve ark., 2009)	2009	Pubmed	Önemli cümleleri filtrelemek, cümle içerisindeki terimleri farklı renklerle vurgulamak
SciMiner(Hur ve ark., 2009)	2009	Medline	Sonuçları sıralamak, terimleri ve protein-protein ilişkilerini belirlemek
RefMed(Yu ve ark., 2010)	2010	Pubmed	Sonuçları kullanıcıya oylama formu ile sunarak kullanıcının yaptığı oylamaya göre tekrardan arama yapmak
iPubMed(Wang ve ark., 2010)	2010	Medline	Sonuçları sıralamak, sonuçlar içerisindeki arama kelimelerini ve benzer kelimeleri vurgulamak
CoPub(Frijters ve ark., 2008)	2008	Medline	Sonuçları sıralamak, arama terimlerinin farklı kategorilerdeki kelimeler ile ilişkilerini göstermek
FACTA+(Tsuruoka ve ark., 2008)	2008	Medline	Sonuçları sıralamak, terimler arasındaki doğrudan olmayan ilişkileri göstermek,arama terimlerinin farklı kategorilerdeki kelimeler ile ilişkilerini göstermek
EBIMED(Rebholz-Schuhmann ve ark., 2007)	2007	Medline	Sonuçları sıralamak, arama terimlerinin farklı kategorilerdeki kelimeler ile ilişkilerini göstermek
PubFocus(Plikus ve ark., 2006)	2006	Pubmed	Sonuçları sıralamak
eGIFT(Tudor ve ark., 2010)	2010	Pubmed	Sonuçları sıralamak, arama kelimelerinin birlikte bulunduğu kelimeleri göstermek
Quertle(Giglia, 2011)	2011	Pubmed	Sonuçları sıralamak, sonuçlar içerisindeki arama kelimelerini ve benzer kelimeleri vurgulamak, sonuçları filtrelemek
PubAnatomy(Xuan ve ark., 2010)	2010	Medline	Nörolojik yapı ile moleküler veriyi bir arada kullanmak, arama kelimelerinin etkileşimde olduğu hastalık ve genlerle ilişkisini ve beyinde ilgili olduğu bölgeleri beyin haritası üzerinde göstermek
MEDIE(Kim ve ark., 2008)	2008	Pubmed	Özne-yüklem-nesne yapısı kullanılarak semantik bir arama yapmak, sonuçları sıralamak, sonuçlar içerisindeki arama kelimelerini vurgulamak,
PolySearch(Cheng ve ark., 2008)	2008	Pubmed	Sorgu yapılandırılması, iki farklı kategorideki varlıkları içeren sonuçların elde edilmesi
GoPubMed(Doms ve Schroeder, 2005)	2005	Pubmed	Sonuçları sıralamak, arama ile ilgili yıllara, ülkelere, dergilere göre makale sayısı gibi istatistikler vermek
RLIMS-P(Torii ve ark., 2014)	2014	Pubmed	Protein fosforilasyonu ile ilgili dokümanlara erişmek
PubNet(Douglas ve ark., 2005)	2005	Pubmed	Sonuçlar ağ yapısı ile göstermek
askMEDLINE(Fontelo ve ark., 2005)	2005	Medline	Sorgunun genişletilmesi
PubTator(Wei ve ark., 2013)	2013	Pubmed	Farklı kategorilerdeki varlıkların sonuçlarda etiketlenmesi
BMExpert(Wang ve ark., 2015)	2015	Medline	Sonuçları sıralamak, konuyla ilgili uzman kişileri bulmak
PubstractHelper(Chen ve Ho, 2014)	2014	Pubmed	Arama kelimelerinin geçtiği cümleleri etiketlemek

2.3.2. Var Olan Sistemlerdeki Eksiklikler

Daha öncede bahsedildiği gibi literatürden bilgi çıkarımı için geliştirilen sistemlerin daha çok sorgu sonuçlarının yeniden sıralanması (Errami ve ark., 2007; Fontaine ve ark., 2009; States ve ark., 2009; Yu ve ark., 2010) ve arama ara yüzünün iyileştirilmesi (Fontelo ve ark., 2005; Muin ve ark., 2005; Schardt ve ark., 2007) üzerine geliştirildikleri görülmüştür. Diğer çalışma alanları ise biyomedikal varlıklar arasındaki ilişkinin gösterilmesi (Rebholz-Schuhmann ve ark., 2007; Tsuruoka ve ark., 2008; Hur ve ark., 2009; Fleuren ve ark., 2011) veya makalelerin başlıklara göre gruplandırılması (Perez-Iratxeta ve ark., 2002; Doms ve Schroeder, 2005; Smalheiser ve ark., 2008) üzerinedir.

Tüm erişilebilen sistemler incelendiğinde sistemlerdeki üç önemli eksikliğin ortaya çıktığı görülmektedir. Bunlardan ilki, erişim sonuçlarının sadece Pubmed web servisinin sunduğu değişkenlere göre kullanıcıya sunulmasıdır. Xuan ve arkadaşları (2010) tarafından geliştirilen PubAnatomy isimli sistem buna örnek olarak verilebilir. Bu sistem sadece, makale ile ilgili yıl, başlık, yazar ve özet gibi Pubmed web servislerini kullanılarak erişilebilecek özellikleri kullanıcılara sunmaktadır. Bir araştırmacının makaleyi daha iyi anlayabilmesini sağlayan makalenin içeriği ile ilgili özellikler sistem çıktısı olarak kullanıcıya verilmemektedir. Pubmed web servislerinin sağladığı temel özellikler dışında makale içerisinde gizli kalan ve makalenin içeriği ile ilgili daha açıklayıcı bilgi veren özelliklerin sistem çıktısı olarak kullanıcıya sunulmasının makalenin kullanıcı tarafından daha az emekle yorumlanabilmesinde ve medikal araştırmalar için kullanılabilecek güncel bir veri seti oluşumunda katkı sağlayacağı düşünülmektedir.

Var olan sistemlerle ilgili ortaya çıkan diğer bir eksiklik ise aynı veya ayrı kategorilerde (hastalık, ilaç vb.) bulunan iki veya daha fazla varlık arasındaki ilişkiyi bütünsel olarak gösteren bir sistem olmamasıdır. Genel olarak geliştirilen sistemlerde kullanıcı tarafından girilen sorgu terimleri ile ilişkili kavramların, erişilen makalelerde ne kadar sıklıkla birlikte bulunduğu ile ilgili sonuçlar her kategori için ayrı ayrı gösterilmektedir. Fleuren ve arkadaşları (2011) tarafından geliştirilen CoPub 5.0 isimli sistem bahsedilen probleme örnek olarak gösterilebilir. Bu sistem kullanıcı sorgusunu almakta, sorgu terimlerinin geçtiği makaleleri bulmakta ve bu makaleler içerisinde geçen sorgu terimleri ile ilişkili kavramları çok fazla sayıda kategori için ayrı ayrı olarak kullanıcıya sunmaktadır. Sorgu çıktısının değişik

uzunluk ve yapılarda verilmesi, kullanıcının farklı kategorilerde bulunan kavramlar arasındaki ilişkileri bütünsel olarak yorumlayabilmesini ve aradığı bilgiyi kolaylıkla bulmasını zorlaştırmaktadır. Benzer olarak, Tsuruoka ve arkadaşları (Tsuruoka ve ark., 2008) tarafından geliştirilen FACTA+ isimli sistemde de kullanıcı girdiği sorgu terimleri ile ilişki kavramları ayrı ayrı kategorilerde elde etmektedir. Diğer sistemden farklı olarak FACTA+ sistemi sonuçları daha az kategoride özetlemiştir. Kullanıcıların sorgularıyla ilişkili kavramların her kategori için ayrı ayrı analiz edilmesi yerine daha bütünsel bir yöntemle analiz edilerek kullanıcıya sunulmasını sağlayan sistemlerin gerekliliği ön plana çıkmaktadır.

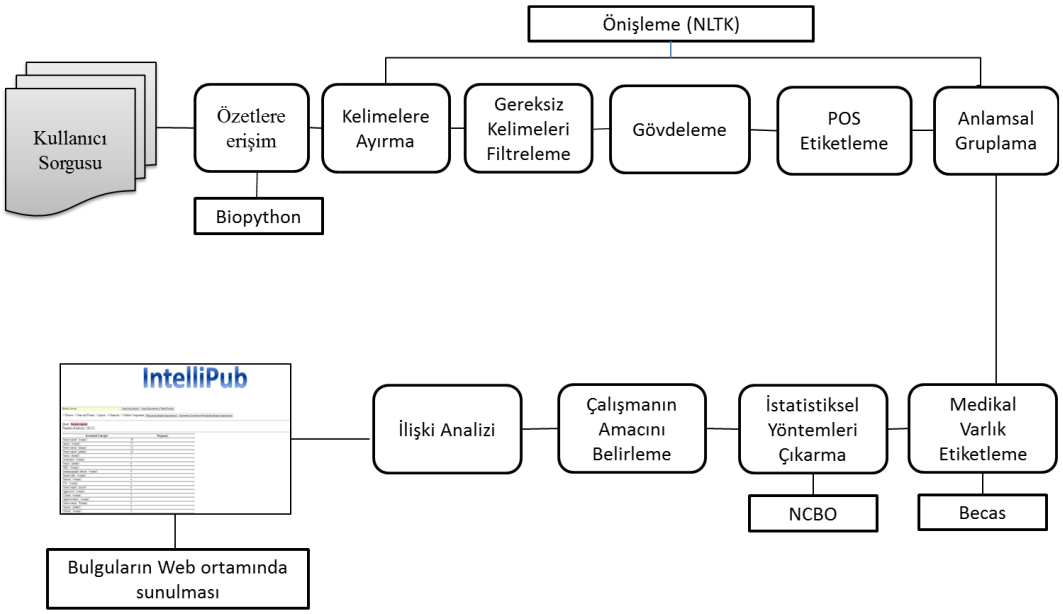
Literatürde birçok sistem, biyomedikal varlıkların metin içinde tanınması veya kullanıcı tarafından girilen sorgunun, eş anlamlı (http://www.ebi.ac.uk/citexplore/, Erişim Tarihi: 29 Şubat 2016), sıklıkla kullanıldığı terim (Eaton, 2006) veya kısaltmaların tam yazımlarıyla (http://twease.apps.campagnelab.org/medline/app, Erişim Tarihi:29 Şubat 2016) genişletilmesini veya çeşitlendirilmesini sağlamak amacıyla terminoloji kullanmaktadır. Böylelikle kullanıcının metin koleksiyonunda bulunan sorguyla ilişkili daha fazla sayıda ve doğrulukta dokümana erişimi ve erişilen dokümanlardan daha farklı tip medikal varlığın çıkarılması sağlanmaktadır (Lu, 2011). Fakat yapılan araştırmalardan sonra birçok sistemin daha çok genetik (gen-protein-hastalık ilişkisi) alanına yönelik olarak geliştirildiği görülmüştür (Rubinstein ve Simon, 2005; Rebholz-Schuhmann ve ark., 2007; Tudor ve ark., 2010; Xuan ve ark., 2010; Arighi ve ark., 2011; Bartsch ve ark., 2011). Bu da diğer alanlarda araştırma yapan bireyler için daha kısıtlı sayıda ve çeşitte sistemin geliştirildiğini göstermekte ve alandan bağımsız sistemlere olan ihtiyacı ön plana çıkarmaktadır.

3. GEREÇ VE YÖNTEM

Bu çalışma, sistem tasarımı ve geliştirme ile sistem performansını değerlendirme olmak üzere iki aşamada tamamlanmıştır.

3.1. Sistem tasarımı ve geliştirme süreci

Şekil 3.1'de sistemin işleyişine ait ardışık düzen verilmiştir. Şekil 3.1'de gösterilen ardışık düzende yer alan modüller ve kütüphaneler kullanılarak erişilen özetlerin işlenmesi ve ilgili bilginin çıkartılması sağlanmaktadır.



Şekil 3.1. Ardışık düzen

3.1.1. Programlama Dili: Python

Programlama dili, yazılımcının bir algoritmayı ifade etmek amacıyla, bir bilgisayara ne yapmasını istediğini anlatmasının tek tipleştirilmiş yoludur. Programlama dilleri, yazılımcının bilgisayara hangi veri üzerinde işlem yapacağını, verinin nasıl depolanıp iletileceğini, hangi koşullarda hangi işlemlerin yapılacağını tam olarak anlatmasını sağlar (https://tr.wikipedia.org/wiki/Programlama_dili, Erişim Tarihi: 03 Mart 2016).

Şu ana kadar 150'den fazla programlama dili yapılmıştır. Bunlardan bazıları Pascal, Basic, C, C#, C++, Java, Perl ve Python'dur (http://archive.oreilly.com/pub/a/oreilly//news/languageposter_0504.html, Erişim Tarihi: 03 Mart 2016). Bu tez çalışmasında, sistemin geliştirilmesi sürecinde hızlı

işlem yapabilme kapasitesi ve çok sayıda kütüphane içeriğine sahip olması göz önünde bulundurularak Python programlama dili kullanılmıştır. Python, ticari ve akademik amaçlı kullanılan yüksek seviyeli bir programlama dilidir.

Python, 1990 yılında Guido van Rossum tarafından, Amsterdam'da Centrum voor Wiskunde en Informatica (CWI) isimli araştırma enstitüsündeki Amoeba dağıtık işletim sistemi üzerinde çalışırken ABC dili yapısına benzeyen bir betik dili ve sistem yönetimi için de C'den veya kabuk betiklerinden daha etkin bir dile ihtiyaç duyulmasıyla geliştirilmeye başlanmıştır. Adını sanılanın aksine bir yılandan değil Guido van Rossum'un çok sevdiği, "Monty Python" adlı altı kişilik bir İngiliz komedi grubunun Monty Python's Flying Circus adlı gösterisinden almıştır (https://docs.python.org/3/faq/general.html, Erişim Tarihi: 03 Mart 2016). Python, nesne yönelimli, yorumlanabilen, birimsel (modüler) ve etkileşimli bir programlama dilidir. Python dilinin özellikleri;

- Nesneye yönelik
- Yorumlamalı ve derlemeli
- Taşınabilir
- Güçlü
- Hızlı
- Ticari uygulamalar geliştirmeye uygun
- Yazılımı kolay
- Öğrenmesi kolay

olarak sıralanabilir (Kuhlman, 2009).

Python ile sistem programlama, kullanıcı arabirimi programlama, ağ programlama, uygulama ve veri tabanı yazılımı programlama gibi birçok alanda yazılım geliştirebilmektedir. Büyük yazılımların hızlı bir şekilde prototiplerinin üretilmesi ve denenmesi gerektiği durumlarda zengin ve genişleyen kütüphane desteğiyle ve farklı platformlara entegre edilebilmesi özelliğiyle Python tercih edilmektedir. Youtube, Google, NASA ve CERN Python programlama dilini kullanan kurumlar arasındadır.

3.1.2. Özetlere Erişim

Geliştirilen sistem, literatür veri tabanı olarak biyomedikal alanda araştırma yapan bireylerin sık olarak tercih ettiği PubMed veri tabanını kullanmaktadır. Kullanıcı sorguları ile ilişkili Pubmed'de yer alan özetlere erişim için ise Biopython kütüphanesinden yararlanmaktadır.

Pubmed

PubMed, Amerikan devletine bağlı Ulusal Sağlık Enstitüsünün (İng. National Institute of Health, NIH) bir alt kuruluşu Ulusal Tıp Kütüphanesi (İng. National Library of Medicine, NLM) bünyesinde geliştirilen, MEDLINE veri tabanı, dergi ve çevrimiçi kitaplardan elde edilen 25 milyondan fazla makaleyi içeren bir veri tabanı ve arama motorudur (https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html, Erişim Tarihi: 01 Mart 2016). MEDLINE ise 1960'lı yıllarda NLM tarafından geliştirilen, şu anda 1946 yılından beri fen bilimleri alanında yazılmış tüm basılı makaleleri dizinleyen bir dergi makale veri tabanıdır. MEDLINE veri tabanı ile ilgili başlıca disiplinler şunlardır; tıp, hemşirelik, diş hekimliği, veterinerlik ve klinik öncesi bilimler. MEDLINE, NLM tarafından oluşturulmuş Medikal Konu Başlıkları (Medical Subject Headings, MeSH) yani tıbbi konu terimlerini kullanır. Ağaç diyagramı şeklinde erişilebilen alt başlıklar aracılığıyla 5,600'den fazla 39 dilde güncel biyomedikal dergide yer alan künye bilgileri taranabilmektedir. PubMed'e 1996 yılından beri aktif olarak erişilebilmekte ve kullanıcılara, MEDLINE'da dizinlenen makalelere ve diğer fen bilimleri alanında çıkan dokümanlara erişim imkanı sunmaktadır. PubMed son zamanlarda yapmış olduğu değişikliklerle özet ve tam metinlerin yansıra makalelere yapılan atıfları da veri tabanına dahil etmiştir. Bu veriler ışığında PubMed dünyada diğer sağlık arama motorları ile kıyaslandığında en avantajlı konumda yer almaktadır.

PubMed ara yüzünü kullanarak araştırmacılar, girdikleri anahtar kelimeler ile veri tabanında bulunan makaleleri sorgulayabilmekte, sonuçları belli filtrelerle (Tarih, makale türü, dil vb.) sınırlandırabilmekte ve sonuçları varsayılan olarak kronolojik sıraya göre liste şeklinde görebilmektedirler. Fakat araştırmacılar belli sınırlandırmalardan sonra bile uzun bir makale listesi ile karşılaşmakta ve tüm bu makaleleri elle incelemek zorunda kalmaktadırlar. Bu da araştırmacılara iş yükü ve zaman kaybına yol açmaktadır. Bu problemden yola çıkarak çalışmada veri tabanı olarak PubMed veri tabanında dizinlenen makale özetleri kullanılmakta ve özetlere

PubMed'e ait web servisleri kullanılarak erişilmektedir. Kullanıcıların girdikleri sorgu sonucu PubMed'den gerçek zamanlı olarak elde edilen makaleler analiz edilerek çalışmanın parametrelerine uygun şekilde kullanıcıya sunulmaktadır.

Biopython

Biopython, moleküler biyoloji ve biyoinformatik alanında kullanılmak üzere kod tekrarını önlemek için Python ile yazılmış birçok aracı içeren ve ücretsiz olarak erişilebilen bir kütüphanedir. Bu proje ile yüksek kaliteli yeniden kullanılabilir modül ve kod parçacıklarını yaratılarak Python programlama dilinin biyoinformatik alanında kullanılmasını kolaylaştırmak hedeflenmektedir (Cock ve ark., 2009). Bu çalışmada, girilen anahtar kelimelerle ilişkili Pubmed'de bulunan dokümanlara erişilmesi, kullanıcı sorgularındaki yazım hatalarının düzeltilmesi ve erişilen dokümanların Python'a uygun veri yapılarına dönüştürülmesi amacıyla Biopython kütüphanesi kullanılmıştır.

```
#!C:/Python34/python.exe
from Bio import Entrez
from Bio import Medline

Entrez.email = "basakoguz@akdeniz.edu.tr"
handle = Entrez.espell(term=query)
record = Entrez.read(handle)
handle.close()
if record["CorrectedQuery"] != "":
    query = record["CorrectedQuery"]
else:
    query = record["Query"]

handlee = Entrez.esearch(db="pubmed",sort='relevance',
                         usehistory="y",term=query+" "+"NOT Review[ptyp]",retmax=500)
record = Entrez.read(handlee)
idlist = record["IdList"]
count = record["Count"]
search_results = Entrez.read(Entrez.epost("pubmed", id=",".join(idlist)))
webenv = search_results["WebEnv"]
querykey = search_results["QueryKey"]
handlee.close()
handleee = Entrez.efetch(db="pubmed",webenv=webenv,query_key=querykey,rettype="medline",
                         retmax=number, retstart=startvalue, retmode="text")
records = Medline.parse(handleee)
records = list(records)
```

Şekil 3.2. Biopython kütüphanesi kullanılarak PubMed özetlerine erişim sağlayan kod bloğu

Şekil 3.2'de Pubmed özetlerine erişimi sağlayan, sorgu kelimelerindeki yazım hatalarını düzelten kod bloğu verilmektedir. Bu aşamada öncelikle kodlamanın başında biopython kütüphanesinden kullanılacak modüller çağırılmaktadır. Daha sonra sisteme kayıtlı eposta adresi verilmektedir. Kullanıcı sorgusunun "espell" özelliği ile yazım hatası kontrolü yapılmakta ve eğer hata varsa otomatik olarak düzeltilmektedir. Düzeltilmiş olan sorgu kelimeleri "esearch" özelliği ile belirli parametreler girilerek Pubmed veri tabanında aratılmaktadır. Daha öncede bahsedilği gibi Pubmed varsayılan olarak sonuçları kronolojik sırayla kişilere sunmaktadır.

Kişiler isterse bu sıralama şeklini ara yüzden değiştirebilir. Bu çalışmada sonuçlar "ilgi (İng. relevance)" kriterine göre sıralanmaktadır. "esearch" özelliğinden elde edilen ilgili özetlere ait Pubmed ID'ler (PMID) kullanılarak bir ID listesi oluşturulmaktadır. Bu liste "efetch" özelliğinde girdi olarak verilmekte ve özetlere erişim sağlanmaktadır.

3.1.3. Metin Ön İşleme

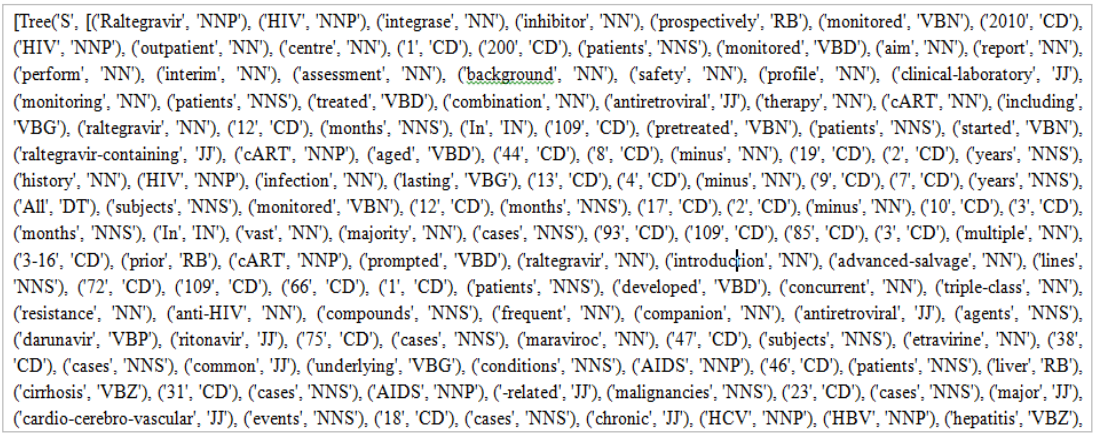
Veri madenciliğinde analiz edilecek giriş verilerinin belirli bir formata sahip olması ve bozuk veya gereksiz verilerden temizlenmiş olması gerekmektedir. Metin madenciliğinin en büyük sorunu, işleyeceği veri kümesinin yapılandırılmış olmamasıdır. Genellikle doğal dil kullanılarak yazılmış dokümanlar üzerinde çalışılan metin madenciliği alanında ön işleme aşaması, veri temizlemenin yanında veriyi uygun formata getirme işlemini de gerçekleştirmektedir (Feldman ve Sanger, 2007). Bu çalışmada, özetlerin DDİ yöntemleri ile işlenmesi aşamasında Python için oluşturulmuş Natural Language Toolkit (http://nltk.org, Erişim Tarihi: 22 Şubat 2016) isimli kütüphane kullanılmıştır. NLTK, ham metin verilerin işlenmesini sağlayan metin işleme ve analiz algoritmaları ile ilgili modülleri bünyesinde barındırmaktadır. NLTK kullanımı kolay, 50'den fazla metin koleksiyonunu ve sözlük yapısını (Örn: WordNet) içeren bir araçtır.

Metin önişleme adımları genel olarak aşağıda belirtildiği gibi beş adımdan oluşmaktadır;

- Özetleri kelimelere ayrıştırma (tokenization): Bir metnin incelenebilmesi için öncelikle kelimelere ve noktalama işaretlerine ayrılmış olması gerekir. Yani kelimeler ve noktalama işaretleri her biri bir öğeymiş gibi listelenir. NLTK içerisinde bu işlem için çeşitli modüller bulunmaktadır. Geliştirilen sistemde özetler, çeşitli modüllerin denenmesi sonucunda en yüksek performansı gösteren "Regexp Tokenizer" modülü kullanılarak kelime ve noktalama işaretlerine ayrıştırılmaktdır.
- Gereksiz kelimeleri filtreleme (stopwords filtering): Metinlerin boyutunu azaltmak ve performansı arttırmak için "ve, veya, bazen" gibi önemsiz olarak kabul edilen kelimelerin metinlerden çıkartılması gerekmektedir. Sistem NLTK kütüphanesinin içerisinde yer alan ve İngilizce için oluşturulan gereksiz kelime listesini kullanarak tüm önemsiz kelimeleri özetlerden çıkarmaktadır.

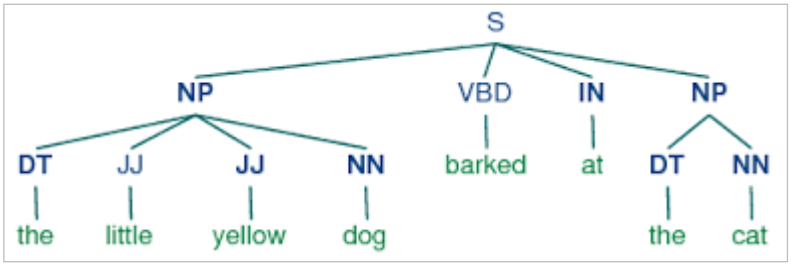
- Kelime gövdeleme (lemmatization): Kelimelerin esas formlarının bulunması, kelimenin gövde formuna dönüştürülmesi analizlerden doğru sonucu alabilmek için çok önemlidir. Örneğin; "cancer" ile "cancers" aynı varlığı ifade etmekte, fakat kelimenin aldığı ekten dolayı farklı varlıklarmış gibi algılanmaktadır. Bu durumun önlenebilmesi için de kelimelerin ana formuna, en temel yapısına indirgenmesi gerekmektedir. Sistem, NLTK modüllerinden performans olarak en iyi sonucu veren WordNetLemmatizer modülünü kullanarak kelimeleri temel formlarına dönüştürmektedir.

- Kelime sınıflarını belirleme (part-of-speech tagging (POS)): Her kelimenin anlamına veya görevine göre bir sınıfı bulunmaktadır. Türkçede isimler (isim, sıfat, zarf, zamir), fiiller ve edatlar (çekim edatları, bağlama edatları, ünlem edatları) olmak üzere üç ana grupta toplamak mümkündür. Bu kelimelerin otomatik olarak türlerine ayrılması ve sınıflandırılması doğal dilin yapısının anlaşılması için önemli bir yerde durmaktadır. Özellikle isimlendirilmiş varlıkların metin içerisinde tanınması için gerekli bir süreçtir. Geliştirilen sistem, Stanford Tagger modülünü kullanarak özetlerde bulunan kelimeleri ait olduğu kategorilere (isim, fiil, sıfat vb.) göre etiketlemektedir. Şekil 3.3'de örnek POS etiketleme sonucu verilmiştir.


[Tree('S', [('Raltegravir', 'NNP'), ('HIV', 'NNP'), ('integrase', 'NN'), ('inhibitor', 'NN'), ('prospectively', 'RB'), ('monitored', 'VBN'), ('2010', 'CD'), ('HIV', 'NNP'), ('outpatient', 'NN'), ('centre', 'NN'), ('1', 'CD'), ('200', 'CD'), ('patients', 'NNS'), ('monitored', 'VBD'), ('aim', 'NN'), ('report', 'NN'), ('perform', 'NN'), ('interim', 'NN'), ('assessment', 'NN'), ('background', 'NN'), ('safety', 'NN'), ('profile', 'NN'), ('clinical-laboratory', 'JJ'), ('monitoring', 'NN'), ('patients', 'NNS'), ('treated', 'VBD'), ('combination', 'NN'), ('antiretroviral', 'JJ'), ('therapy', 'NN'), ('cART', 'NN'), ('including', 'VBG'), ('raltegravir', 'NN'), ('12', 'CD'), ('months', 'NNS'), ('In', 'IN'), ('109', 'CD'), ('pretreated', 'VBN'), ('patients', 'NNS'), ('started', 'VBN'), ('raltegravir-containing', 'JJ'), ('cART', 'NNP'), ('aged', 'VBD'), ('44', 'CD'), ('8', 'CD'), ('minus', 'NN'), ('19', 'CD'), ('2', 'CD'), ('years', 'NNS'), ('history', 'NN'), ('HIV', 'NNP'), ('infection', 'NN'), ('lasting', 'VBG'), ('13', 'CD'), ('4', 'CD'), ('minus', 'NN'), ('9', 'CD'), ('7', 'CD'), ('years', 'NNS'), ('All', 'DT'), ('subjects', 'NNS'), ('monitored', 'VBN'), ('12', 'CD'), ('months', 'NNS'), ('17', 'CD'), ('2', 'CD'), ('minus', 'NN'), ('10', 'CD'), ('3', 'CD'), ('months', 'NNS'), ('In', 'IN'), ('vast', 'NN'), ('majority', 'NN'), ('cases', 'NNS'), ('93', 'CD'), ('109', 'CD'), ('85', 'CD'), ('3', 'CD'), ('multiple', 'NN'), ('3-16', 'CD'), ('prior', 'RB'), ('cART', 'NNP'), ('prompted', 'VBD'), ('raltegravir', 'NN'), ('introduction', 'NN'), ('advanced-salvage', 'NN'), ('lines', 'NNS'), ('72', 'CD'), ('109', 'CD'), ('66', 'CD'), ('1', 'CD'), ('patients', 'NNS'), ('developed', 'VBD'), ('concurrent', 'NN'), ('triple-class', 'NN'), ('resistance', 'NN'), ('anti-HIV', 'NN'), ('compounds', 'NNS'), ('frequent', 'NN'), ('companion', 'NN'), ('antiretroviral', 'JJ'), ('agents', 'NNS'), ('darunavir', 'VBP'), ('ritonavir', 'JJ'), ('75', 'CD'), ('cases', 'NNS'), ('maraviroc', 'NN'), ('47', 'CD'), ('subjects', 'NNS'), ('etravirine', 'NN'), ('38', 'CD'), ('cases', 'NNS'), ('common', 'JJ'), ('underlying', 'VBG'), ('conditions', 'NNS'), ('AIDS', 'NNP'), ('46', 'CD'), ('patients', 'NNS'), ('liver', 'RB'), ('cirrhosis', 'VBZ'), ('31', 'CD'), ('cases', 'NNS'), ('AIDS', 'NNP'), ('-related', 'JJ'), ('malignancies', 'NNS'), ('23', 'CD'), ('cases', 'NNS'), ('major', 'JJ'), ('cardio-cerebro-vascular', 'JJ'), ('events', 'NNS'), ('18', 'CD'), ('cases', 'NNS'), ('chronic', 'JJ'), ('HCV', 'NNP'), ('HBV', 'NNP'), ('hepatitis', 'VBZ'),

Şekil 3.3. POS etiketleme örneği

- Kelimelerin anlamsal olarak gruplanması (chunking): POS etiketleme kelime temelli olup kelimenin isim, fiil vb. olup olmadığını kontrol ederek kelimeleri etiketler. Chunking işleminde ise kelimeler anlamsal olarak gruplanmakta böylelikle isim veya sıfat tamlamaları gibi birden fazla kelimeden oluşan anlamlı kelime öbekleri metin içerisinde etiketlenebilmektedir (Şekil 3.4). RegexpParser modülü kullanılarak kelime öbeklerinin tanımlanması sağlanmaktadır.



Şekil 3.4. Chunking örneği

3.1.4. Biyomedikal Varlıkların ve İstatistiksel Terimlerin Etiketlenmesi

Birçok veri tabanı makalelerle ilgili önemli bilgileri içerse de, bu makaleler içerisindeki biyomedikal varlıkların tanınması ve bunların metinlerden çıkartılması araştırmacılar için önemli bir yerde durmaktadır. Biyomedikal varlıkların belirlenmesi için literatürde birçok yazılım (Savova ve ark., 2010a; Ananiadou ve ark., 2011; Papanikolaou ve ark., 2011; Torii ve ark., 2011) ve algoritma (Zhang ve Elhadad, 2013) geliştirilmiştir. Genel olarak geliştirilen sistemler sözlük tabanlı, kural tabanlı ve makine öğrenmesi algoritmaları tabanlı olmak üzere üç kategoride toplanabilir. Geliştirilen yazılımlarda, doğal dili anlayabilmek ve varlıklar arasındaki anlamsal ilişkileri ortaya çıkartılabilmek için kavramsal ontolojiler/bilgi tabanları kullanılmaktadır. Genel olarak terminolojiler, kavramların sınıf-altsınıf ve parça-bütün ilişkilerine göre oluşturulmaktadır. DrugBank (Wishart ve ark., 2006), UMLS (https://www.nlm.nih.gov/research/umls/, Erişim Tarihi: 12 Şubat 2016), MESH (https://www.nlm.nih.gov/pubs/factsheets/mesh.html, Erişim Tarihi: 23 Şubat 2016), BioThesaurus (Liu ve ark., 2006), LexEBI (Rebholz-Schuhmann ve ark., 2013) ve BioLexicon (Thompson ve ark., 2011) literatürde geliştirilen sistemlerde sıklıkla kullanılan kaynaklardır. MESH (Medical Subject Headings), tıbbi konu başlıkları terimlerini indeksleyen kavramsal sözlüktür. UMLS, yaygın olarak kullanılan, varlıklar arasındaki hiyerarşik ve anlamsal ilişkilerin tanımlandığı en önemli kavram dizinlerinden biridir. NLM tarafından geliştirilmektedir. Temel klinik kodlama ve referans sistemlerinin terminoloji, semantik ve formatları arasında bağlantılar kuran bir "metathesaurus" içeren bir sistemdir. Uzman bir "lexicon", bir "semantic" ağ ve bir enformasyon kaynakları haritalaması içermektedir. DrugBank, ilaçlar ile ilgili kapsamlı bilgileri içeren biyoenformatik veri tabanıdır. Bu veri tabanı, 1447 FDA (Food and Drug Administration) onaylı küçük moleküllü, 131 FDA onaylı biyoteknolojik, 85 beslenme ile ilgili ve 5080 deneysel ilaçlar olmak üzere 6711 ilaç bulunmaktadır. BioLexicon; UniProtKb, ChEBI ve NCBI taksonomi gibi

biyoenformatik ile ilgili veri kaynaklarını bir araya getiren terminolojidir. BioThesaurus, gen ve protein isimlerini içeren geniş ve kapsamlı bir terminoloji veri tabanıdır. LexEBI; Biothes, InterPro, JoChem gibi hastalıkları, enzimleri ve dokuları kapsayan birçok terminolojiyi içerisinde barındıran kavram veri kaynağıdır.

BeCAS Annotator

Bu çalışmada, sağladığı web servis olanağıyla sistemler tarafından rahatlıkla kullanılabilecek olan BeCAS (the Biomedical Concept Annotation System) Annotator (http://bioinformatics.ua.pt/becas/, Erişim Tarihi: 12 Ocak 2016) kullanılmıştır. BeCAS, sistemlerin metin işleme süreçlerinde doküman analiz etme ve terimleri sınıflarına göre metin içerisinde etiketleme işlemlerini yapmak amacıyla sistemlere entegre edilebilen veya kullanıcı dostu interaktif web ara yüzü ile normal kişiler tarafından da kullanılabilen bir web tabanlı araçtır. PubMed makalelerine erişim, cümle bölme, kelimelere ayırma, gövde formuna dönüştürme, POS etiketleme, chunking, varlık belirleme, kısaltmaları çözümleme ve interaktif görsel varlık vurgulama işlemlerini entegre eden bir araçtır. Metin işleme modülü Java programlama dili ile geliştirilmiş olup, makale erişimi ve web servis ayağı Python'da geliştirilmiştir (Nunes ve ark., 2013). Varlıkların belirlenmesi için geliştirilen modüller, Tablo 3.1'de de görüldüğü gibi türler, anatomik varlıklar, miRNA, enzimler, kimyasallar, ilaçlar, hastalıklar, metabolik yollar, hücresel bileşenler, biyolojik süreçler ve moleküler fonksiyonların tanınması için sözlük eşleştirme yöntemini kullanmaktadır. Bunun için, UMLS, NCBI BioSystems, LexEBI, ChEBI, miRBase ve Gen Ontoloji olmak üzere birçok terminolojiyi içeren bir veri tabanı oluşturulmuştur. Gen ve proteinlerin tanınması için Conditional Random Fields algoritması ile geliştirilmiş bir etiketleyici kullanmaktadır. BeCAS bu özellikleri ile araştırmacılara, sağlık bakım uzmanlarına ve geliştiricilere 1.200.000 biyomedikal varlığın tanımlanmasında yardımcı olmaktadır. BeCAS, CRAFT, AnEM ve NCBI hastalıklar metin koleksiyonlarında test edilerek, gen ve proteinler için %76, türler için %95, kimyasallar için %65, hücresel bileşenler için %83, hücreler için %92, moleküler fonksiyonlar ve biyolojik süreçler için %63, anatomik varlıklar için %83 ve hastalıklar için %85 f-ölçütü başarı oranına ulaşmıştır.

Tablo 3.1. BeCAS içerisinde yer alan varlık tipleri ve veri kaynakları (Nunes ve ark., 2013)

Semantik grup	Belirlenen varlık tipi	Veri Kaynağı
Türler	Türler	UMLS
Anatomy	Anatomik yapı	UMLS
	Lokasyon veya bölge	UMLS
	Organ ve Organ Bileşenleri	UMLS
	Vücut boşlukları veya eklemler	UMLS
	Vücut sıvısı	UMLS
	Vücut Sistemi	UMLS
	Hücre	UMLS
	Hücre Bileşenleri	UMLS
	Embriyo yapısı	UMLS
	Doku	UMLS
Hastalıklar	Edinilmiş Bozukluk	UMLS
	Anatomik Bozukluk	UMLS
	Hücre ve moleküler disfonksiyon	UMLS
	Konjenital Bozukluk	UMLS
	Hastalık veya sendrom	UMLS
	Zihinsel ve Davranışsal Bozukluk	UMLS
	Neoplastik Süreç	UMLS
	Patolojik Fonksiyon	UMLS
	Belirti ve semptomlar	UMLS
Yollar	Yol	NCBI BioSystems
Kimyasallar	Kimyasal	ChEBI
Enzimler	Enzim	lexEBI
miRNA	microRNA	miRBase
Genler ve proteinler	Gen protein	lexEBI Biothesaurus
Hücresel Bileşenler	Hücresel Bileşenler	GO+UMLS
Moleküler Fonksiyonlar	Moleküler Fonksiyonlar	GO
Biyolojik Süreçler	Biyolojik Süreçler	GO
	Hücre fonksiyonu	UMLS
	Genetik fonksiyon	UMLS
	Moleküler fonksiyon	UMLS
	Organ veya doku fonksiyonu	UMLS
	Physiologic Function	UMLS

Şekil 3.5'te verilen fonksiyon ile sistem kullanıcı sorgusu ile erişilen özetlerdeki biyomedikal varlıkları etiketleyebilmektedir. Öncelikle BeCAS web servisini kullanabilmek için sistemde kayıtlı bulunan eposta adresini ve geliştirilen aracın ismini vermek gerekmektedir. Servise erişim izni sağlandıktan sonra özetler modülle etiketlenmekte ve sonuçlar web ara yüzünde gösterilmektedir.

```
def becas_annotator(abstract):

    becas.email = 'basakoguz@akdeniz.edu.tr'
    becas.tool = 'becas-python'
    result = becas.annotate_text(abstract)
    result = result.get("entities")
    return result
```

Şekil 3.5. BeCAS erişim ve etiketleme fonksiyonu

Özetlerden İstatistiksel Terimlerin Çıkartılması

Halka açık bulunan biyomedikal verilerin çeşitliliği çok fazladır ve günden güne büyümektedir. Biyomedikal alanda çalışan araştırmacılar daha iyi arama ve erişim için verilerini yapılandırmada ve varlıkları metin içerisinde etiketlemede ontolojileri ve terminolojileri kullanmaktadır. Fakat bu süreç kolay bir şekilde otomatik hala getirilememekte ve uzman kişilere ihtiyaç duyulmaktadır. Ayrıca ontolojileri uygulamada kullanımını kolaylaştırmak için geliştirilen sistemlerin kullanım zorluğu bulunmaktadır. NCBO Annotator (diğer adıyla Open Biomedical Annotator (OBA)) (https://bioportal.bioontology.org/annotator, Erişim Tarihi: 12 Ocak 2016) veri setlerinde bulunan biyomedikal varlıkları etiketleyen halka açık ontoloji tabanlı bir web servisidir. Araştırmacılar veya sistem geliştiriciler bu servisi kullanarak kendi verilerindeki ontoloji varlıklarını otomatik olarak etiketleyebilmektedir. Bu varlıklar, UMLS ve NCBO Biyoportal içerisinde yer alan ontolojilerden gelmektedir (Jonquet ve ark., 2009). NCBO BiyoPortal veri havuzu (Noy ve ark., 2009) yaklaşık 300 terminoloji ve 5,4 milyon terim içermektedir (Bodenreider, 2004).

Bu çalışmada özetler içerisindeki istatistiksel terimleri etiketlemek ve özetlerden çıkartmak için NCBO Biyoportal veri havuzu içerisinde yer alan Ontology of Biological and Clinical Statistics (OBCS) ve Statistics Ontology (STATO) kaynakları kullanılmaktadır. Ayrıca NCBO Annotator tarafından etiketlenmeyen veya eksik etiketlenen terimlerin (Örneğin; "Per Protocol Analysis" NCBO Annotator tarafından "Protocol" olarak etiketlenmiş) saptanması ve sistemin performansını arttırmak için medikal istatistik terimlerini içeren bir sözlük (Everitt,

2006) kullanılarak bir anahtar kelime listesi oluşturulmuştur. Herhangi bir özet içerisindeki istatistiksel terimler ilk olarak NCBO Annotator ile etiketlenmektedir. Sonraki aşamada oluşturulan anahtar kelime listesindeki kelimeler kullanılarak özette bu kelimeler aratılmaktadır. Eğer NCBO Annotator tarafından eksik etiketlenmiş veya etiketlenmemiş fakat anahtar kelime listesinde yer alan bir istatistiksel terim bulunuyorsa bu terim sistem tarafından etiketlenmektedir.

```
def get_json(url):

    REST_URL = "http://data.bioontology.org"
    API_KEY = "809b28a6-f3d8-4a19-bdb8-e3895d001dd3"
    req = urllib.request.Request(url)
    req.add_header('Authorization', 'apikey token=' + API_KEY)
    r = urllib.request.urlopen(req)
    return json.loads(r.read().decode('utf8'))

def get_annotations(annotations, get_class=True):

    statistics = ""
    for result in annotations:
        class_details = get_json(result["annotatedClass"]["links"]["self"]) if get_class else result["annotatedClass"]
        if class_details["prefLabel"] not in methods:
            statistics += class_details["prefLabel"] +"|"+" "
    return statistics

def statistic_annotate(abstract):

    annotations = get_json(REST_URL + "/annotator?text=" + urllib.parse.quote(abstract))
    statistics = get_annotations(annotations)
    return statistics
```

Şekil 3.6. NCBO Annotator erişim ve etiketleme fonksiyonu

Şekil 3.6'da NCBO annotator kullanımı ve özetlerdeki istatistiksel terimlerin etiketlenmesi için oluşturulan kodlar verilmektedir. NCBO Biyoportal'da üyelik işlemleri tamamlandıktan sonra sistem her üye için bir API_KEY üretmektedir. Bu anahtar numara kullanılarak Annotator'ı kullanacak sistemler için gerekli özelliklere erişim izni verilmektedir. Bu yüzden ilk fonksiyonda görüldüğü gibi Biyoportal'da geliştirilen sistem için üretilen anahtar, parametre olarak verilmektedir. Annotator'a ait URL (İng. Uniform Resource Locator, Standart Kaynak Bulucu) adresi de parametre olarak verilerek istek gönderilmektedir. NCBO annotator web servisinden gelen sonuçlar JSON formatındadır ve ilgili JSON Kütüphanesi kullanılarak sonuçlar okutulmakta ve bu fonksiyonun döndürdüğü değişken olarak sunulmaktadır. İkinci fonksiyon ise döndürülen sonuçların çözümlenmesi işlevini yapmaktadır. JSON formatındaki sonuçlar ayrıştırıldıktan sonra istatistiksel terimin geçtiği sınıf bulunarak terim etiketlenmekte ve fonksiyon tarafından çıktı olarak verilmektedir. Son fonksiyonda ise önceki iki fonksiyonun çağırılması ve özetlerin girdi olarak verilerek içerisindeki istatistiksel terimlerin çıkartılması işlemleri yapılmaktadır.

```
def extractStatistics(methods, abstract):
    file = open("C:\documents\statisticsmethods.txt","r")
    listindex = file.read().splitlines()
    file.close()
    method = methods
    sentences = sent_tokenize(abstract.lower())
    for sentence in sentences:
        for term in listindex:
            if fuzz.token_set_ratio(term.lower(), sentence) >= 95 and
            fuzz.partial_token_set_ratio(term.lower(), method.lower())< 95:
                method += term + "|"+ " "
    return method
```

Şekil 3.7. Anahtar kelime listesi kullanılarak istatistiksel terim etiketleme fonksiyonu

Şekil 3.7'de anahtar kelime listesi kullanılarak istatistiksel terimleri çıkartan fonksiyon verilmektedir. Bu fonksiyonda, özet ve NCBO annotator tarafından çıkartılan terimler, girdi parametresi olarak verilmektedir. İlk olarak anahtar kelimeleri içeren metin dosyası açılmakta ve içerisindeki öğeler kullanılarak bir liste oluşturulmaktadır. Daha sonra özet, cümlelere ayrılarak her cümlede listedeki terimlerle eşleşen terim olup olmadığı FuzzyWuzzy kütüphanesinde bulunan "token_set_ratio" fonksiyonu ile kontrol edilmektedir. Eğer terim listede var ise ve NCBO Annotator çıktısında bulunmuyorsa etiketlenen istatistiksel terimlere eklenmekte ve fonksiyonun çıktısı olarak sunulmaktadır.

3.1.5. Özetlerden Makale Amacının Çıkarılması

Özet, bilimsel makalelerin yapıtaşlarından biridir. İyi yapılandırılmış bir özette çalışmanın amacının sunulması, kullanılan yöntemlerin ve çalışma tasarımının kısaca açıklanması ve çalışma sonucunda elde edilen önemli bulguların verilmesi gerekmektedir. İyi yapılandırılmış bir özet, araştırmacının ilk bakışta makalenin niteliği ve içeriği ile ilgili fikir sahibi olmasını sağlar. Bu yüzden, araştırmacılara çalıştıkları konu ile ilgili literatürü gözden geçirmelerine yardımcı olması ve genel olarak makalenin içeriği ile ilgili bilgiye erişebilmesi için bu tez çalışmasında, çalışmaların amacı, çalışmalarda kullanılan istatistiksel terimlerin ve içerisinde yer alan biyomedikal kavramların çıkartılması ve araştırmacılara tablo şeklinde sunulması hedeflenmiştir. Daha önceki kısımlarda istatistiksel terimlerin ve medikal varlıkların çıkartılması için kullanılan araçlar ve sistemde nasıl uygulanıldığı anlatılmıştır. Bu bölümde ise sistem tarafından özetlerden çalışmanın amacının nasıl çıkartıldığına değinilecektir.

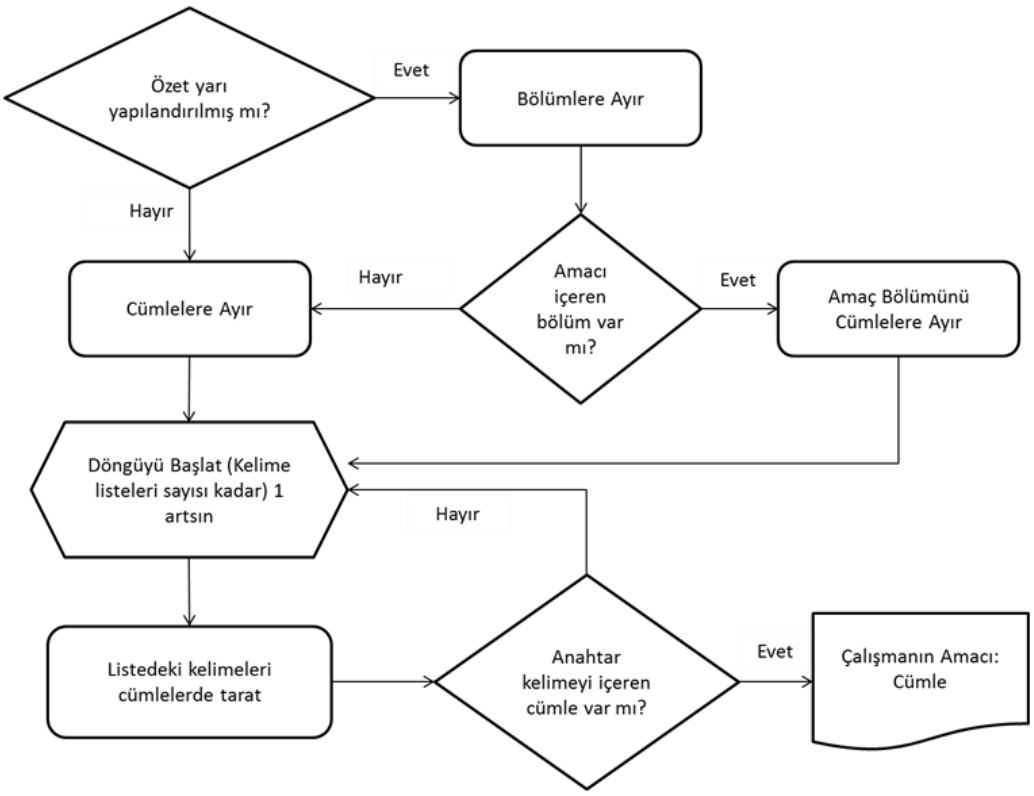
Özetlerde çalışmaların amaçlarının nasıl yazıldığı, hangi kelimelerle ifade edildiği ve sık kullanılan kelimelerin belirlenmesi için PubMed'de "breast cancer" anahtar

kelimeleri kullanılarak arama yapılmış ve son beş yılda yayınlanan makalelere ait özetlerden 1000 tanesi rastgele seçilerek bir geliştirme veri seti oluşturulmuştur. İlk olarak bu özetlerdeki amaçların yapısı incelenmiş ve elde edilen anahtar kelimeler ve yazım örüntüleri kullanılarak amaç çıkarma modülü geliştirilmiştir. Özetler yarı yapılandırılmış ve yapılandırılmamış olmak üzere ikiye ayrılabilir. Yarı yapılandırılmış formatta bulunan özetlerin (Purpose-Objectives, Methods-Materyal and Methods, Results, Conclusions) amaçlarının ilgili alanda verildiği, yapılandırılmamış paragraf şeklinde yazılmış özetlerde ise genellikle ilk üç cümle içerisinde amaçların yer aldığı görülmüştür. Amaçların ifade edilmesinde sıklıkla kullanılan kelimeler EK 1'de gösterilmektedir.

Şekil 3.8'de modülün, özetlerden çalışmaların amacını çıkartmak için kullandığı algoritma verilmektedir. Bu modülde ilk olarak özetlerin yarı yapılandırılmış veya yapılandırılmamaış formatta olup olmadığı incelenmektedir. Eğer yarı yapılandırılmış formatta yazılmışsa amacı içeren özel bir bölüm olup olmadığına bakılmaktadır. Yarı yapılandırılmış özetlerde amaç genellikle "OBJECTIVE", "PURPOSE:", "AIM:", "OBJECT:" ve "BACKGROUND:" alanlarında verilmektedir. Bazen bu alanlarda tek cümle ile amaç belirtilirken bazen de iki veya daha fazla cümle kullanılabilmektedir. Yarı yapılandırılmış özetlerde modül ilk olarak özetin bölümlerini belirler. İkinci adımda eğer var ise OBJECTIVE, PURPOSE, AIM, OBJECT ve BACKGROUND bölümlerindeki cümleler birbirinden ayrılır ve bu cümleler tek tek anahtar kelime listeleri kullanılarak taratılır ve anahtar kelimelerden herhangi birini içeriyorsa çalışmanın amacı olarak belirlenir. Yapılandırılmamış özetlerde ise modül ilk olarak özeti cümlelere ayırır ve ilk kelime listesindeki anahtar kelimeleri tüm cümlelerde arar eğer bulamazsa diğer kelime listelerini de sırasıyla kullanarak bu kelimelerden herhangi birini içeren cümleyi bulmaya çalışır. Herhangi bir cümlede anahtar kelimeler bulunamazsa sistem boş bir metin döndürür.

Anahtar kelimelerin ve benzer formatlarının sistem tarafından özetlerde tespit edilmesi için bulanık metin eşleştirme (Fuzzy String Matching) algoritması tercih edilmiş ve bu amaçla Python'a özgü geliştirilmiş FuzzyWuzzy kütüphanesi (https://github.com/seatgeek/fuzzywuzzy, Erişim Tarihi: 12 Mart 2016) kullanılmıştır. Bulanık metin eşleştirme pratikte yazım hatası düzeltme, metinin yeniden kullanımının tespit edilmesi (intihal), istenmeyen epostaların filtrelenmesi,

DNA dizilişlerinin eşleştirilmesi gibi birçok alanda kullanılmaktadır. Bu algoritma verilen bir örüntüde kelime ve kelime grubuna veya herhangi bir metne en yakın benzerlikteki öğeleri bulmaktadır. Amaç çıkarma modülünde FuzzyWuzzy kütüphanesinde bulunan "partial_token_set_ratio" metodu kullanılmıştır.



Şekil 3.8. Amaç çıkarma modülü

3.1.6. Web Ara yüzü ve Sonuçların Sunumu

Web, internet üzerinde yayınlanan birbirleriyle bağlantılı hipermetin dokümanlarından oluşan bir bilgi sistemidir. Bu dokümanların her birine Web sayfası adı verilir ve Web sayfalarına internet kullanıcısının bilgisayarında çalışan Web tarayıcısı adı verilen bilgisayar programları aracılığıyla erişilir (https://tr.wikipedia.org/wiki/World_Wide_Web, Erişim Tarihi: 09 Mart 2016). Bu çalışmada, web sayfalarının tasarımı HTML kodları kullanılarak basit bir metin editöründe oluşturulmuş ve kullanıcı sorgularının Python modüllerine gönderilmesi amacıyla bir Javascript kütüphanesi olan JQuery ve web programlama tekniği olan AJAX kullanılmıştır.

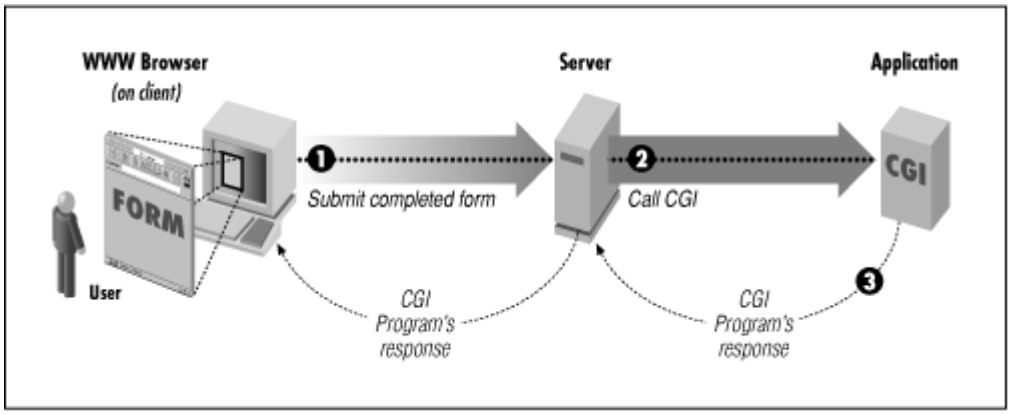
Web Sayfasından Python Modüllerine Erişim

İnternet yoluyla insanların kullanımına sunulmuş olan her dokümanın kendine ait ve tek olan bir adres yani URL'si vardır. URL'ler statik kaynaklarla ilişkilendirilebileceği gibi kullanıcı isteklerini işleyip sonucu kullanıcının tarayıcısına döndüren programlara da işaret edebilir. URL'lerde belirtilen programlara ağ geçidi programcıkları (gateway scripts) adı verilir. Ağ geçidi, bir arabirim sunan program veya cihazlardır. Ağ geçidi tarayıcı ile sunucu arasında bulunmaktadır. Ağ geçidi programcıklarının sunuculara bilgi iletmesi CGI'ler (İng. Common Gateway Interface, Ortak Ağ geçidi Arabirimi) ile yapılmaktadır (Stanek, 2000). CGI, web sunucuları ile bu sunucuların dışındaki programlar arasında etkileşim, diğer deyişle ortak çalışma platformu oluşturmak için geliştirilmiş bir standarttır, bir programdır. Web'in statik yapısına, HTML kodu içerisinden çağırılan CGI programları dinamik bir yapı kazandırmaktadır. CGI web kullanıcılarının web sunucusunun çalıştığı makine üzerinde belirlenen programları çalıştırmasını sağlayan bir ara yüz programıdır (https://tr.wikipedia.org/wiki/CGI, Erişim Tarihi: 09 Mart 2016). İşlemlerin akışı sırasıyla:

- Kullanıcının tarayıcıdan yaptıkları girişleri web sunucusuna aktarır.
- Sunucu sırası geldiğinde girişleri CGI programcığına aktarır.
- CGI programcığı girişi işler ve sonucu web sunucusuna gönderir.
- Sunucu çıktıyı kullanıcının tarayıcısına gönderir.

şeklinde olmaktadır (Şekil 3.9).

Programcıkların yazımında Java, C/C++, Perl, Python ve Visual Basic en çok kullanılan dillerdir. En popüler CGI uygulamalarından birisi Web Sayaçlarıdır. Web sayfalarını kaç kişinin ziyaret ettiğini saptayan küçük uygulamalar dış program olarak, bir web sayfası içinden çağrılabilir. CGI programları gerçek zamanlı çalışırlar. CGI'lerin en önemli uygulama alanlarından biri de, web üzerinde doldurup gönderilen formlar üzerindeki bilgileri sunucu tarafında değerlendirip kullanıcıya cevabı göndermektir.



Şekil 3.9. CGI mimarisi (Gundavaram, 1996)

Bu çalışmada, CGI yapısı kullanılarak HTML sayfasında kullanıcı tarafından girilen sorgu ilgili Python modülünde girdi olarak alınmakta ve istenilen işlem sunucu tarafında yapılmaktadır. Elde edilen program çıktıları yine aynı altyapı kullanılarak kullanıcıya bir HTML ara yüzü ile sunulmaktadır.

Birlikte Bulunabilirlik İstatistikleri ve Grafiksel Gösterim

Çalışmada etiketlenen varlıklar arasındaki ilişki örüntülerini ve birlikte bulunma istatistiklerini kullanıcılara sunmak için bir ara yüz tasarlanmıştır. Kişiler istedikleri kategorileri seçerek (hastalık, gen ve protein vb.) sorgu sonucunda elde edilen özetlerde ne kadar birlikte kullanıldıklarını tablo formatında görebilmektedir. Birlikte bulunma frekanslarının hesaplanmasında NLTK kütüphanesinde yer alan FreqDist modülü kullanılmıştır. Ara yüzde beş adet kategoriye ait seçim imkanı sunulmaktadır. Kullanıcılar bu kategorilerden istedikleri sayıda seçim yaparak, sorgu sonucunda erişilen özetlerde bu kategoriler içerisinde yer alan terimlerin birlikte bulunma frekansları özet başına hesaplanmakta ve ilk 25 tanesi sonuç olarak web ara yüzünde gösterilmektedir. Sistemden elde edilen sonuçların kullanıcılara grafiksel olarak sunulabilmesi için bir Python kütüphanesi olan Matplotlib kullanılmıştır. Matplotlib (Hunter, 2007), Matlab'ı andıran, 2 boyutlu grafik çizimi ve gösterimi için geliştirilmiş açık kaynak kodlu bir kütüphanedir. Bu kütüphane içerisinde yer alan modüller kullanılarak varlıkların birlikte bulunma istatistiklerine ait sonuçlar sütun grafikle web ara yüzünde gösterilmektedir.

3.2. Sistem Performansının Değerlendirilmesi

3.2.1. Amaç Çıkarma Modülünün Değerlendirilmesi

Sistem tarafından belirlenen amaçların doğruluğunun değerlendirilmesi için PubMed'de "helicobacter pylori" anahtar kelimeleri kullanılarak arama yapılmış ve

son 10 yılda yayınlanan makalelere (2007-2016) ait her yıldan 50 tane olmak üzere toplam 500 özet rastgele seçilerek bir değerlendirme veri seti oluşturulmuştur. Sorgu kelimeleri farklı olsa da geliştirme bölümünde kullanılan özetlerle benzersizliği sağlayabilmek için değerlendirme veri setindeki özetlere ait PMID'ler geliştirme veri setindeki özetlere ait PMID'ler ile karşılaştırılmış aynı olan özetler değerlendirme veri setinden çıkartılarak farklı bir özet eklenmiştir. Değerlendirme veri setindeki özetlere ait amaçlar hem modül tarafından hem de bir alan uzmanı tarafından belirlenmiş ve daha sonra birbirleriyle karşılaştırılmıştır. Uzman tarafından özetlerde hangi cümlenin amaç cümlesi olduğuna karar verilmesi sürecinde araştırmacıdan iki unsura dikkat edilmesi istenmiştir: 1) özette açık bir şekilde ifade edilen bir amaç cümlesi var mı? 2) amaç cümlesi dolaylı bir şekilde mi ifade edilmiş? Eğer amaç cümlesi açık bir şekilde verilmişse (Örneğin; "*The present study was undertaken to investigate whether H. pylori-induced activation of NF-kappaB and AP-1 mediates the expression of oncogenes and hyperproliferation of gastric epithelial cells*") veya dolaylı bir şekilde de olsa amaç olarak ifade edilmişse (Örneğin; "*The resistance rate in a random population has not been studied previously*") uzman bu cümleleri özetin amacı olarak nitelendirmiştir. Amaç cümlesi yapısına uygun veya herhangi bir amacı ifade eden bir cümle özette yer almıyorsa uzman bu özetlere ait amaç bölümlerini boş bırakmıştır.

Karşılaştırma için veri girişi yapılırken sistem ve uzman olmak üzere iki sütun oluşturulmuştur. Eğer modülün çıkardığı amaç ile uzmanın çıkardığı amaç aynı ise her iki sütuna da "1", hem uzman hem modül amaç cümlesi alanını boş bırakmışsa her iki sütuna da "0" ve uzman amacı boş bırakırken sistem yanlış bir cümleyi çıkartıyorsa sistem için "1" uzman için "0" girilmiştir.

3.2.2. İstatistiksel Terimleri Çıkarma Modülünün Değerlendirilmesi

Bu aşamada da amaç çıkarma modülünün değerlendirilmesinde oluşturulan veri seti kullanılmıştır. Bu veri setinde bulunan özetlere ait istatistiksel terimler hem bir alan uzmanı hem de modül tarafından çıkartılmış ve bulgular birbiriyle karşılaştırılmıştır.

Uzman tarafından özetlerde herhangi bir istatistiksel terim olup olmadığına karar verilmesi aşamasında üç unsura dikkat edilmesi istenmiştir:

1) Özette açık bir şekilde ifade edilen bir istatistiksel terim (çalışma türü, test adı, bulguların sunumunda kullanılan frekans, odds ratio vb. terimler) var mı?

2) Özette kaç tane istatistiksel terim verilmiş?

3) Terim adı verilmese bile kullanılan kelimeler (correlated, associated, frequent vb.) terim hakkında fikir veriyor mu?

Uzman değerlendirmesinde, etiketlenmemiş makale özetlerinde istatistiksel terim bulunma durumu ve terim sayısı incelenmiş, terimler açık ya da dolaylı bir şekilde verilmiş ise, özette terim bulunma durumunu "1" diğer durumlarda "0" olarak skorlanmış, terim sayıları ve etiketlenen istatistiksel terimler belirlenmiştir.

Sistemin ürettiği çıktı ve uzman değerlendirmesi tam eşleşmeli ve kısmi eşleşmeli olmak üzere iki farklı şekilde skorlanmıştır. Tam eşleşmede hem sistem hem de uzman tarafından etiketlenen tüm terimlerin aynı ve eşit sayıda olması durumunda sistemin etiketleme durumu "1", diğer durumlarda "0"olarak kodlanmıştır. Kısmi eşleşmede ise eğer sistemin etiketlediği terim sayısı uzman tarafından etiketlenen terim sayısının %50'sinden fazla ise sistemin terim etiketleme durumu"1" olarak diğer durumlarda ise "0" olarak kodlanmıştır.

3.2.3. Değerlendirme Aşamasında Kullanılan Performans Ölçütleri

Her iki modülün de değerlendirilmesi aşamasında, elde edilen bulgular ile sisteminin performansını ölçmek amacıyla aşağıda verilen ve bilgi erişim-çıkarım sistemlerinin değerlendirilmesinde kullanılan başarı ölçütleri hesaplanmıştır.

Tablo 3.2. Değerlendirme matrisi

	Uzman Pozitif	**Uzman Negatif**
Sistem Pozitif	a	b
Sistem Negatif	c	d

Kesinlik (**Precision**): Sistem tarafından doğru sınıflandırılmış örnek sayısının sistemin toplamda doğru dediği örnek sayısına oranıdır.

$$Kesinlik\ (K) = a/(a+b)$$

Hassasiyet (**Recall**): Sistem tarafından doğru sınıflandırılmış örnek sayısının, gerçekte pozitif olarak belirlenmiş tüm örnek sayısına oranıdır.

$$Hassasiyet\ (H) = a/(a+c)$$

F ölçütü (F measure): Kesinlik ve hassasiyet ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verir. Bunun için F ölçütü (F) tanımlanmıştır. F ölçütü, kesinlik (K) ve hassasiyetin (H) harmonik ortalamasıdır.

$$F\ ölçütü = 2KH/(K+H)$$

4. BULGULAR

Bu bölümde, geliştirilen web tabanlı sistemin işleyişini gösterebilmek için "cardiovascular diseases" anahtar kelimeleri kullanılarak yapılan sorgu ile birlikte sistemin özetlerden bilgi çıkarımı sürecine ve bu süreç sonunda elde edilen çıktıların kullanıcılara sunulması için tasarlanmış ara yüzlere değinilecektir.

4.1. Bilgi Çıkarım Süreci

4.1.1. Kullanıcı Sorgularının Sunucuya Gönderilmesi

Kullanıcılar tarafından web ara yüzünde bir sorgu girildiğinde sorgu kelimeleri ve sayfa numarası Jquery ve Ajax komutları ile sunucu tarafında bulunan ilgili python modüllerine gönderilmektedir. Şekil 4.1'de kullanıcı tarafından girilen sorguyu Python modüllerine iletmek için oluşturulan komutlar verilmektedir.

```
$("#buton").click(function() {
        $("#results").html("processing...");
        document.getElementById('sayfano').value = 1;
        $.ajax({
                url: "cgi-bin/SearchNormal.py",
                type: "get",
                data:{query:$('#query').val(),sayfano:$('#sayfano').val()},
                success: function(response) {
                    $("#results").html(response);
                    if ($("#results").text().indexOf("Service Unavailable") > 0){
                        $("#prev").hide();
                        $("#sayfano").hide();
                        $("#next").hide();
                    }
                    else {
                        $("#prev").hide();
                        $("#sayfano").show();
                        $("#next").show();

                    }

                },
                error: function(xhr) {
                alert(xhr)
                }
        });
});
```

Şekil 4.1. Jquery ve Ajax Komutları

"cardiovascular diseasess" anahtar kelimeleri kullanılarak sorgulama yapıldığında Ajax komutlarına "cardiovascular diseases" anahtar kelimeleri ve sayfa numarası parametre olarak verilmekte ve bu parametreler ilgili Python modülüne gönderilmektedir. Sayfa numarası kullanılarak modülde hesaplama yapılmakta ve elde edilen özet listesinde bulunan özetlerden hangi aralıktakilerinin gösterileceği

belirlenmektedir. Sunucu tarafında işlemler yapıldıktan sonra web sayfasında yer alan bir içerisinde sonuçlar gösterilmektedir. Eğer herhangi bir sıkıntıdan dolayı modülden cevap gelmezse bu içerisinde "Service Unavailable" metni gösterilmektedir.

4.1.2. Özetlere Erişim

Ajax komutu ile gönderilen "cardiovascular diseases" sorgu kelimeleri ve sayfa numarası, ilgili Python modülünde girdi parametre olarak alınmaktadır. Sistem, Biopython Kütüphanesi aracılığıyla Pubmed'de yer alan ve sorgu kelimesiyle "ilgili" olan ilk 500 özete erişim sağlamaktadır. Şekil 4.2'de Pubmed'den web servisinden çıktı olarak sunulan özet formatına örnek verilmiştir. Özetler, kendilerine ait yayın yılı, yazarlar, dili vb. özelliklerle Pubmed'de saklanmakta ve web servisi aracılığıyla erişim sağlandığında yine bu formatta sistem geliştiricilere sunulmaktadır. Elde edilen özetler liste formatına dönüştürülerek özete ait TI: Title, PMID: Pubmed ID, TA: Journal, DP: Publication Date, AU: Authors ve AB: Abstract özellikleri kaydedilerek sistem tarafından kullanılmaktadır.

4.1.3. Medikal Varlıkları Etiketleme

Medikal varlıkların özetlerde etketlenmesi amacıyla Becas Annotator web servisi kullanılmıştır. Pubmed'den erişilen özetler liste formatına dönüştürüldükten sonra etiketleme işleminin daha hızlı yapılabilmesi için öncelikle gereksiz kelimeler özetlerden çıkarılmakta ve kelimeler gövde formatına dönüştürülmektedir. Listede bulunan her bir özet Becas Annotator ile tek tek etiketlenerek Şekil 4.3'teki gibi etiketlenen varlıkları, sınıfları ve ait olduğu terminolojileri içeren liste formatında bir çıktı elde edilmektedir. Varlıkların etiketlenme biçimleri "Cardiovascular diseases|UMLS:C0007222:T047:DISO|370" örneğinde görüldüğü üzere varlığın ismi|terminolojisi, terminoloji ID'si, sınıfı|sıra numarası şeklindedir. Sistem öncelikle etiketlenen her varlığı "|" işaretini kullanarak parçalamakta ve bu parçalardan ikinci sırada olan yani sınıfı içeren öğeyi kullanarak varlığın ait olduğu sınıfı belirlemektedir. Örneğin; Eğer 2. öğe "DISO" kelimesini içeriyorsa sistem bu varlığı "hastalık" olarak tanımlamaktadır. Sınıflarına ayrıştırılan varlıklar sistemin web ara yüzlerinde gösterim şekillerine göre biçimlendirilerek kullanıcılara sunulmaktadır.

```
'TI': '[Risk factors of erectile dysfunction in patients with cardiovascular diseases].',
'MHDA': '2016/06/11 06:00', 'JT': 'Zhonghua nan ke xue = National journal of andrology',
'PG': '219-24', 'IS': '1009-3591 (Print) 1009-3591 (Linking)','TA': 'Zhonghua Nan Ke Xue', 'CRDT': ['2016/05/14 06:00'],
'AB': "OBJECTIVE: To investigate the penile erectile function of hospitalized male patients with cardiovascular diseases,
the incidence of erectile dysfunction (ED) in this cohort, and the relationship of ED with cardiovascular diseases and
its risk factors. METHODS: Using a self-designed questionnaire, we conducted an investigation among the hospitalized patients
in the Department of Cardiovascular Diseases of the First and Second Affiliated Hospitals of Xi'an Jiaotong University.
We measured their body height, body mass index (BMI), waist circumference, hip circumference, and blood pressure,
obtained their personal data, past history, metabolic indexes, and erectile function scores by IIEF-5, and analyzed
the risk factors of ED using univariate and multivariate logistic regression and OR analyses.
RESULTS: Totally, 225 valid questionnaires were included in this investigation, which showed a 66.7% incidence of ED,
15.8% mild, 27.0% mild to moderate, 17.6% moderate, and 6.3% severe. The incident rates of ED in the 18-35 yr, 36-49 yr,
50-65 yr, and > 65 yr age groups were 13.6%, 39.1%, 89.2%, and 91.2%, respectively. Univariate logistic regression analysis
manifested that the risk factors of ED in the patients with cardiovascular diseases included age (OR = 3.122, 95% CI 2.040-4.779),
smoking (OR = 1.768, 95% CI 1.209-2.584), BMI (OR = 1.261, 95% CI 1.114-1.427), total cholesterol (OR = 1.77, 95% CI 1.339-2.340),
TC/HDL (OR =1.715, 95% CI 1.349-2.181), hypertension (OR = 1.717, 95% CI 1.110-2.658), and coronary heart disease (OR = 2.235, 95% CI 1.
while multivariate logistic regression analysis showed the risk factors to be age (OR = 4.99, 95% CI 2.264-10.998), financial condition,
(OR = 2.804, 95% CI 1.127-6.976), smoking (OR = 2.109, 95% CI 1.179-3.772), BMI (OR = 1.414, 95% CI 1.136-1.760), and TC/HDL
(OR = 2.001, 95% CI 1.016-3.943). CONCLUSION: The incidence of ED is high in hospitalized patients with cardiovascular diseases
and rises with the increase of age. Age, smoking, financial condition, BMI, and TC/HDL are the risk factors of both ED and c
ardiovascular diseases, and financial condition is closely associated with ED.",
'SB': 'IM', 'OWN': 'NLM','AU': ['Xing JP', 'Ning L', 'Chen HM', 'Tan T'], 'PT': ['English Abstract', 'Journal Article'],
'DP': '2016 Mar', 'FAU': ['Xing, Jun-ping', 'Ning, Liang', 'Chen, Hui-ming', 'Tan, Tan'],
'PMID': '27172660', 'STAT': 'MEDLINE','EDAT': '2016/05/14 06:00','SO': 'Zhonghua Nan Ke Xue. 2016 Mar;22(3):219-24.','PL': 'China',
'RN': ['0 (Imidazoles)', '0 (Pyrimidines)', '89239-35-0 (CI 943)'], 'VI': '22', 'IP': '3', 'JID': '1010935921', 'PST': 'ppublish',
'MH': ['Adult', 'Aged', 'Blood Pressure', 'Body Height', 'Body Mass Index', 'Cardiovascular Diseases/*complications',
'Erectile Dysfunction/epidemiology/*etiology', 'Hospitalization', 'Humans', 'Hypertension/complications', 'Imidazoles',
'Incidence', 'Male', 'Middle Aged', 'Pyrimidines', 'Regression Analysis', 'Risk Factors', 'Smoking/adverse effects',
'Waist Circumference', 'Young Adult'], 'DA': '20160513', 'LA': ['chi'], 'DCOM': '20160610'}
```


Şekil 4.2. Erişilen özet formatı örneği

```
['Cardiovascular Diseases|UMLS:C0007222:T047:DISO|370', 'cardiovascular diseases|UMLS:C0007222:T047:DISO|1932',
'coronary heart disease|UMLS:C0010054:T047:DISO;UMLS:C0010068:T047:DISO|1509', 'ED|UMLS:C0242350:T047:DISO|1002',
'erectile|GO:0043084::PROC|650', 'penile|UMLS:C0030851:T023:ANAT|19', 'body|UMLS:C0460148:T017:ANAT;UMLS:C1268086:T017:ANAT|500',
'cardiovascular|UMLS:C0007226:T022:ANAT|79', 'ED|UMLS:C0242350:T047:DISO|143', 'cardiovascular|UMLS:C0007226:T022:ANAT|1216',
'hypertension|UMLS:C0020538:T047:DISO|1458', 'ED|UMLS:C0242350:T047:DISO|187', 'metabolic|GO:0044237::PROC|627',
'blood|UMLS:C0005767:T024:ANAT;UMLS:C0229664:T031:ANAT|567', 'HDL|UNIPROT:P28845:T116:PRGE;CHEBI:39025:T103:CHED|2039',
'condition|UMLS:C0012634:T047:DISO|2016', 'cardiovascular|UMLS:C0007226:T022:ANAT|2079', 'ED|UMLS:C0242350:T047:DISO|2072',
'cardiovascular|UMLS:C0007226:T022:ANAT|195', 'body|UMLS:C0460148:T017:ANAT;UMLS:C1268086:T017:ANAT|487',
'Cardiovascular|UMLS:C0007226:T022:ANAT|370', 'coronary|UMLS:C0018787:T023:ANAT|1509', 'ED|UMLS:C0242350:T047:DISO|1891',
'waist|UMLS:C1280087:T029:ANAT;UMLS:C0230097:T029:ANAT|523', 'erectile dysfunction|UMLS:C0242350:T047:DISO|121',
'ED|UMLS:C0242350:T047:DISO|719', 'condition|UMLS:C0012634:T047:DISO|1692', 'HDL|UNIPROT:P28845:T116:PRGE;CHEBI:39025:T103:CHED|1421',
'erectile|GO:0043084::PROC|121', 'ED|UMLS:C0242350:T047:DISO|907', 'cardiovascular diseases|UMLS:C0007222:T047:DISO|79',
'heart|UMLS:C0018787:T023:ANAT|1518', 'ED|UMLS:C0242350:T047:DISO|1192', 'cardiovascular diseases|UMLS:C0007222:T047:DISO|1216',
'HDL|UNIPROT:P28845:T116:PRGE;CHEBI:39025:T103:CHED|1824', 'cardiovascular diseases|UMLS:C0007222:T047:DISO|195',
'blood pressure|UMLS:C0005823:T040:PROC|567', 'cardiovascular diseases|UMLS:C0007222:T047:DISO|2079',
'hip|UMLS:C0019552:T023:ANAT;UMLS:C0022122:T023:ANAT|544', 'cholesterol|CHEBI:16113:T103:CHED|1373', 'condition|UMLS:C0012634:T047:DISO|2118',
'erectile|GO:0043084::PROC|26', 'cardiovascular|UMLS:C0007226:T022:ANAT|1932', 'ED|UMLS:C0242350:T047:DISO|2155']
```

Şekil 4.3. Becas Annotator ile etiketlenen özet örneği

Pubmed benzeri ara yüzde her varlık ait olduğu sınıfa göre Şekil 4.4'te gösterilen "vurgulama" fonksiyonu ile önceden belirlenen arka plan rengi ile vurgulanmaktadır.

```
def highlightdoc(doc, concepts, color):

    term = re.sub('\s+', '|', concepts)
    regex = re.compile(r'(\s*)((?:\b\s*(?:%s)\b)+)' % term, re.I)
    return regex.sub(r'\1<span style="background:%s">\2</span>' % color, doc)
```

Şekil 4.4. Vurgulama fonksiyonu

Sistem, içerisindeki varlıkları etiketlenmiş olan her bir özeti, etiketlenen varlıkları ve varlığın sınıfına ait önceden belirlenen rengi bu fonsiyona girdi olarak vermektedir. İlk olarak etiketlenen varlıkta kelime çözümelemesi yapılmakta ve kelimeler arasına "|" işareti konarak yeniden formatlanmaktadır. Örneğin; "cardiovascular diseases" varlığı "cardiovascular|disease" olarak biçimlendirilmektedir. Bu biçimlendirme ile birden fazla kelimeye sahip varlıklar tek kelime yapısına dönüştürülmektedir. İkinci satırda renklendirilecek varlık için büyük-küçük harf duyarlılığına (re.I) bakılmaksızın diziliş örüntüsü tanımlanmaktadır. Son satırda ise özet içerisinde bir önceki satırda belirtilen formatta bulunan varlıklar "\2' % color" HTML koduna dönüştürülerek fonksiyondan çıktı olarak verilmektedir. Bu HTML kodu ile web ara yüzünde varlıklar farklı renklerle gösterilebilmektedir.

4.1.4. İstatistiksel Terimleri Çıkarma

Makalelerde kullanılan istatistiksel terimler NCBO Annotator web servisi kullanılarak çıkartılmaktadır. İstatistiksel terimler sadece tablo formatında sonuçların sunulduğu ara yüzde verilmektedir. Sistem liste haline dönüştürülen özetleri web servise tek tek girdi olarak vermekte ve web servisten JSON formatında çıktı almaktadır. Şekil 4.5'te "cardiovascular diseases" arama terimleri kullanılarak yapılan sorgu sonucunda elde edilen bir özetten çıkartılan istatistiksel terimler gösterilmektedir.

Sistem JSON formatında bulunan çıktıyı çözümlemek için "json" adlı Python kütüphanesini kullanarak çıktıyı öğelerine ayırmaktadır. Etiketlenen her bir istatistiksel terim terminoloji ID'si, terimin adı ve ait olduğu terminoloji öğelerini içermektedir. "prefLabel" etiketi ile verilen istatistiksel terimler sistem tarafından her

özet için liste halinde tutulmakta ve tablo formatında sonuçların sunulduğu ara yüzde gösterilmektedir.

```
Class details
    id: http://purl.obolibrary.org/obo/BFO_0000034
    prefLabel: function
    ontology: http://data.bioontology.org/ontologies/OBCS
Class details
    id: http://purl.obolibrary.org/obo/BFO_0000034
    prefLabel: function
    ontology: http://data.bioontology.org/ontologies/STATO
Class details
    id: http://purl.obolibrary.org/obo/OBCS_0000052
    prefLabel: Incidence rate
    ontology: http://data.bioontology.org/ontologies/OBCS
Class details
    id: http://purl.obolibrary.org/obo/STATO_0000203
    prefLabel: cohort
    ontology: http://data.bioontology.org/ontologies/STATO
Class details
    id: http://purl.obolibrary.org/obo/OBI_0001000
    prefLabel: questionnaire
    ontology: http://data.bioontology.org/ontologies/OBCS
Class details
    id: http://purl.obolibrary.org/obo/OBI_0000066
    prefLabel: investigation
    ontology: http://data.bioontology.org/ontologies/OBCS
Class details
    id: http://purl.obolibrary.org/obo/OBI_0000066
    prefLabel: investigation
    ontology: http://data.bioontology.org/ontologies/STATO
```

Şekil 4.5. NCBO Annotator ile elde edilen çıktı örneği

4.1.5. Makalenin Amacını Belirleme

Makalelerin amaç cümleleri tablo formatında sonuçların sunulduğu ara yüzde gösterilmektedir. “cardiovascular diseases” arama terimleri kullanılarak yapılan sorgu sonucunda elde edilen özet listesindeki özetleri tek tek inceleyen sistem, özetin ilk cümlesinden başlayarak EK-1’de verilen amaç kelimelerini sırasıyla aramaktadır. Daha öncede bahsedildiği gibi eğer özet yarı yapılandırılmış formattaysa “OBJECTIVE”, “AIM” ve benzeri alanlarda anahtar kelimeleri aramaktadır.

Aşağıda, “cardiovascular diseases” arama terimleri kullanılarak elde edilen ve yarı yapılandırılmış formatta bulunan bir özet örnek olarak verilmiştir;

"OBJECTIVES: Reducing mortality due to cardiovascular diseasess especially in people less than 65 years is one of the main targets of WHO preventive programs. **<AIM>***This work aimed to analyse recent trends in cardiovascular mortality rates in Slovakia*.**<AIM>**
STUDY DESIGN: A descriptive study was implemented with a Joinpoint analysis.
METHODS: Analysis was of annual all circulatory, acute myocardial infarction mortality, and cerebrovascular disease mortality rates, between 1980 and 2010 for Slovakia. Data were stratified by sex and 10-year age group (age 25-85 years). The annual percentage change (APC) and significant changes in the trend were identified using joinpoint Poisson regression.
RESULTS: The standardized mortality rate for all cardiovascular diseasess declined in Slovakia between 1980 and 2010 by 25.7% and 30.5% for men and women, respectively. Joinpoint analysis of all cardiovascular diseases mortality rates demonstrated statistically significant changes in trends of APC decline for both genders. For men, acceleration in the rate of decline between 2001 and 2010 was observed APC -2.2 (95% CI = -3.5, -1.2) following a slowing of the rate of decline between 1980 and 2001, when the APC reached -0.5 (95% CI = -0.8, -0.3). For women the trend was similar. Between 2003 and 2010 acceleration in the decline was demonstrated APC -2.8 (95% CI = -4.3, -1.4).
CONCLUSION: The results of our analysis demonstrate the need to constantly address issues of cardiovascular diseasess, as mortality rates in Slovakia are among the highest within the European Union countries in the long term."

Bu özette bulunan amaç cümlesini belirlemek için sistem öncelikle özeti bölümlerine ayırmaktadır. Sistemin, özetin formatını tanımlayabilmesi ve bölümlere ayırma işlemini yapabilmesi için tüm bölüm isimlerini içeren bir liste oluşturulmuştur. Yukarıdaki özette amaç cümlesi "OBJECTIVES" bölümünde verilmiştir. Bu özet için sistem sadece bu bölümdeki cümlelerde anahtar kelimeleri aratmaktadır. Anahtar kelimelerin metinde bulunup bulunmadığını kontrol etmek için FuzzyWuzzy Kütüphanesi kullanılmakta ve her bir kelime için skor elde edilmektedir. Eğer bu skor belirlenen eşik değerden yüksek ise kelimenin bulunduğu cümle amaç cümlesi olarak belirlenmekte ve listedeki diğer özete geçilmektedir. Anahtar kelime listesinde bulunan "aim to analyse" kelimeleri yukarıdaki örnekte bulunan "aimed to analyse" kelimeleri ile %100 bir eşleşme sağladığı için bu cümle sistem tarafından amaç cümlesi olarak etiketlenmiştir.

4.1.6. Birlikte Bulunma Frekansları ve Grafiksel Gösterim

Özetlerdeki varlıklar arasındaki ilişkilerin tespiti için sistem varlıkların birlikte bulunma frekanslarını kullanmaktadır. "cardiovascular diseases" anahtar kelimeleri ile sisteme sorgu gönderildiğinde ilk olarak AJAX komutları ile Python modüllerine sorgu kelimeleri ve kullanıcı tarafından seçilen varlık sınıfları parametre olarak

gönderilmektedir. Ara yüzde beş varlık sınıfı için seçim olanağı sunulmaktadır. Kullanıcı isterse tek sınıf isterse birden fazla sınıfı seçerek sorgusunu gönderebilmektedir. Örneğin; kullanıcının "cardiovascular diseases" ile birlikte "Disease" ve "Chemicals" sınıflarını seçtiği varsayılırsa modüle "cardiovascular diseases" ve "disease, chemicals" ayrı ayrı girdi olarak verilmektedir. Sistem sorgu sonucunda elde edilen özetlerin ilk 100'ünde bu sınıflara ait varlıkların olup olmadığını araştırmakta ve eğer varsa her sınıfa özel, etiketlenen varlıkları içeren liste oluşturmaktadır. Bir özette bulunan varlık o özet için sadece bir kere listeye eklenmektedir. Örneğin Şekil 4.3'te örnek olarak verilen özette "Cardiovascular diseases|UMLS:C0007222:T047:DISO|370" varlığı birden çok sayıda bulunmaktadır. Sistem bu özet için sadece bir kere "Cardiovascular diseases|UMLS:C0007222:T047:DISO|370" varlığını listeye eklemekte yani hesaplamalar varlığın geçtiği özet sayılarına göre yapılmaktadır. Oluşturulan varlık listeleri Şekil 4.6'da verilen kod bloğu ile birleştirilmektedir.

```
def getwordmatrix(list1, list2):

come = []
for i in range(len(list1)):
    for j in range(len(list2)):
        w1, w2 = list1[i], list2[j]
        if w1 != w2:
            come.append((w1,w2))
return come

def getwordmatrix3(list1, list2, list3):

come = []
for i in range(len(list1)):
    for j in range(len(list2)):
        for h in range(len(list3)):
            w1, w2, w3 = list1[i], list2[j], list3[h]
            if w1 != w2 and w1 != w3 and w2 != w3:
                come.append((w1,w2,w3))
return come

def getwordmatrix4(list1, list2, list3, list4):

come = []
for i in range(len(list1)):
    for j in range(len(list2)):
        for h in range(len(list3)):
            for k in range(len(list4)):
                w1, w2, w3, w4 = list1[i], list2[j], list3[h], list4[k]
                if w1 != w2 and w1 != w3 and w1 != w4 and w2 != w3 and w2 != w4 and w3 != w4:
                    come.append((w1,w2,w3,w4))
return come

def getwordmatrix5(list1, list2, list3, list4, list5):

come = []
for i in range(len(list1)):
    for j in range(len(list2)):
        for h in range(len(list3)):
            for k in range(len(list4)):
                for l in range(len(list5)):
                    w1, w2, w3, w4, w5 = list1[i], list2[j], list3[h], list4[k], list5[l]
                    if w1 != w2 and w1 != w3 and w1 != w4 and w1 != w5 and w2 != w3 and w2 != w4 and w2 != w5 and w3 != w4 and w3 != w5 and w4 != w5:
                        come.append((w1,w2,w3,w4,w5))
return come
```

Şekil 4.6. Varlık listelerini birleştiren kod bloğu

Kod bloğu incelendiğinde liste sayısına göre ayrı ayrı fonksiyon yazıldığı görülmektedir. Örneğin; "disease, chemicals" sınıflarının seçildiği bir sorguda sistem bu sınıflara ait varlıkları içeren "disease" ve "chemicals" isimli iki tane liste oluşturmaktadır. Bu listeler "getwordmatrix(list1,list2)" isimli fonksiyona girdi olarak verilmektedir. Listelerin eleman aralıklarına göre döngü oluşturularak elemanlar sırasıyla benzerlik kontrolü yapılarak yeni bir listeye eleman olarak eklenmektedir. Bu liste Şekil 4.7'de verilen kod bloğunda kullanılarak varlıkların

birlikte geçtiği özet sayıları hesaplanmakta ve ilk 25 sonuç HTML kodu içerisine gömülerek web ara yüzünde sunulmaktadır.

```
fdist = nltk.FreqDist(cooccuredWords)
wordfreq = []
for word, frequency in fdist.most_common(25):
    wordfreq.append((word,frequency))
for word in wordfreq:
    results += "<tr><td>%s</td><td>%s</td></tr>"%(word[0], word[1])
```

Şekil 4.7. Frekans dağılımlarını hesaplayan kod bloğu

Şekil 4.8'de frekans dağılımları hesaplanan varlık gruplarının grafiğinin oluşturulması için yazılan kodlar verilmektedir. Frekans dağılımlarının gösteriminde "sütun grafik" türünün kullanılması tercih edilmiştir. Oluşturulan her bir grafik dosya "png" uzantılı resim dosyası şeklinde sunucuda bulunan "/images" dizinine kaydedilmektedir. Sorgu kelimelerine 0-1000 arasında rastgele üretilen bir sayı da eklenerek resim dosyasının ismi oluşturulmaktadır. Oluşturulan resim dosyası frekans dağılımlarının verildiği bir tabloyla birlikte web ara yüzünde kullanıcılara sunulmaktadır.

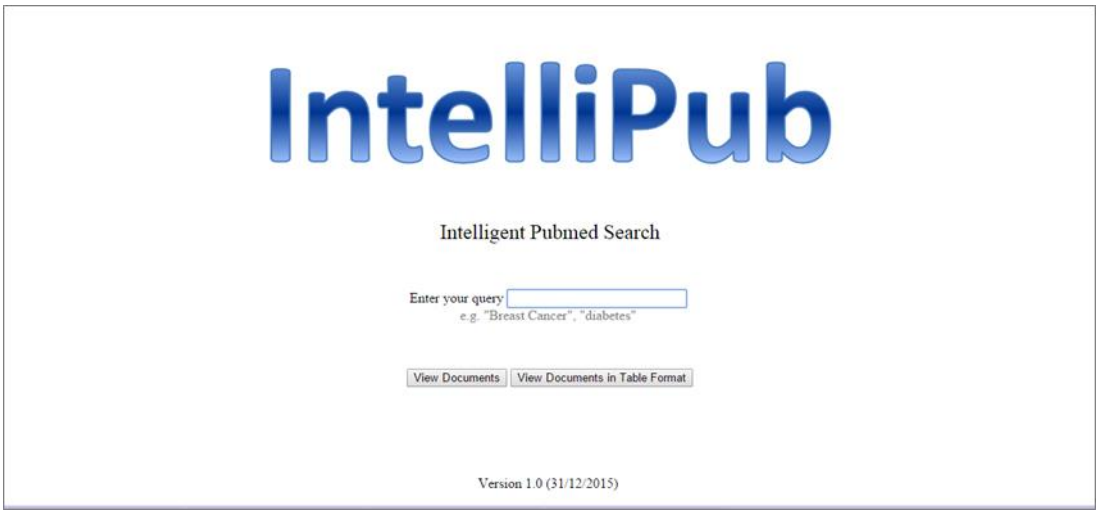
```
words = [x[0] for x in wordfreq]
values = [int(x[1]) for x in wordfreq]
figure = plt.figure()
indexes = np.arange(len(words))
width = 0.7
plt.bar(indexes, values, width)
plt.xlabel('Word Index')
plt.ylabel('Frequency')
plt.title('Word Frequency Chart')
randsayi = random.randint(0,1000)
outputfile = "images/" + query + str(randsayi) + ".png"
figure.savefig(outputfile)
results += "<div style='width: 50%; right: 1px; position: absolute; height: 100%; vertical-align: middle;'><img src='../"
results += outputfile + "' margin=auto display=block height=100% width=100%></div>"
print (results)
```

Şekil 4.8. Frekans dağılım grafiğini oluşturan kod bloğu

4.2. Web Ara yüzleri

Kullanıcıların sorgularını girebilmeleri ve sonuçları görüntüleyebilmeleri için 4 farklı ara yüz tasarlanmıştır. Şekil 4.9'da kullanıcılar sisteme girdiklerinde karşılarına çıkan giriş ara yüzü verilmektedir. Kullanıcılar istedikleri konuyla ilgili anahtar kelimeleri girerek sorgu yapabilmektedir. Bu ara yüzde; "View documents" ve "View documents in table format" olmak üzere iki adet buton bulunmaktadır. "View documents" butonuna tıklanıldığında Pubmed benzeri bir ara yüzle sonuçlar gösterilmekte, "View documents in Table format" butonu ise sonuçları tablo formatında sunmaktadır.

Giriş ara yüzünde bulunan metin kutusuna “cardiovascular diseases” anahtar kelimeleri girilip “View documents” butonuna tıklanılarak sunucu tarafında bulunan Python modülüne sorgu kelimeleri gönderilmektedir.



Şekil 4.9. Giriş ara yüzü

Şekil 4.10’da “cardiovascular diseases” sorgusu sonucunda Pubmed benzeri bir ara yüzle sonuçların verildiği ara yüz gösterilmektedir. Bu sayfada kullanıcının yaptığı sorgu sonucu elde edilen sonuçlar makalenin “adı”, “yazarları”, “yayın yılı”, “Pubmed numarası”, “dergisi” ve “özeti” ile birlikte sunulmaktadır. Özetler içerisinde geçen medikal varlıklar ekranın sağ tarafında verilen renklere göre etiketlenmektedir. Kullanıcının daha iyi ayırt edebilmesi için her sınıfa birbirinden farklı renkler atanmıştır. Bu ara yüzde kullanıcıya her sayfada 10 adet sonuç sunulmaktadır. Kullanıcılar makalenin ismine tıklayarak Pubmed’in kendi web ara yüzünde görüntüleme yapabilmektedirler. Kullanıcı “Next” veya “Previous” butonlarını kullanarak bir sonraki sayfadaki veya bir önceki sayfadaki sonuçları görebilmektedir. Bunun yanı sıra bu ara yüzün üst tarafında kullanıcının farklı arama kelimeleriyle sorgu yapabileceği, giriş ara yüzündeki özellikleri içeren bir form bulunmaktadır. Giriş ara yüzünden farklı olarak elde edilen sonuçlardaki farklı gruplardaki varlıkların birlikte bulunma frekanslarını hesaplayan “Frequency Based Association” butonu bulunmaktadır. Ayrıca bu formun alt kısmında bulunan alanda sorgu kelimesi ve erişilen makale sayısı da verilmektedir. Mesela “cardiovascular diseases” sorgusu ile erişilen özet sayısı sistem tarafından 1.824.438 olarak gösterilmektedir.

Şekil 4.10. Pubmed benzeri sonuç ara yüzü

Şekil 4.11.'de "cardiovascular diseases" sorgusundan elde edilen tablo formatında ara yüz gösterilmektedir. Bu ara yüzde de bir önceki ara yüzde bahsedilen form, aynı özellik ve nesnelerle yer almaktadır. Bu ara yüz sorgu sonucunda elde edilen özetlerin ve özetlere ait özelliklerin tablo formatında sunulduğu ara yüzdür. Bu tabloda sırasıyla makalenin Pubmed numarası, başlığı, geliştirilen sistem tarafından çıkartılan çalışmanın amacı, makalenin içerisinde geçen varlıklar ve kullanılan istatitistiksel terimler gösterilmektedir. Kullanıcılar PMID sütununda bulunan numaralara tıklayarak makalenin özetine Pubmed'ten erişebilirler. Pubmed bezneri ara yüzde olduğu gibi bu ara yüzde de her sayfada 10 özete ait sonuçlar sunulmaktadır. Kullanıcılar "Next" ve "Previous" butonuna tıklayarak bir sonraki sayfadaki veya bir önceki sayfadaki sonuçları görebilmektedir.

Şekil 4.12'de "cardiovascular diseases" anahtar kelimeleri kullanılarak yapılan sorgu sonucunda elde edilen ilk 100 özetten çıkartılan ve "Disease" ve "Chemicals" sınıflarında bulunan varlıkların birlikte bulunma frekanslarının ve bu analize ait grafiksel gösterimin sunulduğu ara yüz gösterilmektedir. Kullanıcılar "checkbox" öğelerine tıklayarak istedikleri sınıfı seçebilmekte ve bu sınıftaki varlıklara ait birlikte bulunma frekanslarını görüntülüyebilmektedirler. Sistem ilişkili varlıkların isimlerini ve frekaslarını içeren bir tablo ile sonuçları sunmaktadır. Ayrıca grafiksel gösterimle görsel olarak da kullanıcıların faydalanabileceği bir özellik sağlanmaktadır.

No	PMID	Title	Aim of Study	Disorder	Anatomy	Species	Chemical	Gene and Protein	Pathway	Enzyme	mRNA	Cellular Components	Statistical Methods
1	27172660	[Risk factors of erectile dysfunction in patients with cardiovascular diseases].	To investigate the penile erectile function of hospitalized male patients with cardiovascular diseases, the incidence of erectile dysfunction (ED) in this cohort, and the relationship of ED with cardiovascular diseases and its risk factors.	coronary heart disease| ED| condition| erectile dysfunction| hypertension| Cardiovascular Diseases|	waist| body| hip| penile| cardiovascular| coronary| heart| blood|		HDL| cholesterol|	HDL|				95% CI| Cohort| Logistic regression| Regression analysis|
2	26195612	Cardiovascular health in Italy: Ten-year surveillance of cardiovascular diseases and risk factors: Osservatorio Epidemiologico Cardiovascolare/Health Examination Survey 1998-2012.	The aim of this paper is to assess time trends of cardiovascular diseases, lifestyles, risk factors and high risk conditions in different socioeconomic levels.	myocardial infarction| obesity| diabetes| cardiovascular disease| dyslipidemia| stroke|	cardiovascular| blood|	men| woman|						Mean| Prevalence|
3	27004348	KNOWLEDGE AWARENESS AND BEHAVIOUR OF NON-MEDICAL STUDENTS ABOUT CARDIOVASCULAR DISEASES.	The purpose of this study was to evaluate the awareness, knowledge and the preventive measures taken to avoid the health issues related to cardiovascular diseases.	cardiovascular disease| diabetes|	cardiovascular| blood| coronary artery| heart|		salt| cholesterol|	SPSS-16|				Target population|
4	26507637	Trends in standardized mortality rates for select groups of cardiovascular diseases in Slovakia between 1980 and 2010.	This work aimed to analyse recent trends in cardiovascular mortality rates in Slovakia.	annual percentage change| cardiovascular disease| APC| acute myocardial infarction|	annual percentage change| cardiovascular| APC|	woman| men| people|				annual percentage change| APC|	95% CI| Mortality| Poisson regression| Standardized mortality rate| study design|
5	26978604	[The stages of development of cardiovascular diseases and the evolution of their pattern in the veterans of the Great Patriotic War (according to the 1946-2015 records of the Saint Petersburg War Veterans Hospital)].	To determine the stages of development of cardiovascular diseases and the evolution of their pattern in the veterans of the Great Patriotic War during 70 postwar years.	hypertensive disease| myocardial infarction|	cardiovascular| coronary| cardiopulmonary| blood| heart|				gas|			Average| Comorbidity| Morbidity|	

Şekil 4.11. Tablo formatında sonuç ara yüzü

Associated Concepts	Frequency
('cardiovascular disease', 'lipid')	7
('cardiovascular disease', 'cholesterol')	6
('cardiovascular disease', 'air')	6
('cardiovascular disease', 'glucose')	4
('diabetes', 'cholesterol')	3
('cardiovascular disease', 'NO')	3
('diabetes', 'lipid')	3
('metabolic syndrome', 'cholesterol')	3
('emergency', 'air')	3
('cardiovascular disease', 'ozone')	2
('CVD', 'high-density lipoprotein cholesterol')	2
('cardiovascular disease', 'lost')	2
('atherosclerosis', 'lipid')	2
('obesity', 'angiotensin')	2
('cardiovascular disease', 'uric acid')	2
('cardiovascular disease', 'inhibitor')	2
('endothelial dysfunction', 'nitric oxide')	2
('cardiovascular disease', 'copper')	2
('stroke', 'angiotensin')	2
('diabetes mellitus', 'glucose')	2
('hypertension', 'air')	2
('adhesion', 'nitric oxide')	2
('CVD', 'glucose')	2
('cardiovascular disease', 'LDL')	2

Şekil 4.12. Frekans tabanlı ilişki örüntüleri

4.3. Sistem Değerlendirme Sonuçları

4.3.1. Amaç Çıkartma Modülünün Değerlendirme Sonuçları

Sistem tarafından belirlenen amaçların doğruluğunun değerlendirilmesi için PubMed'de "helicobacter pylori" anahtar kelimeleri kullanılarak arama yapılmış ve son 10 yılda yayınlanan makalelere (2007-2016) ait her yıldan 50 tane olmak üzere toplam 500 özet rastgele seçilerek bir değerlendirme veri seti oluşturulmuştur. Karşılaştırma sonucunda elde edilen bulgular ile sisteminin performansını ölçmek amacıyla bilgi erişim-çıkarım sistemlerinin değerlendirilmesinde kullanılan kesinlik, hassasiyet ve f-ölçütü başarı ölçütleri hesaplanmıştır. Tablo 4.1.'de sistem ve uzman tarafından verilen kararlara ait 2x2 tablo gösterilmektedir. 394 özete ait amaç sistem tarafından doğru çıkartılmıştır. 78 özette ise sistem ya bir amaç cümlesi bulamamış ya da yanlış cümleyi amaç cümlesi olarak atamıştır. 28 özette uzman bir amaç cümlesi bulamamış, sistem ise bu özetlerin 21 tanesinde yer alan herhangi bir cümleyi amaç cümlesi olarak belirlemiş, 7 özette ise uzman ile aynı karara vararak amaç yok olarak çıktı vermiştir.

Tablo 4.1. Amaç çıkarma modülünün değerlendirme sonuçları

		Uzman	
	Amaçlar	Var	Yok
Sistem	Var	394	21
	Yok	78	7
	Toplam	472	28

Tablo 4.2.'de hesaplamalar sonucunda elde edilen başarı ölçütleri yüzde olarak verilmektedir. Özellikle f-ölçütü değeri (%88,9) sistemin amaçları çıkarmada oldukça başarılı olduğunu göstermektedir.

Tablo 4.2. Amaç çıkarma modülünün performans yüzdeleri

Ölçütler	**Sonuçlar (%)**
Kesinlik	94,9
Hassasiyet	83,5
F-ölçütü	88,9
Doğruluk	80,2

4.3.2. İstatistiksel Terimleri Çıkartma Modülünün Değerlendirme Sonuçları

Sistem tarafından belirlenen istatistiksel terimlerin doğruluğunun değerlendirilmesi için amaç çıkarma modülünün değerlendirilmesi aşamasında oluşturulan veri seti kullanılmıştır. Değerlendirme sonucunda elde edilen karşılaştırma bulgularının yorumlanabilmesi için değerlendirme sonuçları tam eşleşme ve kısmi eşleşme olarak ikiye ayrılarak kesinlik, hassasiyet ve f-ölçütü başarı ölçütleri hesaplanmıştır. Uzman 295 özette toplam 632 adet istatistikle ilgili terim etiketlemiştir. Bu terimlerden 517 tanesini sistem de doğru olarak etiketlemiştir. 205 özette hem uzman hem de sistem tarafından herhangi bir istatistiksel terime rastlanmamıştır. Tablo 4.3'te tam eşleşmeli değerlendirme sonuçlarınaa ait 2x2 tablo gösterilmektedir. Bu tabloda eğer sistem araştırmacının çıkardığı terimlerden bir tanesini bile bulamamışsa uzman için "var", sistem için ise "yok" girilmiştir. 192 özetin içinde geçen istatistiksel terimlerin hepsi sistem tarafından doğru çıkartılmıştır. 91 özette sistem herhangi bir istatistiksel terim bulamamış, 12 özette ise araştırmacı bir terim bulamazken, sistem yanlış terimi etiketlemiştir.

Tablo 4.3. İstatistiksel terimleri çıkarma modülü tam eşleşme değerlendirme sonuçları

		Uzman	
Sistem	**Terimler**	Var	Yok
	Var	192	12
	Yok	91	205
	Toplam	283	217

Kısmi eşleşmede eğer sistem herhangi bir özette geçen terimlerin yarısından fazlasını buluyorsa sistemin tüm terimleri bulduğu kabul edilmiş ve Tablo 4.4'te sonuçlar verilmiştir. Bu varsayımdan yola çıkarak 250 özette sistem terimleri doğru belirlemiş, 33 tanesinde ise herhangi bir terim bulamamıştır.

Tablo 4.4. İstatistiksel terimleri çıkarma modülü kısmi eşleşme değerlendirme sonuçları

		Uzman	
Sistem	**Terimler**	Var	Yok
	Var	250	12
	Yok	33	205
	Toplam	283	217

Tablo 4.5.'te hem kısmi hem de tam eşleşmeli değerlendirme sonucunda elde edilen başarı ölçütlerine ait yüzdeler verilmektedir. Sonuçlar incelendiğinde kısmi eşleşmede %91,7 f-ölçüt yüzdesi yakalanırken, tam eşleşme de bu değer %78,8'e düşmüştür. Benzer şekilde sistemin hassasiyet ölçütü %88,3'ten %67,8'e düşmüştür. Her iki sonuç ele alındığında kesinlik ölçütünde büyük bir azalış olmadığı görülmektedir.

Tablo 4.5. İstatistiksel terimleri çıkarma modülü performans yüzdeleri

Ölçütler	**Kısmi Eşleşme (%)**	**Tam Eşleşme (%)**
Kesinlik	95,4	94,1
Hassasiyet	88,3	67,8
F-ölçütü	91,7	78,8
Doğruluk	79,2	90,8

5. TARTIŞMA

Bu çalışmada, sağlık bakım uzmanları ve araştırmacılar tarafından klinik veya akademik çalışmalarda güncel literatür bilgisine ulaşmak amacıyla sıklıkla kullanılan Pubmed makale veri tabanından gerçek zamanlı olarak erişilen özetleri metin madenciliği teknikleri ile analiz eden, özetlerde yer alan tıbbi varlıkları etiketleyen, bu varlıklar arasındaki birlikte bulunma frekanslarını hesaplayan ve literatür gözden geçirme yapan araştırmacılara farklı web ara yüzleri ile sistem çıktılarını sunan web tabanlı bir literatür madenciliği uygulaması geliştirilmiştir.

Literatür tarama, her biyomedikal araştırmacının kendi bilimsel keşif süreçlerinde uyguladıkları temel adımlardan biridir. Ayrıca, sağlık bakım uzmanlarının sağlıkla ilgili bilgi arama ve yeni bulgularla önemli klinik kararlar verme sürecinde de literatür tarama önemli bir yere sahiptir. Var olan biyomedikal literatüre tam erişim ve alanla ilgili istenilen gerçek bilgiyi elde etmek, akademik ve klinik uzmanlığın önemli bir unsuru haline gelmiştir (Kumar ve ark., 2012). Pubmed, sağlık bakım uzmanları ve sağlık alanında araştırma yapan kişiler için günümüzde en sık kullanılan literatür veri tabanlarından biridir. Sağlık alanında araştırma yapan kişiler Pubmed'te arama yaparken ilgilendikleri alanla alakalı anahtar kelimeleri kullanarak sorgu oluştururlar ve bu sorgu sonucunda Pubmed sunucuları dizinledikleri makale özetlerinden ilgili olanları kişilere liste halinde sunar. Kişiler bu listede ilgilendiği makalenin ismine tıklayarak makalenin özetine ulaşmakta ve özet, ilgilendiği konu ile alakalı ise makalenin tam metnini okumaktadır.

Bilgi teknolojilerindeki gelişmelere rağmen, biyomedikal literatürün büyüklüğü, hızla büyümeye devam etmesi ve biyomedikal araştırmanın giderek multidisipliner olması nedenleriyle biyomedikal literatür taramayı kolaylaştırıcı çalışmaların sayısında aynı oranda gelişme sağlanamamıştır (Khare ve ark., 2014). Bu nedenle, bilgi erişim, veri madenciliği, doğal dil işleme ve bilgi çıkarımı alanlarındaki gelişmiş bilgi teknolojilerinin kullanılarak literatürdeki bilgiye erişimi hızlandıracak çalışmalar yapılmaya başlanmıştır (Khare ve ark., 2014). Biyomedikal literatürdeki makalelere ait özetleri analiz eden, tıbbi varlıkları çıkartan ve bu varlıklar arasındaki ilişki örüntülerini saptayan sistemlerin sayısı özellikle son 10 yılda hızla artmıştır ve hala artmaktadır.

Bu çalışmada metin madenciliği teknikleri kullanılarak geliştirilen web tabanlı sistem, makale özetlerine ve makalenin dergi adı, basım yılı vb. özelliklerine Pubmed web servislerini kullanarak gerçek zamanlı olarak erişmektedir. Sorgu sonucu elde edilen özetler ve makale özellikleri hem Pubmed benzeri bir ara yüzle hem de tablo formatında sunulmaktadır. Pubmed benzeri ara yüzde liste halinde özetler gösterilmekte ve özetler içerisinde bulunan farklı katagorilerdeki tıbbi varlıklar ait oldukları sınıflara göre farklı renklerde vurgulanarak sunulmaktadır. Kullanıcılar bu ara yüzle herhangi bir linke tıklamadan direkt olarak özetlerdeki en çarpıcı öğeleri daha net bir şekilde görebilmektedir. Tsuruoka ve arkadaşları (2008) tarafından geliştirilen FACTA adlı sistemde kullanıcı sorgusu sonucunda elde edilen özetler kullanıcılara benzer bir ara yüzle sunulmaktadır. FACTA, özetler içerisinde geçen sorgu kelimelerini farklı renkle vurgularken bu çalışmada geliştirilen sistemde kullanıcı sorgusuna göre elde edilen özetler içerisinde geçen medikal terimler sınıflarına göre farklı renklerle vurgulanmaktadır. Bu şekilde kullanıcı daha açık ve hızlı bir şekilde özetleri gözden geçirebilmekte ve yorumlayabilmektedir. Benzer bir ara yüz Wei ve arkadaşları (2013) tarafından geliştirilen PubTator isimli web tabanlı sistemde de kullanılmıştır. PubTator kullanıcı sorgusu sonucunda elde edilen özetleri kapalı liste (özet içerikleri verilmemekte) halinde sunmakta, eğer kullanıcı özeti görmek isterse "ABSTRACT" yazısına tıklayarak özetleri içerisindeki medikal terimler etiketlenmiş olarak görebilmektedir. Bu çalışmada kullanıcılara doğrudan sonuçları göstermek ve ekstra iş yükü oluşturmamak ve zamandan tasarruf sağlamak amacıyla özetleri açık olarak sunmak tercih edilmiştir. PubTator beş sınıfa (hastalık, tür, mutasyon, gen ve kimyasal) göre özetler içerisindeki terimleri etiketlerken, geliştirilen sistem dokuz sınıfa (hastalık, tür, gen ve protein, kimyasal, yol, anatomi, enzim, mikroRNA ve hücresel bileşenler) ait medikal terimleri özetler içerisinde etiketleyerek daha kapsamlı sonuçlar sunmaktadır.

Sistemin literatüre en büyük katkısı ve diğer sistemlerden farklı olarak tasarlanan özelliği kullanıcıya sorgu sonuçlarını tablo görünümünde sunmasıdır. Literatürdeki sistemler incelendiğinde genel olarak medikal terimler arasındaki ilişkileri çıkaran (Rebholz-Schuhmann ve ark., 2007; Tsuruoka ve ark., 2008; Frijters ve ark., 2010), özetleri yeniden dizinleyerek daha ilgili özetleri kullanıcılara sunan (Fontaine ve ark., 2009; Yu ve ark., 2010) veya kullanıcı sorgularını genişleterek daha iyi sonuçlara erişim sağlayan (Eaton, 2006) sistemler geliştirildiği görülmüştür. Fakat

araştırmacılar için elde edilen sonuçların daha anlaşılır ve yapılandırılmış bir şekilde sunulması da çok önemlidir. Bu nedenle, geliştirilen sistemde araştırmacıların elde edilen özetleri daha ayrıntılı görebilmesi ve yorumlayabilmesi için sonuçların tablo formatında sunulmaktadır. Tablo formatında yer alan özellikler; Pubmed ID (PMID), başlık, özet, amaç, istatistiksel terimler ve makale içerisinde yer alan medikal varlıkların sınıflarına göre ayrı sütunlarda gösterimi şeklindedir. Daha önce yapılan çalışmalarda daha çok sonuçların Pubmed benzeri ara yüzle sunulup özetler içerisindeki kelimelerin vurgulanması şeklinde tasarımlar kullanılmıştır (Tsuruoka ve ark., 2008; Wei ve ark., 2013). Geliştirilen sistem ile sorgu yapan bir kullanıcı ise hem makalelerin özetlerini liste şeklinde hem de tablo şeklinde özetin bir anlamda yapılandırılmış hali olarak görebilmektedir. Kullanıcının makaleye istediği zaman erişimini ve daha ayrıntılı olarak görebilmesini sağlayabilmek amacıyla Pubmed web adresleri kullanılarak tabloda sunulan pmid numaralarına köprü verilmiştir. Ayrıca kişiler bu tabloyu hesap tablosuna aktararak çalışmalarının literatür gözden geçirme kısmında veya doğrudan sistematik gözden geçirme çalışmalarında ön hazırlık verisi olarak kullanabilirler.

Çalışmanın en önemli katkılarından biri de erişilen özetlerdeki amaç cümlelerinin otomatik olarak çıkartılmasıdır. İyi yazılmış bir özet ister yapılandırılmış formatta olsun ister yapılandırılmamış formatta, çalışmanın amacına yer vermelidir (Andrade, 2011). Çalışmanın amacını belirten cümle, okuyucuya çalışmanın hedeflerini, hangi hipotezlerin test edileceğini ve hangi türde çalışmaların ve literatürün refere edileceğini göstermektedir. Eğer çalışmada amaç cümlesine yer verilmezse, okuyucu teknik konularda ve veriler içerisinde kaybolabilir veya çalışmada hangi unsurun önemli olduğunu anlamazsa çalışmayı okumadan geçebilir. Bu nedenle erişilen makalelerin amacını hızlıca gözden geçirmek araştırmacılara araştırmanın hipotezleri ve ilgilendiği değişkenler hakkında bilgi sağlamakta, araştırmacının ilgilendiği konuyla örtüşüp örtüşmediği hakkında fikir vererek, gözden geçirme işlemini kolaylaştırmaktadır. Geliştirilen literatür madenciliği sistemleri incelendiğinde çalışma amaçlarını çıkaran bir özelliğe rastlanmamıştır. Fakat bazı çalışmalarda Pubmed veri tabanındaki özetler kullanılarak özet cümlelerini sınıflandırma veya özet cümlelerinin içeriklerini analiz etme gibi konuların işlendiği görülmüştür. Chung (2009) randomize kontrollü çalışmalara ait yapılandırılmış özetlerde bulunan bölüm isimlerini belirleyip önceden belirlenen standart bir formata (aim, methods,

results, conclusion) dönüştürmüştür. Kim ve arkadaşları (2011) Conditional Random Fields yöntemini kullanarak önceden belirlenen kategorilere (background, population, intervention, outcome, study design, other) göre özetlerdeki cümleleri sınıflandırmışlardır. Benzer bir çalışmada destek vektör makinaları ile Medline veri tabanından elde edilen yapılandırılmamış formatta bulunan randomize kontrollü çalışma özetlerindeki cümleler Introduction, Methods, Results ve Conclusions olmak üzere dört kategoride sınıflandırılmıştır (McKnight ve Srinivasan, 2003). Her üç çalışmada da çalışmanın amacını ifade eden cümle çıkartılmamış, ilgili kategorilere yöntemler tarafından belirlenen cümle veya cümleler atanmıştır. Bu çalışmada makalelerin özetlerinden makalenin amacını çıkartan sözlük tabanlı bir modül geliştirilmiştir. Modül çalışmanın amacını belirlemek için oluşturulan anahtar kelime listelerini kullanmakta ve sadece bir cümleyi amaç cümlesi olarak belirlemektedir. Hsu ve arkadaşları (2012) küçük olmayan hücreli akciğer kanseri ile alakalı 42 randomize konrollü çalışmaların tam metinlerini kullanarak hipotezleri çıkartan bir çalışma yapmışlardır. 7 tam metinli makalenin bulunduğu değerlendirme veri setini kullanarak %83 kesinlik, %91 hassasiyet ve %86 f-ölçütü skoruna ulaşmıştır. Bu çalışmada geliştirilen modül alandan bağımsız olarak tasarlanmış ve değerlendirme aşamasında da "helicobacter pylori" ile alakalı rastgele seçilen 1000 özet kullanılmıştır. Modülün performansının değerlendirilmesi aşamasında elde edilen sonuçlar analiz edilerek modülün performans ölçütleri (kesinlik, hassasiyet ve f-ölçütü) hesaplanmıştır. Amaç çıkarma modülünün kesinlik, hassasiyet ve f-ölçütü değerleri sırasıyla %94,9, %83,5, %90'dır. Her iki çalışma kıyaslandığında ve kullanılan değerlendirme veri seti boyutları düşünüldüğünde daha iyi bir sonuç elde edildiğini söyleyebiliriz.

Amaç çıkarma modülünün performans ölçütleri iyi olmasına rağmen modülde geliştirilmesi gereken bazı unsurlar bulunmaktadır. Modülün en önemli kısıtlarından biri; eğer bir özet modül içerisinde yer alan anahtar kelimelerden herhangi birini içermiyorsa modül o özet için bir amaç cümlesi belirleyememektedir. Diğer bir problem ise modül içerisinde bulunan anahtar kelimeler cümle sırasına göre aratıldığı için eğer o kelimeleri içeren ve amaç cümlesinden önce verilen bir cümle varsa (Örn; bazı özetlerde önceki çalışmalara ait amaç cümleleri verilmiş) sistem yanlış etiketleyebilmektedir. Örneğin; "*Then in a previous study, we demonstrated that chloroform extract of Cistus laurifolius possessed a significant anti-Helicobacter*

pylori activity" cümlesinde olduğu gibi yazar özette bir önceki çalışmasına ait amaç cümlesine yer vermiş, fakat modül tarafından var olan çalışmanın amaç cümlesi olarak belirlenmiştir. Diğer bir kısıt ise bazı özetlerde amaç cümlelerinin olumsuz kelimelerle veya varolan makalelerdeki eksikleri ifade eden cümlelerle verilmesidir. Örneğin; "*Helicobacter pylori infection has been consistently associated with lack of access to clean water and proper sanitation, but no studies have demonstrated that the transmission of H. pylori can occur from drinking contaminated water.*" cümlesinde amacı ifade etmek için literatürdeki eksiklikten bahsedilmiştir. Bu cümle anahtar kelimeleri içerdiği için sistem tarafından etiketlense de buna benzer başka örneklerde sistemin amaç cümlesini belirleyemediği görülmüştür. Bu tarz hataları önleyebilmek amacıyla ilerleyen aşamalarda sadece sözlük tabanlı yaklaşım yerine hem sözlük hem de kural tabanlı yaklaşımın birlikte kullanıldığı yeni bir yöntem geliştirilmesi planlanmaktadır.

Sistemin diğer bir artı yönü ise özetlerde kullanılan istatistiksel terimlerin çıkartılmasıdır. İstatistik, verilerin toplanması, analiz edilmesi ve sonuçta elde edilen bulgulardan çıkarımlar yapılması sürecidir. Hem çalışmanın tasarımı aşamasında hem de toplanan verinin analiz edilmesi sürecinde istatistik bilimi önemli bir yere sahiptir (Davidian ve Louis, 2012). Bu çalışmada, makalelerde kullanılan istatistiksel terimler belirlenerek, araştırmacılara çalışmanın önemi hakkında bilgi verilmesi ve güncel yöntemlerin takibini kolaylaştırarak kendi çalışmalarının tasarım aşamasında ve çalışmalarında kullanılabilecek istatistiksel yöntemlerin belirlenmesi konusunda fikir verilmesi amaçlanmıştır. Çalışmada, özetlerde kullanılan istatistiksel terimlerin çıkartılabilmesi için NCBO annotator ve araştırmacılar tarafından oluşturulan anahtar kelime listesi kullanılmış olup literatür madenciliği için geliştirilen mevcut sistemlerde bu tarz bir özelliğe rastlanmamıştır. Sadece Hsu ve arkadaşlarının (2012) yaptıkları çalışmada benzer yöntemler kullanılarak çalışmalardaki istatistiksel terimlerin çıkartıldığı görülmüş ve çalışmanın performans ölçütleri %95 kesinlik, %76 hassasiyet ve %84 f-ölçütü skoru olarak bildirilmiştir. Bu çalışmada geliştirilen modülün değerlendirme sonuçları kısmi ve tam eşleşme olmak üzere ikiye ayrılmış ve performans ölçütleri hesaplanmıştır. Çalışmanın kısmi eşleşme değerlendirme sonuçları %95,4 kesinlik, %88,3 hassasiyet ve %91,7 f-ölçüt olarak hesaplanırken, tam eşleşme değerlendirme sonuçları sırasıyla %94,1, %67,8 ve %78,8 şeklindedir. Hsu ve arkadaşlarının (2012) yaptıkları çalışmadan elde edilen sonuçlar umut vaat

edici olsa da daha öncede bahsedildiği gibi çalışmada sadece küçük olmayan hücreli akciğer kanseri ile ilgili randomize kontrollü çalışmalara ait 42 tam metin kullanılmış olup herhangi bir ara yüzle kullanıcılara sonuçlar sunulmamaktadır.

Geliştirilen sistem medikal varlıkları etiketlemek için Becas annotator web servisini kullanmaktadır. Becas annotator Python dili ile birçok sınıftaki varlıkları metinlerde etiketlemek için geliştirilmiş bir sistemdir. Hem web tabanlı bir ara yüzle normal kullanıcılar erişebilmekte, hem de sistem geliştiriciler tarafından web servisi aracılığıyla kendi geliştirdikleri modüller içerisinden kullanılabilmektedir. Bu araç birçok ontoloji ve terminolojiyi bünyesinde barındırdığı için tercih edilmiştir. Çünkü çalışmanın en önemli hedeflerinden biri medikal terimlere ait sınıf sayısını (Becas annotator 11 kategoride etiketlemektedir (Tablo 3.1).) arttırarak daha ayrıntılı bir bilginin kullanıcıya sunulmasıdır. Ayrıca Becas Annotator'ın performans değerlendirilmesinde hesaplanan f-ölçütü değerleri (gen ve protein-%76, türler-%95, kimyasallar-%65, hücresel bileşenler-%83, hücreler-%92, moleküler fonksiyonlar ve biyolojik süreçler-%63, anatomik varlıklar-%83 ve hastalıklar-%85) literatürde yer alan birçok sisteme göre daha iyi olduğu görülmektedir (Nunes ve ark., 2013). Bu web servis kullanılırken yaşanılan en büyük problem, servisin hizmet dışı olduğu durumlarda kullanıcıya herhangi bir sonuç döndürülememekte, bunun yerine hata mesajı gönderilmektedir. Ayrıca bu servis 10 özeti yaklaşık 40 saniyede etiketliyebilmekte ve özet sayısı arttıkça işlem süresi uzamaktadır. Eğer çok fazla özet için işlem yapılması istenirse (örneğin: 100 ve üzeri) sistem yanıt veremeyebilmektedir. Literatürdeki sistemler incelendiğinde genel olarak var olan terminoloji veya ontolojiler kullanılarak geliştirilen sisteme ait annotator tasarlandığı görülmüştür (Egorov ve ark., 2004; Plake ve ark., 2006; Hur ve ark., 2009; Tudor ve ark., 2010; Wei ve ark., 2013). Bu sebeple çalışmanın ilerleyen aşamalarında sistemin kendisine ait bir annotator geliştirilerek hem sürenin kısaltılması hem de daha çok dokümanı analiz edebilmesi planlanmaktadır.

Var olan sistemler incelendiğinde varlıklar arasındaki birlikte bulunma ilişkilerinin ikili (hastalık-ilaç vb.) (Tsuruoka ve ark., 2008; Frijters ve ark., 2008) olarak verildiği görülmüştür. Bu çalışmada, hastalık, gen ve protein, kimyasallar, hücresel bileşenler ve tür olmak üzere toplam beş sınıf belirlenmiş ve kullanıcılara istedikleri sayıda sınıfı seçme olanağı sağlanmış olup seçilen sınıflarda yer alan varlıkların sorgu sonucunda elde edilen özetlerde birlikte bulunma frekansları hesaplanarak en

yüksek skora sahip ilk 25 sonuç kullanıcılara sunulmaktadır. Böylelikle kullanıcılar kendi ilgilendikleri ve sorguları sonucunda elde ettikleri özetler içerisindeki farklı sınıflardaki varlıkların birlikte bulunma örüntülerini görebilmekte ve hatta daha önceden bilmedikleri bilgilere erişebilmektedir. Bu bölümde sistemin en büyük eksikliği, sadece ilk 100 özet içerisinde yer alan birliktelikleri kullanıcılara sunuyor olmasıdır. Daha öncede bahsedildiği gibi sistem varlıkların etiketlenmesinde Becas Annotator web servisini kullanmakta ve bu servis 100 adetten fazla özeti etiketlerken ya çok uzun sürede yanıt vermekte ya da zaman aşımı uyarısı vermektedir. Bu yüzden ilerleyen aşamalarda Pubmed'de yer alan özetlerin etiketlenmiş versiyonlarını içeren bir metin koleksiyonu oluşturularak daha fazla özetin analiz edilmesi planlanmaktadır.

Çalışmanın en önemli kısıtlılıklarından biri işlemlerin yapılabilmesi için geçen süredir. Daha öncede değinildiği gibi sistem tüm işlemlerini gerçek zamanlı yapmaktadır. Literatürde geliştirilen birçok sistemde makale özetleri ve istenilen özellikler gecelik olarak bir veri tabanında toplanmakta, özetlere ait hesaplamalar bu veri tabanında tutulmakta ve kullanıcı sorgusunu girdiğinde ilgili sorguyla alakalı sonuçlar kullanıcıya web ara yüzüyle sunulmaktadır (Hristovski ve ark., 2005; Plake ve ark., 2006; Rebholz-Schuhmann ve ark., 2007; Frijters ve ark., 2008; Tsuruoka ve ark., 2008; Fontaine ve ark., 2009; Barbosa-Silva ve ark., 2010; Wei ve ark., 2013). Fakat bu tarz bir altyapının kullanılması maliyetli bir iştir ve çalışmalarda genelde bu vurgulanmıştır. Geliştirilen sistem pubmed özetlerine gerçek zamanlı erişmekte ve özetler içerisindeki kavramlar o süre içerisinde etiketlenmektedir. Bu işlem çok zaman aldığı için ve pubmed web servisleri çok fazla istek gönderildiğinde cevap vermediği için geliştirilen sistemin de kullanıcıya sonuçları döndürmesi uzun sürmekte, hatta bazı zamanlarda süre aşımından veya servisteki yoğunluktan dolayı hata mesajı gönderilmektedir. Bu tarz hataları en aza indirmek ve işlem süresini kısaltmak için her sayfada 10 özete ait sonuçlar kullanıcılara sunulmaktadır. Kişiler navigasyon butonları ile bir sonraki sayfadaki sonuçlara veya önceki sayfadaki sonuçlara erişebilmektedir. Ayrıca her sorgu için o sorguya ait ilk 500 sonuç (içerisinde özetleri ve özetlere ait özellikleri içeren) sorgu kelimeleri kullanılarak hesap tablosu olarak kaydedilmekte ve 7 gün sonunda klasörden silinmektedir. Eğer başka bir kullanıcı 7 gün içerisinde aynı sorgu kelimeleri ile arama yaparsa sistemde var olan hesap tablosu kullanılarak istenilen sonuçlar sağlanmaktadır. Örneğin; bir

kullanıcı “type 2 diabet” kelimeleriyle sisteme bir sorgu gönderdiğinde sistem öncelikle aynı isimde ve son 7 gün içinde üretilmiş bir hesap tablosu olup olmadığını kontrol etmekte eğer yoksa Pubmed web servislerine sorguyu göndermektedir. Böylelikle yinelenen sorgular için tekrar tekrar Pubmed web servisleri meşgul edilmemekte ve işlem yükü azaltılmaktadır.

6. SONUÇ VE ÖNERİLER

Son 10 yılda modern teknolojilerin getirdiği avantajlarla birlikte geniş çapta biyolojik verilerin üretilmesinde büyük bir artış yaşanmış ve bununla birlikte medikal literatürde hızlı bir artış görülmüştür. Akademik bilginin bu kadar zenginleşmesi, araştırmacıların bilimsel çalışmalarında ve sağlık bakım uzmanlarının günlük rutinlerinde büyük bir öneme sahiptir. Kanıt temelli hasta bakım sürecinin sağlanabilmesi amacıyla sağlık bakım uzmanları internet tabanlı veri tabanlarını, özellikle Pubmed'i, güncel bilgilere erişmek için sıklıkla kullanmaktadır. Fakat yüksek hacimlerde olan ve hızla büyüyen bilginin toplanması ve analiz edilmesi giderek zorlaşmaktadır. Bu sebeple araştırmacılara makaleleri otomatik olarak analiz eden, farklı ara yüzlerle sonuçları gösteren ve farklı metotlar kullanılarak en ilgili makaleleri bulmasını sağlayan sistemlerin geliştirilmesi gerekmektedir. Buradan yola çıkarak bu çalışmada, bir gerçek zamanlı web tabanlı literatür madenciliği sistemi geliştirilmiştir.

Bu çalışmanın en önemli katkılarından biri, Pubmed'den erişilen özetleri analiz ederek içerisindeki önemli unsurlarla birlikte hem Pubmed benzeri bir arayüzle hem de tablo formatında kullanıcılara sunmasıdır. Böylelikle araştırmacılar literatürde yer alan özetleri ayrı ayrı gözden geçirmek yerine daha açık ve kolay okunabilir bir şekilde görebilmekte ve yorumlayabilmektedir. Ayrıca özetlerden amaç cümlelerinin ve kullanılan istatistiksel terimlerin çıkartılması özelliği ile araştırmacılara çalışmalarının tasarım ve analiz aşamasında yardımcı olunacağı düşünülmektedir. Sistemin diğer önemli bir özelliği ise çeşitli sınıflardaki varlıklar arasındaki ilişkileri bularak kullanıcılara sunmasıdır. Böylelikle sağlık bakım uzmanlarının veya araştırmacıların bilinen veya daha önceden bilinmeyen ilişkileri görerek bu bilgileri günlük rutinlerinde ve araştırmalarında kullanabilecekleri veya mevcut literatür hakkında bilgi vererek yeni araştırma konuları bulmalarında yardımcı olabileceği düşünülmektedir.

Özetle, geliştirilen sistem ile sağlık bakım uzmanlarının istedikleri kanıta hızlı bir şekilde erişimlerinin sağlanması ve makalelerin gözden geçirilmesi için ayrılan sürenin en aza indirgenmesi beklenmektedir. İlerleyen aşamalarda gelişmiş analiz yöntemleri kullanılarak veri madenciliği çalışmalarında kullanılabilecek veriyi

sağlayan ve hatta bazı veri madenciliği yöntemlerini sisteme entegre ederek otomatik olarak analiz yapan modüllerin tasarlanması planlanmaktadır. Ayrıca kural tabanlı veya makine öğrenmesi gibi yöntemler kullanılarak özetlerin bulgular bölümündeki istatistiksel sonuçları yorumlayan, önemli faktörleri belirleyerek kullanıcılara sunan öğelerin sisteme eklenmesi hedeflenmektedir. Bunun yanı sıra sistemin kullanıcı değerlendirilmesinin ve ihtiyaç analizlerinin yapılarak kullanıcıların ihtiyaçları ve görüşleri doğrultusunda yeni özelliklerin, ara yüz tasarımlarının veya yeni yöntemlerin sisteme eklenmesi planlanmaktadır.

KAYNAKLAR

Alicante A,Corazza A,Isgro F,Silvestri S. Unsupervised entity and relation extraction from clinical records in Italian. Comput Biol Med. 2016 May 1;72:263-275.

Alvaro N,Conway M,Doan S,Lofi C,Overington J,Collier N. Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use. J Biomed Inform. 2015 Dec;58:280-287.

Ananiadou S,Sullivan D,Black W,Levow GA,Gillespie JJ,Mao C,Pyysalo S,Kolluru B,Tsujii J,Sobral B. Named entity recognition for bacterial Type IV secretion systems. PLoS One. 2011;6(3):e14780.

Andrade C. How to write a good abstract for a scientific paper or conference presentation. Indian J Psychiatry. 2011 Apr;53(2):172-175.

Ao H,Takagi T. ALICE: an algorithm to extract abbreviations from MEDLINE. J Am Med Inform Assoc. 2005 Sep-Oct;12(5):576-586.

Arighi CN,Siu AY,Tudor CO,Nchoutmboube JA,Wu CH,Shanker VK. eFIP: a tool for mining functional impact of phosphorylation from literature. Methods Mol Biol. 2011;694:63-75.

B. Rosario MAH. Classifying semantic relations in bioscience texts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04). Barcelona, Spain; 2002.

Barbosa-Silva A,Soldatos TG,Magalhaes IL,Pavlopoulos GA,Fontaine JF,Andrade-Navarro MA,Schneider R,Ortega JM. LAITOR--Literature Assistant for Identification of Terms co-Occurrences and Relationships. BMC Bioinformatics. 2010;11:70.

Bartsch A,Bunk B,Haddad I,Klein J,Munch R,Johl T,Karst U,Jansch L,Jahn D,Retter I. GeneReporter--sequence-based document retrieval and annotation. Bioinformatics. 2011 Apr 1;27(7):1034-1035.

Bellafqira R,Coatrieux G,Bouslimi D,Quellec G. Content-based image retrieval in homomorphic encryption domain. Conf Proc IEEE Eng Med Biol Soc. 2015 Aug;2015:2944-2947.

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2003;32(1):267-270.

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-270.

Bui DD,Jonnalagadda S,Del Fiol G. Automatically finding relevant citations for clinical guideline development. J Biomed Inform. 2015 Sep 10.

Bui DD,Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc. 2014 Sep-Oct;21(5):850-857.

Chen CC,Ho CL. PubstractHelper: A Web-based Text-Mining Tool for Marking Sentences in Abstracts from PubMed Using Multiple User-Defined Keywords. Bioinformation. 2014;10(11):708-710.

Chen P,Hinote D,Chen G. A rule based solution to co-reference resolution in clinical text. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):891-897.

Cheng D,Knox C,Young N,Stothard P,Damaraju S,Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W399-405.

Chowdhury GG. Introduction to Modern Information Retrieval. Third ed: ALA Neal-Schuman; 2010.

Chowdhury MF,Zweigenbaum P. A controlled greedy supervised approach for co-reference resolution on clinical text. J Biomed Inform. 2013 Jun;46(3):506-515.

Chung G. Sentence retrieval for abstracts of randomized controlled trials. BMC Medical Informatics and Decision Making. 2009;9(10).

Cock PJ,Antao T,Chang JT,Chapman BA,Cox CJ,Dalke A,Friedberg I,Hamelryck T,Kauff F,Wilczynski B,de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422-1423.

Cohen AM. Using symbolic network logical analysis as a knowledge extraction method on Medline abstracts. Scholar Archive. 2004; Paper 3031.

Cohen AM,Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005 Mar;6(1):57-71.

Cowie J,Lehnert W. Information extraction. Commun ACM. 1996;39(1):80-91.

Crowley RS,Tseytlin E,Jukic D. ReportTutor - an intelligent tutoring system that uses a natural language interface. AMIA Annu Symp Proc. 2005:171-175.

Cui B-J,Lin H-F,Zhang X. Research of protein-protein interaction extraction based on semi-supervised learning. Journal of Shandong University: Engineering Science. 2009;39(3):16–21.

Cui W,Liu S,Tan L,Shi C,Song Y,Gao ZJ,Tong X,Qu H. TextFlow: towards better understanding of evolving topics in text. IEEE Trans Vis Comput Graph. 2011 Dec;17(12):2412-2421.

Dai HJ,Chang YC,Tsai RT,Hsu WL. Integration of gene normalization stages and co-reference resolution using a Markov logic network. Bioinformatics. 2011 Sep 15;27(18):2586-2594.

Dai HJ,Wu JC,Tsai RT,Pan WH,Hsu WL. T-HOD: a literature-based candidate gene database for hypertension, obesity and diabetes. Database (Oxford). 2013;2013:bas061.

Dai J,Liu X. Approach for text classification based on the similarity measurement between normal cloud models. ScientificWorldJournal. 2014;2014:784392.

Davidian M,Louis TA. Why statistics? Science. 2012 Apr 6;336(6077):12.

De Comité F,Gilleron R,Tommasi M. Learning multi-label alternating decision trees from texts and data. Machine Learning and Data Mining in Pattern Recognition. Berlin, Germany: Springer; 2003. p. 35-49.

Doan S,Maehara CK,Chaparro JD,Lu S,Liu R,Graham A,Berry E,Hsu CN,Kanegaye JT,Lloyd DD,Ohno-Machado L,Burns JC,Tremoulet AH. Building a Natural Language Processing Tool to Identify Patients with High Clinical Suspicion for Kawasaki Disease from Emergency Department Notes. Acad Emerg Med. 2016 Jan 30.

Doms A,Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W783-786.

Douglas SM,Montelione GT,Gerstein M. PubNet: a flexible system for visualizing literature derived networks. Genome Biol. 2005;6(9):R80.

Drolet BC,Lorenzi NM. Registries and evidence-based medicine in craniofacial and plastic surgery. J Craniofac Surg. 2012 Jan;23(1):301-303.

Dubey A,Keller F,Sturt P. Probabilistic modeling of discourse-aware sentence processing. Top Cogn Sci. 2013 Jul;5(3):425-451.

Eaton AD. HubMed: a web-based biomedical literature search interface. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W745-747.

E-BMW Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992 Nov 4;268(17):2420-2425.

Egorov S,Yuryev A,Daraselia N. A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. J Am Med Inform Assoc. 2004 May-Jun;11(3):174-178.

Eichler K,Hemsen H,Neumann G. Unsupervised Relation Extraction From Web Documents. International Conference on Language Resources and Evaluation; 2008. p. 1674-1679.

Erhardt RA,Schneider R,Blaschke C. Status of text-mining techniques applied to biomedical text. Drug Discov Today. 2006 Apr;11(7-8):315-325.

Errami M,Wren JD,Hicks JM,Garner HR. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W12-15.

Everitt B. Medical Statistics from A to Z. 2 ed. Newyork: Cambridge University Press; 2006.

Feldman R,Regev Y,Finkelstein-Landau M,Hurvitz E,Kogan B. Mining biomedical literature using information extraction. Current Drug Discovery. 2002;2(10):19-23.

Feldman R,Regev Y,Gorodetsky M. A modular information extraction system. Intelligent Data Analysis. 2008;12(1):51-71.

Feldman R,Sanger J. The text mining handbook: advanced approaches in analyzing unstructured data: Cambridge University Press; 2007.

Fleuren WW,Verhoeven S,Frijters R,Heupers B,Polman J,van Schaik R,de Vlieg J,Alkema W. CoPub update: CoPub 5.0 a text mining system to answer biological questions. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W450-454.

Fontaine JF,Barbosa-Silva A,Schaefer M,Huska MR,Muro EM,Andrade-Navarro MA. MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W141-146.

Fontelo P,Liu F,Ackerman M. askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. BMC Med Inform Decis Mak. 2005;5:5.

Friedman C,Alderson PO,Austin JH,Cimino JJ,Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994 Mar-Apr;1(2):161-174.

Frijters R,Heupers B,van Beek P,Bouwhuis M,van Schaik R,de Vlieg J,Polman J,Alkema W. CoPub: a literature-based keyword enrichment tool for microarray data analysis. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W406-410.

Frijters R,van Vugt M,Smeets R,van Schaik R,de Vlieg J,Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol. 2010;6(9).

Frisch M,Klocke B,Haltmeier M,Frech K. LitInspector: literature and signal transduction pathway mining in PubMed abstracts. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W135-140.

Fukuda K,Tamura A,Tsunoda T,Takagi T. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput. 1998:707-718.

Gambrill E. Evidence-based clinical practice, [corrected] evidence-based medicine and the Cochrane collaboration. J Behav Ther Exp Psychiatry. 1999 Mar;30(1):1-14.

Garla V,Taylor C,Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform. 2013 Oct;46(5):869-875.

Giannakopoulos T. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLoS One. 2015;10(12):e0144610.

Giglia E. Quertle and KNALIJ: searching PubMed has never been so easy and effective. Eur J Phys Rehabil Med. 2011 Dec;47(4):687-690.

Griffon N,Charlet J,Darmoni SJ. Managing free text for secondary use of health data. Yearb Med Inform. 2014;9:167-169.

Gundavaram S. CGI Programming on the World Wide Web. First ed: O'Reilly Media; 1996.

Guo L,Liu Y,Ding Z,Sun W,Yuan M. Signal transduction by M3 muscarinic acetylcholine receptor in prostate cancer. Oncol Lett. 2016 Jan;11(1):385-392.

Güven A. Türkçe Belgelerin Anlam Tabanlı Yöntemlerle Madenciliği [Doktora Tezi]. [İstanbul]; 2007.

Han D,Wang S,Jiang C,Jiang X,Kim HE,Sun J,Ohno-Machado L. Trends in biomedical informatics: automated topic analysis of JAMIA articles. J Am Med Inform Assoc. 2015 Nov;22(6):1153-1163.

He M,Wang Y,Li W. PPI finder: a mining tool for human protein-protein interactions. PLoS One. 2009;4(2):e4554.

He T,Xu C,Li J. Named entity relation extraction method based on seed self-expansion. Computer Engineering. 2006;32(21):183–193.

He Y,Kayaalp M. Biological entity recognition with conditional random fields. AMIA Annu Symp Proc. 2008:293-297.

Henriksson A,Moen H,Skeppstedt M,Daudaravicius V,Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semantics. 2014;5(1):6.

Hotho A,Nürnberger A,Paaß G. A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology 2005;20(1):19-62.

Hristovski D,Peterlin B,Mitchell JA,Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform. 2005 Mar;74(2-4):289-298.

Hsu W,Speier W,Taira RK. Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature. AMIA Annu Symp Proc. 2012;2012:350-359.

Huang M,Zhu X,Li M. A hybrid method for relation extraction from biomedical literature. Int J Med Inform. 2006 Jun;75(6):443-455.

Hung SY,Ku YC,Chien JC. Understanding physicians' acceptance of the Medline system for practicing evidence-based medicine: a decomposed TPB model. Int J Med Inform. 2012 Feb;81(2):130-142.

Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science & Engine. 2007;9(3):90 - 95.

Hur J,Schuyler AD,States DJ,Feldman EL. SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics. 2009 Mar 15;25(6):838-840.

Jakub Piskorski RY. Information Extraction: Past, Present and Future. In: Poibeau, T., editor, Multi-source, Multilingual Information Extraction and Summarization: Springer Berlin Heidelberg; 2013. p. 23-49.

Jiang S,Pang G,Wu M,Kuang L. An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications. 2012;39(1):1503–1509.

Jonquet C,Shah NH,Musen MA. The open biomedical annotator. Summit on Translat Bioinforma. 2009;2009:56-60.

Khare R,Leaman R,Lu Z. Accessing biomedical literature in the current information landscape. Methods Mol Biol. 2014;1159:11-31.

Kim JD,Ohta T,Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics. 2008;9:10.

Kim JJ,Rebholz-Schuhmann D. Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. Brief Bioinform. 2008 Nov;9(6):452-465.

Kim SN,Martinez D,Cavedon L,Yencken L. Automatic classification of sentences to support Evidence Based Medicine. BMC Bioinformatics. 2011;12 Suppl 2:S5.

Konchady M, editor. Text Mining Application Programming. 1 ed. Boston: Charles River Media; 2006.

Krallinger M,Vazquez M,Leitner F,Salgado D,Chatr-Aryamontri A,Winter A,Perfetto L,Briganti L,Licata L,Iannuccelli M,Castagnoli L,Cesareni G,Tyers M,Schneider G,Rinaldi F,Leaman R,Gonzalez G,Matos S,Kim S,Wilbur WJ,Rocha L,Shatkay H,Tendulkar AV,Agarwal S,Liu F,Wang X,Rak R,Noto K,Elkan C,Lu Z,Dogan RI,Fontaine JF,Andrade-Navarro MA,Valencia A. The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinformatics. 2011;12 Suppl 8:S3.

Kuhlman D. A Python Book: Beginning Python, Advanced Python, and Python Exercises. Dave Kuhlman; 2009.

Kumar M,Gopal M. A comparison study on multiple binary-class SVM methods for unilabel text categorization. Pattern Recognition Letters. 2010;31(11):1437–1444.

Kumar P,Goel R,Jain C,Kumar A,Parashar A,Gond AR. An overview of biomedical literature search on the World Wide Web in the third millennium. Oral Health Dent Manag. 2012 Jun;11(2):83-89.

Lavergne T,Grouin C,Zweigenbaum P. The contribution of co-reference resolution to supervised relation detection between bacteria and biotopes entities. BMC Bioinformatics. 2015;16 Suppl 10:S6.

Lee L,Isa D,Choo W,Chue W. High Relevance Keyword Extraction facility for Bayesian text classification on different domains of varying characteristic. Expert Systems with Applications. 2012;39(1):1147–1155.

Li L,Zhang J,Jin L,Guo R,Huang D. A distributed meta-learning system for Chinese entity relation extraction. Neurocomputing. 2015;149:1135–1142.

Lindsay R,Gordon M. Literature-based discovery by lexical statistics. J Am Soc Inf Sci. 1999;50(7):574–587.

Liu H,Hu ZZ,Zhang J,Wu C. BioThesaurus: a web-based thesaurus of protein and gene names. Bioinformatics. 2006 Jan 1;22(1):103-105.

Liu Q,Song J,Li J. Using contrast patterns between true complexes and random subgraphs in PPI networks to predict unknown protein complexes. Sci Rep. 2016a;6:21223.

Liu X,Fu H,Du Z. Support Vector Machine with Ensemble Tree Kernel for Relation Extraction. Comput Intell Neurosci. 2016b;2016:8495754.

Lu Y,Zhang P,Liu J,Li J,Deng S. Health-related hot topic detection in online communities using text clustering. PLoS One. 2013;8(2):e56221.

Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford). 2011;2011:baq036.

Maki A,Evans R,Ghezzi P. Bad News: Analysis of the Quality of Information on Influenza Prevention Returned by Google in English and Italian. Front Immunol. 2015;6:616.

McDonald R,Pereira F. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics. 2005;6 Suppl 1:S6.

McKnight L,Srinivasan P. Categorization of sentence types in medical abstracts. AMIA Annu Symp Proc. 2003:440-444.

Merabti T,Lelong R,Darmoni S. InfoRoute: the CISMeF Context-specific Search Algorithm. Stud Health Technol Inform. 2015;216:544-548.

Mitsumori T,Fation S,Murata M,Doi K,Doi H. Gene/protein name recognition based on support vector machine using dictionary as features. BMC bioinformatics. 2005;6(1):8.

Miwa M,Saetre R,Miyao Y,Tsujii J. Protein-protein interaction extraction by leveraging multiple kernels and parsers. Int J Med Inform. 2009 Dec;78(12):e39-46.

Mohanty M,Ruke P,Mathew S,Kulkarni G,Alappanavar P. Unsupervised relation extraction. Int J Adv Res Comput Commun Eng. 2014;3(1):4979–4981.

Muin M,Fontelo P,Liu F,Ackerman M. SLIM: an alternative Web interface for MEDLINE/PubMed searches - a preliminary study. BMC Med Inform Decis Mak. 2005;5:37.

Murthy S. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining and Knowledge Discovery. 1998;2(4):345-389.

Myslin M,Zhu SH,Chapman W,Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res. 2013;15(8):e174.

Nguyen NT,Miwa M,Tsuruoka Y,Chikayama T,Tojo S. Wide-coverage relation extraction from MEDLINE using deep syntax. BMC Bioinformatics. 2015;16:107.

Noy NF,Shah NH,Whetzel PL,Dai B,Dorf M,Griffith N,Jonquet C,Rubin DL,Storey MA,Chute CG,Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W170-173.

Nunes T,Campos D,Matos S,Oliveira JL. BeCAS: biomedical concept recognition services and visualization. Bioinformatics. 2013 Aug 1;29(15):1915-1916.

O. Frunza DI. Extraction of disease-treatment semantic relations from biomedical sentences. Proceedings of the Workshop on Biomedical Natural Language Processing (ACL '10). Uppsala, Sweden; 2010. p. 91–98.

Oğuz B. Metin Madenciliği Teknikleri Kullanılarak Kulak Burun Boğaz Hasta Bilgi Formlarının Analizi [Yüksek Lisans]. [Antalya]: Akdeniz Üniversitesi; 2009.

Onur H. Dizinleme Amacı ile Kullanılabilecek Yöntemlerin Kıyaslanması ve Arama Sistemi Geliştirilmesi [Yüksek Lisans]. [Ankara]; 2007.

Papanikolaou N,Pafilis E,Nikolaou S,Ouzounis CA,Iliopoulos I,Promponas VJ. BioTextQuest: a web-based biomedical text mining suite for concept discovery. Bioinformatics. 2011 Dec 1;27(23):3327-3328.

Perez-Iratxeta C,Bork P,Andrade MA. Exploring MEDLINE abstracts with XplorMed. Drugs Today (Barc). 2002 Jun;38(6):381-389.

Petric I,Urbancic T,Cestnik B,Macedoni-Luksic M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. J Biomed Inform. 2009 Apr;42(2):219-227.

Plake C,Schiemann T,Pankalla M,Hakenberg J,Leser U. AliBaba: PubMed as a graph. Bioinformatics. 2006 Oct 1;22(19):2444-2445.

Plikus MV,Zhang Z,Chuong CM. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinformatics. 2006;7:424.

Quan C,Wang M,Ren F. An unsupervised text mining method for relation extraction from biomedical literature. PLoS One. 2014;9(7):e102039.

Rebholz-Schuhmann D,Kim JH,Yan Y,Dixit A,Friteyre C,Hoehndorf R,Backofen R,Lewin I. Evaluation and cross-comparison of lexical entities of biological interest (LexEBI). PLoS One. 2013;8(10):e75185.

Rebholz-Schuhmann D,Kirsch H,Arregui M,Gaudan S,Riethoven M,Stoehr P. EBIMed--text crunching to gather facts for proteins from Medline. Bioinformatics. 2007 Jan 15;23(2):e237-244.

Rindflesch TC,Tanabe L,Weinstein JN,Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput. 2000:517-528.

Rink B,Harabagiu S,Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):594-600.

Roberts A,Gaizauskas R,Hepple M,Guo Y. Mining clinical relationships from patient narratives. BMC Bioinformatics. 2008;9 Suppl 11:S3.

Rosenberg W,Donald A. Evidence based medicine: an approach to clinical problem-solving. BMJ. 1995 Apr 29;310(6987):1122-1126.

Rozenfeld B,Feldman R. Self-supervised relation extraction from the Web. Knowledge and Information Systems. 2008;17(1):17–33.

Rubinstein R,Simon I. MILANO--custom annotation of microarray results using automatic literature searches. BMC Bioinformatics. 2005;6:12.

Ruiz M,Srinivasan P. Hierarchical text categorization using neural networks. Information Retrieval. 2002;5(1):87-118.

Sackett DL,Rosenberg WM,Gray JA,Haynes RB,Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996 Jan 13;312(7023):71-72.

Sackett DL,Rosenberg WM,Gray JA,Haynes RB,Richardson WS. Evidence based medicine: what it is and what it isn't. 1996. Clin Orthop Relat Res. 2007 Feb;455:3-5.

Sarker A,Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform. 2015 Feb;53:196-207.

Sarker A,Nikfarjam A,Gonzalez G. Social Media Mining Shared Task Workshop. Pac Symp Biocomput. 2016;21:581-592.

Savova GK,Fan J,Ye Z,Murphy SP,Zheng J,Chute CG,Kullo IJ. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA Annu Symp Proc. 2010a;2010:722-726.

Savova GK,Masanz JJ,Ogren PV,Zheng J,Sohn S,Kipper-Schuler KC,Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010b Sep-Oct;17(5):507-513.

Schardt C,Adams MB,Owens T,Keitz S,Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak. 2007;7:16.

Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys. 2002;34:1-47.

Segura-Bedmar I,Martinez P,Herrero-Zazo M. Lessons learnt from the DDIExtraction-2013 Shared Task. J Biomed Inform. 2014 Oct;51:152-164.

Segura-Bedmar I,Martinez P,Revert R,Moreno-Schneider J. Exploring Spanish health social media for detecting drug effects. BMC Med Inform Decis Mak. 2015;15 Suppl 2:S6.

Smalheiser NR,Zhou W,Torvik VI. Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab. 2008;3:2.

Smith B,Williams J,Schulze-Kremer S. The ontology of the gene ontology. AMIA Annu Symp Proc. 2003:609-613.

Spasic I,Zhao B,Jones CB,Button K. KneeTex: an ontology-driven system for information extraction from MRI reports. J Biomed Semantics. 2015;6:34.

Srinivasan P,Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics. 2004 Aug 4;20 Suppl 1:i290-296.

Stanek W. HTML, JAVA, CGI, VRML, SGML-UNLEASHED. 2 ed. Yurt, B.; Ötkünç, B.; Karaçayır, A., translator. İstanbul: Sistem Yayıncılık; 2000.

States DJ,Ade AS,Wright ZC,Bookvich AV,Athey BD. MiSearch adaptive pubMed search tool. Bioinformatics. 2009 Apr 1;25(7):974-976.

Su J-S,Zhang B-F,Xu X. Advances in machine learning based text categorization. Journal of Software. 2006;17(9):1848-1859.

Sun A,Lim E-P,Liu Y. On strategies for imbalanced text classification using SVM: a comparative study. Decision Support Systems. 2009;48(1):191–201.

Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986 Autumn;30(1):7-18.

Swanson DR. Medical literature as a potential source of new knowledge. Bull Med Libr Assoc. 1990 Jan;78(1):29-37.

Tanabe L,Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002 Aug;18(8):1124-1132.

Thompson P,McNaught J,Montemagni S,Calzolari N,del Gratta R,Lee V,Marchi S,Monachini M,Pezik P,Quochi V,Rupp CJ,Sasaki Y,Venturi G,Rebholz-Schuhmann D,Ananiadou S. The BioLexicon: a large-scale terminological resource for biomedical text mining. BMC Bioinformatics. 2011;12:397.

Torii M,Li G,Li Z,Oughtred R,Diella F,Celen I,Arighi CN,Huang H,Vijay-Shanker K,Wu CH. RLIMS-P: an online text-mining tool for literature-based extraction of protein phosphorylation information. Database (Oxford). 2014;2014.

Torii M,Wagholikar K,Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):580-587.

Tsuruoka Y,Tsujii J,Ananiadou S. FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics. 2008 Nov 1;24(21):2559-2560.

Tudor CO,Schmidt CJ,Vijay-Shanker K. eGIFT: mining gene information from the literature. BMC Bioinformatics. 2010;11:418.

Wang B,Chen X,Mamitsuka H,Zhu S. BMExpert: Mining MEDLINE for Finding Experts in Biomedical Domains Based on Language Model. IEEE/ACM Trans Comput Biol Bioinform. 2015 Nov-Dec;12(6):1286-1294.

Wang J,Cetindil I,Ji S,Li C,Xie X,Li G,Feng J. Interactive and fuzzy search: a dynamic way to explore MEDLINE. Bioinformatics. 2010 Sep 15;26(18):2321-2327.

Wang T-Y,Chiang H-M. One-against-one fuzzy support vector machine classifier: an approach to text categorization. Expert Systems with Applications. 2009;36(6):10030–10034.

Weeber M, editor. Drug discovery as an example of literature-based discovery; 2007. 290–306 p.

Weeber M,Klein H,Aronson AR,Mork JG,de Jong-van den Berg LT,Vos R. Text-based discovery in biomedicine: the architecture of the DAD-system. Proc AMIA Symp. 2000:903-907.

Wei CH,Kao HY,Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W518-522.

Wishart DS,Knox C,Guo AC,Shrivastava S,Hassanali M,Stothard P,Chang Z,Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-672.

Wu X,Yang Z,Li Z,Lin H,Wang J. Disease Related Knowledge Summarization Based on Deep Graph Search. Biomed Res Int. 2015;2015:428195.

Xuan W,Dai M,Buckner J,Mirel B,Song J,Athey B,Watson SJ,Meng F. Cross-domain neurobiology data integration and exploration. BMC Genomics. 2010;11 Suppl 3:S6.

Yadav K,Sarioglu E,Choi HA, Cartwright WB,Hinds PS,Chamberlain JM. Automated Outcome Classification of Computed Tomography Imaging Reports for Pediatric Traumatic Brain Injury. Acad Emerg Med. 2016 Jan 14.

Yu B,Xu Z-B,Li C-H. Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems. 2008;21(8):900-904.

Yu H,Hatzivassiloglou V,Friedman C,Rzhetsky A,Wilbur WJ. Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. Proc AMIA Symp. 2002:919-923.

Yu H,Kim T,Oh J,Ko I,Kim S,Han WS. Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. BMC Bioinformatics. 2010;11 Suppl 2:S6.

Zhang C,Xu W,Ma Z,Gao S,Li Q,Guo J. Construction of semantic bootstrapping models for relation extraction. Knowledge-Based Systems. 2015;83:128–137.

Zhang J,Shen D,Zhou G,Su J,Tan CL. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. J Biomed Inform. 2004 Dec;37(6):411-422.

Zhang S,Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inform. 2013 Dec;46(6):1088-1098.

Zhao B,Xu S,Lin S,Luo X,Duan L. A new visual navigation system for exploring biomedical Open Educational Resource (OER) videos. J Am Med Inform Assoc. 2015 Sep 2.

Zhao M,Wang KJ,Tan Z,Zheng CM,Liang Z,Zhao JQ. Identification of potential therapeutic targets for papillary thyroid carcinoma by bioinformatics analysis. Oncol Lett. 2016 Jan;11(1):51-58.

Zheng J,Chapman WW,Crowley RS,Savova GK. Coreference resolution: a review of general methodologies and applications in the clinical domain. J Biomed Inform. 2011 Dec;44(6):1113-1122.

Zheng W,Blake C. Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles. J Biomed Inform. 2015 Oct;57:134-144.

Zhou D,Zhong D. A semi-supervised learning framework for biomedical event extraction based on hidden topics. Artif Intell Med. 2015 May;64(1):51-58.

Amaçlarda Sık Kullanılan Kelimeler

Anahtar kelime 1	"the purpose of study", "the aim of study", "the purpose of research", "the aim of research", "the purpose of article", "the aim of article", "the purpose of work", "the aim of work", "the purpose of review", "the aim of review", "the purpose of author", "the aim of author", "the purpose of paper", "the aim of paper", "the objective of study", "the objective of work", "the objective of review", "the objective of research", "the objective of article", "the objective of author", "the objective of paper", "here we carry on", "here we carry out", "aim to carry on", "aim to carry out", "study we carry on", "study we carry out", "work we carry on", "work we carry out", "research we carry on", "research we carry out", "paper we carry on", "paper we carry out", "article we carry on", "article we carry out", "review we carry on", "review we carry out"
Anahtar kelime 2	"here we address", "here we aim", "here we analyse", "here we analyze", "here we answer", "here we assess", "here we characterize", "here we collect", "here we compare", "here we conduct", "here we construct", "here we demonstrate", "here we describe", "here we determine", "here we develop", "here we discuss", "here we employ", "here we establish", "here we estimate", "here we evaluate", "here we examine", "here we explore", "here we find", "here we focus", "here we found", "here we hypothesize", "here we identify", "here we illustrate", "here we introduce", "here we incoporate", "here we investigate", "here we observe", "here we perform", "here we present", "here we propose", "here we provide", "here we purpose", "here we report", "here we review", "here we show", "here we studied", "here we summarize", "here we undertake", "here we undertook", "here we utilize"
Anahtar kelime 3	"aim to address", "aim to analyse", "aim to analyze", "aim to answer", "aim to assess", "aim to characterize", "aim to collect", "aim to compare", "aim to conduct", "aim to construct", "aim to demonstrate", "aim to describe", "aim to determine", "aim to develop", "aim to discuss", "aim to employ", "aim to establish", "aim to estimate", "aim to evaluate", "aim to examine", "aim to explore", "aim to find", "aim to focus", "aim to found", "aim to hypothesize", "aim to identify", "aim to illustrate", "aim to introduce", "aim to incoporate", "aim to investigate", "aim to observe", "aim to perform", "aim to present", "aim to propose", "aim to provide", "aim to purpose", "aim to report", "aim to review", "aim to show", "aim to studied", "aim to summarize", "aim to undertake", "aim to undertook", "aim to utilize", "aim to study"

Anahtar kelime 4	"study we address", "study we aim", "study we analyse", "study we analyze", "study we answer", "study we assess", "study we characterize", "study we collect", "study we compare", "study we conduct", "study we construct", "study we demonstrate", "study we describe", "study we determine", "study we develop", "study we discuss", "study we employ", "study we establish", "study we estimate", "study we evaluate", "study we examine", "study we explore", "study we find", "study we focus", "study we found", "study we hypothesize", "study we identify", "study we illustrate", "study we introduce", "study we investigate", "study we observe", "study we perform", "study we present", "study we propose", "study we provide", "study we purpose", "study we report", "study we review", "study we show", "study we studied", "study we summarize", "study we undertake", "study we undertook", "study we utilize"
Anahtar kelime 5	"work we address", "work we aim", "work we analyse", "work we analyze", "work we answer", "work we assess", "work we characterize", "work we collect", "work we compare", "work we conduct", "work we construct", "work we demonstrate", "work we describe", "work we determine", "work we develop", "work we discuss", "work we employ", "work we establish", "work we estimate", "work we evaluate", "work we examine", "work we explore", "work we find", "work we focus", "work we found", "work we hypothesize", "work we identify", "work we illustrate", "work we study", "paper we study", "article we study", "review we study", "work we introduce", "work we incoporate", "work we investigate", "work we observe", "work we perform", "work we present", "work we propose", "work we provide", "work we purpose", "work we report", "work we review", "work we show", "work we studied", "work we summarize", "work we undertake", "work we undertook", "work we utilize", "work we study"
Anahtar kelime 6	"research we address", "research we aim", "research we analyse", "research we analyze", "research we answer", "research we assess", "research we characterize", "research we collect", "research we compare", "research we conduct", "research we construct", "research we demonstrate", "research we describe", "research we determine", "research we develop", "research we discuss", "research we employ", "research we establish", "research we estimate", "research we evaluate", "research we examine", "research we explore", "research we find", "research we focus", "research we found", "research we hypothesize", "research we identify", "research we illustrate", "research we introduce", "research we investigate", "research we observe", "research we perform", "research we present", "research we propose", "research we provide", "research we purpose", "research we report", "research we review", "research we show", "research we studied", "research we summarize", "research we undertake", "research we undertook", "research we utilize", "research we study"

Anahtar kelime 7	"paper we address", "paper we aim", "paper we analyse", "paper we analyze", "paper we answer", "paper we assess", "paper we characterize", "paper we collect", "paper we compare", "paper we conduct", "paper we construct", "paper we demonstrate", "paper we describe", "paper we determine", "paper we develop", "paper we discuss", "paper we employ", "paper we establish", "paper we estimate", "paper we evaluate", "paper we examine", "paper we explore", "paper we find", "paper we focus", "paper we found", "paper we hypothesize", "paper we identify", "paper we illustrate", "paper we investigate", "paper we incoporate", "paper we observe", "paper we perform", "paper we present", "paper we propose", "paper we provide", "paper we purpose", "paper we report", "paper we review", "paper we show", "paper we studied", "paper we summarize", "paper we undertake", "paper we undertook", "paper we utilize"
Anahtar kelime 8	"article we address", "article we aim", "article we analyse", "article we analyze", "article we answer", "article we assess", "article we characterize", "article we collect", "work carry on", "work carry out", "research carry on", "research carry out", "paper carry on", "paper carry out", "article carry on", "article carry out", "review carry on", "review carry out", "we carry on", "we carry out", "article we compare", "article we conduct", "article we construct", "article we demonstrate", "article we describe", "article we determine", "article we develop", "article we discuss", "article we employ", "article we establish", "article we estimate", "article we evaluate", "article we examine", "article we explore", "article we find", "article we focus", "study carry on", "study carry out", "article we found", "article we hypothesize", "article we identify", "article we illustrate", "article we introduce", "article we incoporate", "article we investigate", "article we observe", "article we perform", "article we present", "article we propose", "article we provide", "article we purpose", "article we report", "article we review", "article we show", "article we studied", "article we summarize", "article we undertake", "article we undertook", "article we utilize", "article we study"
Anahtar kelime 9	"review we address", "review we aim", "review we analyse", "review we analyze", "review we answer", "review we assess", "review we characterize", "review we collect", "review we compare", "review we conduct", "review we construct", "review we demonstrate", "review we describe", "review we determine", "review we develop", "review we discuss", "review we employ", "review we establish", "review we estimate", "review we evaluate", "review we examine", "review we explore", "review we find", "review we focus", "review we found", "review we hypothesize", "review we identify", "review we illustrate", "review we introduce", "review we incoporate", "review we investigate", "review we observe", "review we perform", "review we present", "review we propose", "review we provide", "review we purpose", "review we report", "review we show", "review we studied", "review we summarize", "review we undertake", "review we undertook", "review we utilize", "review we study", "to carry on", "to carry out"

Anahtar kelime 10	"study address”, “study aim”, “study analyse”, “study analyze”, “study answer”, “study assess”, “study characterize”, “study collect”, “study compare”, “study conduct”, “study construct”, “study demonstrate”, “study describe”, “study determine”, “study develop”, “study discuss”, “study employ”, “study establish”, “study estimate”, “study evaluate”, “study examine”, “study explore”, “study find”, “study focus”, “study found”, “study hypothesize”, “study identify”, “study illustrate”, “study introduce”, “study incoporate”, “study investigate”, “study observe”, “study perform”, “study present”, “study propose”, “study provide”, “study purpose”, “study report”, “study review”, “study show”, “study studied”, “study summarize”, “study undertake”, “study undertook”, “study utilize"
Anahtar kelime 11	"work address”, “work aim”, “work analyse”, “work analyze”, “work answer”, “work assess”, “work characterize”, “work collect”, “work compare”, “work conduct”, “work construct”, “work demonstrate”, “work describe”, “work determine”, “work develop”, “work discuss”, “work employ”, “work establish”, “work estimate”, “work evaluate”, “work examine”, “work explore”, “work find”, “work focus”, “work found”, “work hypothesize”, “work identify”, “work illustrate”, “work introduce”, “work investigate”, “work observe”, “work perform”, “work present”, “work propose”, “work provide”, “work purpose”, “work report”, “work review”, “work show”, “work study”, “work studied”, “work summarize”, “work undertake”, “work undertook”, “work utilize"
Anahtar kelime 12	"research address”, “research aim”, “research analyse”, “research analyze”, “research answer”, “research assess”, “research characterize”, “research collect”, “research compare”, “research conduct”, “research construct”, “research demonstrate”, “research describe”, “research determine”, “research develop”, “research discuss”, “research employ”, “research establish”, “research estimate”, “research evaluate”, “research examine”, “research explore”, “research find”, “research focus”, “research found”, “research hypothesize”, “research identify”, “research illustrate”, “research introduce”, “research incoporate”, “research investigate”, “research observe”, “research perform”, “research present”, “research propose”, “research provide”, “research purpose”, “research report”, “research review”, “research show”, “research studied”, “research summarize”, “research undertake”, “research undertook”, “research utilize”, “research study"
Anahtar kelime 13	"paper address”, “paper aim”, “paper analyse”, “paper analyze”, “paper answer”, “paper assess”, “paper characterize”, “paper collect”, “paper incoporate”, “article incoporate”, “paper compare”, “paper conduct”, “paper construct”, “paper demonstrate”, “paper describe”, “paper determine”, “paper develop”, “paper discuss”, “paper employ”, “paper establish”, “paper estimate”, “paper evaluate”, “paper examine”, “paper explore”, “paper find”, “paper focus”, “paper found”, “paper hypothesize”, “paper identify”, “paper illustrate”, “paper introduce”, “paper investigate”, “paper study”, “paper observe”, “paper perform”, “paper present”, “paper propose”, “paper provide”, “paper purpose”, “paper report”, “paper review", "paper show”, “paper studied”, “paper summarize”, “paper undertake”, “paper undertook”, “paper utilize"

Anahtar kelime 14	"article address", “article aim”, “article analyse”, “article analyze”, “article answer”, “article assess”, “article characterize”, “article collect”, “article compare”, “article conduct”, “article construct”, “article demonstrate”, “article describe”, “article determine”, “article develop”, “article discuss”, “article employ”, “article establish”, “article estimate”, “article evaluate”, “article examine”, “article explore”, “article find”, “article focus”, “article found”, “article hypothesize”, “article identify”, “article illustrate”, “article introduce”, “article investigate”, “article observe”, “article perform”, “article present”, “article propose”, “article provide”, “article purpose”, “article report”, “article review”, “article show”, “article studied”, “article summarize”, “article undertake”, “article undertook”, “article utilize"
Anahtar kelime 15	"review address”, “review aim”, “review analyse”, “review analyze”, “review answer”, “review assess”, “review characterize”, “review collect”, “review compare”, “review conduct”, “review construct”, “review demonstrate”, “review describe”, “review determine”, “review develop”, “review discuss”, “review employ”, “review establish”, “review estimate”, “review evaluate”, “review examine”, “review explore”, “review find”, “review focus”, “review found”, “review hypothesize”, “review identify”, “review illustrate”, “review introduce”, “review incoporate”, “review investigate”, “review observe”, “review perform”, “review present”, “review propose”, “review provide”, “review purpose”, “review report”, “review review”, “review show”, “review studied”, “review summarize”, “review undertake”, “review undertook”, “review utilize”, “review study"
Anahtar kelime 16	"we studied”, “we address”, “we aim”, “we analyse”, “we analyze”, “we answer”, “we assess”, “we characterize”, “we collect”, “we compare”, “we conduct”, “we construct”, “we demonstrate”, “we describe”, “we determine”, “we develop”, “we discuss”, “we employ”, “we establish”, “we estimate”, “we evaluate”, “we examine”, “we explore”, “we find”, “we focus”, “we found”, “we hypothesize”, “we identify”, “we illustrate”, “we introduce”, “we incoporate”, “we investigate”, “we observe”, “we perform”, “we present”, “we propose”, “we provide”, “we purpose”, “we report”, “we review”, “we show”, “we summarize”, “we undertake”, “we undertook”, “we utilize”, “to address”, “to aim”, “to analyse”, “to analyze”, “to answer”, “to assess”, “to characterize”, “to collect”, “to compare”, “to conduct”, “to construct”, “to demonstrate”, “to describe”, “to determine”, “to develop”, “to discuss”, “to employ”, “to establish”, “to estimate”, “to evaluate”, “to examine”, “to explore”, “to find”, “to focus”, “to found”, “to hypothesize”, “to identify”, “to illustrate”, “to introduce”, “to investigate”, “to observe”, “to perform”, “to present”, “to propose”, “to provide”, “to purpose”, “to report”, “to review”, “to show”, “to studied”, “to summarize”, “to undertake”, “to undertook”, “to utilize”, “to incoporate", "to study"

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı	**Başak**	**Uyruğu**	**Türkiye Cumhuriyeti**
Soyadı	**Oğuz Yolcular**	**Tel no**	**-**
Doğum tarihi	**03.06.1983**	**e-posta**	**basakoguz@akdeniz.edu.tr**

Eğitim Bilgileri

Mezun olduğu kurum		**Mezuniyet yılı**
Lise	**Antalya Aldemir Atilla Konuk**	**2001**
Lisans	**İstanbul Üniversitesi-İngilizce İktisat**	**2006**
Yüksek Lisans	**Akdeniz Üniversitesi-Tıp Bilişimi**	**2009**
Doktora	**Akdeniz Üniversitesi-Tıp Bilişimi**	**2016**

İş Deneyimi

Görevi	**Kurum**	**Süre (yıl-yıl)**
Araştırma Görevlisi	**Akdeniz Üniversitesi**	**10 yıl**

Yabancı Dilleri	**Sınav türü**	**Puanı**
İngilizce	**YDS**	**85**

Proje Deneyimi

Proje Adı	**Destekleyen kurum**	**Süre (Yıl-Yıl)**

Yayınlar ve Bildiriler:

Uluslararası Yayınlar:

1. Yuce YK, Zayim N, Oguz B, Bozkurt S, Isleyen F, Gulkesen KH. Analysis of social networks among physicians employed at a medical school. Stud Health Technol Inform. 2014;205:543-7.
2. Bilge U, Bozkurt S, Yolcular BO, Ozel D. Can social web help to detect influenza related illnesses in Turkey. Stud Health Technol Inform. 2012;174:100-4.
3. Cekin Y, Cekin AH, Duman A, Yilmaz U, Yesil B, Yolcular BO. The role of serum procalcitonin levels in predicting ascitic fluid infection in hospitalized cirrhotic and non-cirrhotic patients. Int J Med Sci. 2013 Aug 20;10(10):1367-74.
4. Cekin AH, Cekin Y, Adakan Y, Tasdemir E, Koclar FG, Yolcular BO. Blastocystosis in patients with gastrointestinal symptoms: a case-control study. BMC Gastroenterol. 2012 Sep 10;12:122.
5. Cekin AH, Taskoparan M, Duman A, Sezer C, Cekin Y, Yolcular BO, Can H, Pehlivan FS, Cayirci M. The Role of Helicobacter pylori and NSAIDs in the Pathogenesis of Uncomplicated Duodenal Ulcer. Gastroenterol Res Pract. 2012;2012:189373.
6. Cekin AH, Ungor D, Duman A, Akbay FH, Taskin V, Yilmaz U, Yolcular BO, Adakan FY, Sahinturk Y, Yesil B. Evaluation of renal toxicity in chronic hepatitis B patients treated with tenofovir. HealthMed. 2014;8(6):728.

Uluslararası Kongre Bildirileri:

1. Oğuz B, Bilge B, Samur MK. Transformıng Unstructured Otolaryngology Dıscharge Notes Into A Structured Form. Informatica'09, Cuba 9-13 February 2009.
2. Oğuz B, Zayim N, Saka O. Internet Addiction Among First-year Medical Students. Hibit'10.
3. Oğuz B, Bilge B, Samur MK. Association Rules Extraction from the Otolaryngology Discharge Notes. COMPSTAT 2010 Paris France August 22-27, 2010.
4. Isleyen F, Gulkesen KH, Yuce YK, Samur AA, Oguz B. Predicting Experts Quality Ratings of Dermatologic Images Using Bayesian Network. Waset 2011 Venice İtaly 28-30 November 2011.
5. Yolcular BO. Concepts Extraction from Discharge Notes Using Association Rule Mining. Waset 2011 Venice İtaly 28-30 November 2011.
6. Yolcular BO. Concepts Extraction from Discharge Notes Using Association Rule Mining. The International Symposium on Health Informatics and Bioinformatics (HIBIT 2012) 19-22 April 2012 Kapadokya, Nevşehir.
7. Isleyen F, Gulkesen KH, Samur AA, Yuce YK, Oguz B. Comparison of Experts Quality Ratings and Objective Quality Measures Using Compressed

Dermatologic Images. The International Symposium on Health Informatics and Bioinformatics (HIBIT 2012) 19-22 April 2012 Kapadokya, Nevşehir.

8. Bilge U, Bozkurt S, Yolcular BO, Özel D. Can Social Web Help To Detect Influenza Related Illnesses in Turkey? Conference of the European-Federation-for-Medical-Informatics (EFMI), Moscow, RUSSIA, 18-20 April 2012.
9. Yüce YK, Zayim N, Oguz B, Bozkurt S, İşleyen F, Samur AA, Gülkesen KH. Social Networks among Physicians Employed at a Medical School: Protocol of a Study. MSN 2012.
10. Yuce YK, Zayim N, Oguz B, Bozkurt S, Isleyen F, Gulkesen KH. Analysis of social networks among physicians employed at a medical school. The 25th European Medical Informatics Conference - MIE2014 31 August-3 September İstanbul Turkey.

Ulusal Kongre Bildirileri:

1. Oğuz B, Gülkesen KH, Saka O. Sağlık Bilgi Sistemlerinde Maliyet-Fayda Analizi. Akademik Bilişim 2007 Dumlupınar Üniversitesi, Kütahya 31 Ocak-2 Şubat 2007.
2. Oğuz B, Bilge U, Saka O. Tıpta Metin Madenciliği. Tıp Bilişimi ’07, Antalya 15-18 Kasım 2007.
3. Özel D, Zayim N, Oğuz B, Döşemeci L, SAKA O. Yoğun Bakımda Mobilite: Uygulama Öncesi Değerlendirme. Tıp Bilişimi ’07, Antalya 15-18 Kasım 2007.
4. Oğuz B, Bilge U, Samur MK. Yapılandırılmamış Kulak Burun Boğaz Epikriz Notlarının Yapılandırılmış Formata Dönüştürülmesi. Tıp Bilişimi ’08 Antalya 13-16 Kasım 2008.
5. Özel D, Zayim N, Oğuz B, Döşemeci L, Saka O. Yoğun Bakımda Mobilite: Pilot Uygulama Sonrası Görüşler. Tıp Bilişimi ’08 Antalya 13-16 Kasım 2008.
6. Özel D, Zayim N, Oğuz B, Saka O. Tıp Öğrencilerinin İnternette Bilişsel Durumları. Akademik Bilişim 2008 Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 30 Ocak - 01 Şubat 2008.
7. Oğuz B, Bilge U, Samur MK. Kulak Burun Boğaz Epikriz Notlarından Birliktelik Kurallarının Çıkartılması. TurkMIA ‘0912-15 Kasım 2009, Antalya.
8. Oğuz B. Koroner Arter Hastalığı ile İlgili Özetlerden Anahtar Kelimelerin Çıkartılması: Bir Bilgi Çıkarım Uygulaması. VII. Ulusal Tıp Bilişimi Kongresi 14-17 Ekim 2010 Magosa/KKTC.
9. Bilge U, Bozkurt S, Yolcular BO, Özel D. Sosyal medya araçları Türkiye’deki grip benzeri hastalıkları saptayabilmek için kullanılabilir mi? İnet 2011.

10. Isleyen F, Gulkesen KH, Yuce YK, Samur AA, Oguz B. Dermatoloji Görüntülerinin Kalitesinin Objektif Ölçütlerle Değerlendirilmesi. IX. Ulusal Tıp Bilişimi Kongresi 15-18 Kasım 2012 Belek, Antalya.
11. Yüce YK, Zayim N, Oğuz B, Bozkurt S, İşleyen F, Samur AA, Gülkesen KH. Bir Tıp Fakültesinde Hekim Fikir Liderlerinin ve Hekimler Arasındaki Sosyal Ağın Tespiti: Çalışma Protokolü. 8. Ulusal Tıp Bilişimi Kongresi 17-20 Kasım 2011 Belek, Antalya.
12. Yolcular BO, Bozkurt S, Bilge U, Özel D. Sosyal Medya Araçları Türkiye'deki Grip Benzeri Hastalıkları Saptayabilmek için kullanılabilir mi? 8. Ulusal Tıp Bilişimi Kongresi 17-20 Kasım 2011 Belek, Antalya.
13. Yolcular BO, Zayim N. Biyomedikal Literatür Taraması için Geliştirilmiş Web-Tabanlı Araçların Karşılaştırılması. IX. Ulusal Tıp Bilişimi Kongresi 15-17 Kasım 2012 Belek, Antalya.
14. Önder M, Aldemir C, Yolcular BO, Oğuz N. Lunat Fossa'nın Morfometrik Değerlendirilmesi. 16. Ulusal Anatomi Kongresi 11-14 Eylül 2014 Malatya, Türkiye.
15. Cekin AH, Gömceli İ, Orman NS, Yolcular BO, Uyar S, Şimşek M, Özcan F, Gündoğdu M. "Nutrisyon Desteği Verilen Hastalarda Hedef Kaloriye ulaşma Israrı Ne Getirir?", KEPAN 2015, Antalya, Türkiye, 18-22 Mart 2015.
16. Cekin AH, Gömceli İ, Orman NS, Yolcular BO, Uyar S, Şimşek M, Özcan F, Gündoğdu M. "Nutrisyon desteği verilen hastalarda verilmesi amaçlanan kalori hedefine ulaşmada karşılaşılan sorunlar", KEPAN 2015, Antalya, Türkiye, 18-22 Mart 2015.