ESTUDAM Bilişim Dergisi
Cilt 2, Sayı 1 , 11-14, 2021

Journal of ESTUDAM Information
Volume 2, Issue 1, 11-14, 2021

(Geliş Tarihi / Received Date: 30.12.2020, Kabul Tarihi/ Accepted Date: 10.01.21)

Wikipedia'daki Verilere Metin Madenciliği Yöntemlerinin Uygulanması

Ayça Nur KAHYA[*1]
[1]Eskişehir Osmangazi Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, 26480, Eskişehir

Anahtar Kelimeler:
Metin Madenciliği,
Veri Madenciliği,
Veri Görselleştirme,
Wikipedia,
Python,
JupyterLab

Özet: Wikipedia, internetin en geniş kaynak sitelerinden biridir. Herkesin katkıda bulunabildiği ücretsiz bir ansiklopedidir. Wikipedia çok geniş bilgi kaynağı oluşturması sebebiyle birçok insan tarafından tercih edilmektedir. Gün geçtikçe bu veri kaynağı büyümektedir. Ücretsiz, reklamsız ve birçok dil desteği bulunması Wikipedia'nın tercih edilme sebepleri arasındadır. Bu çalışmanın amacı metin madenciliği ile Wikipedia verilerinin işlenmesi ve analiz edilmesidir. Wikipedia üzerinden örnek alınan verilere metin madenciliği yöntemleri uygulanarak daha işlevsel hale getirilmiştir.

Applying Text Mining Methods to Wikipedia Data

Keywords:
Text Mining,
Data Mining,
Data Virtualization,
Wikipedia,
Python,
JupyterLab

Abstract: Wikipedia is one of the internet's largest resource sites. It is a free encyclopedia that anyone can contribute. Wikipedia is preferred by many people due to its wide range of information resources. This data source is growing day by day. It is among the reasons why Wikipedia is preferred because it is free, does not contain advertising and has many language support. The aim of this study is to process and analyze the data of Wikipedia with text mining. Text mining methods have been applied to the data sampled on Wikipedia, making it more functional.

1. GİRİŞ

Wikipedia, kullanıcıları tarafından ortaklaşa olarak birçok dilde hazırlanan, özgür, bağımsız, ücretsiz, reklamsız, kâr amacı gütmeyen bir internet ansiklopedisidir. Kurucularından, Jimmy Wales Wikipedia'yı, "Dünya üzerindeki her insana kendi dilinde, en üst kalitede, bedava bir ansiklopedi oluşturma ve dağıtma uğraşısı" olarak tanımlamaktadır. Wikipedia, gönüllülerin ortaklaşa çabası doğrultusunda ve neredeyse herkesin web sitesine ulaşıp değiştirmesiyle yazılmaktadır. Wikipedia'ya bilgi girişi yapabilmek için üye olma zorunluluğu yoktur. Bu durum bilgilerin herkes tarafından anında eklenebilmesine olanak verir [1].

Metin madenciliği (Text Mining), yapısal olmayan ve düzensiz haldeki elektronik metin yığınlarından; önceden bilinmeyen, potansiyel olarak kullanışlı, yapısal ve düzenli veri elde etme sürecidir. Elde edilen bilgiyle, analiz edilen metin kaynaklarında açık olarak görülmeyen ilişkiler, hipotezler ve eğilimler tespit edilir. Metin madenciliği çalışmaları, metin kaynaklı literatürdeki diğer bir çalışma alanı olan doğal dil işleme (Natural Language Processing / NLP) çalışmaları ile çoğu zaman beraber ele alınmaktadır. Doğal dil işleme çalışmaları daha çok yapay zeka altındaki dil bilim bilgisine dayalı çalışmalarını kapsamaktadır. Metin madenciliği çalışmaları ise daha çok istatistiksel olarak metin üzerinden sonuçlara

*İlgili yazar: Ayça Nur KAHYA, yilmazayca26@gmail.com

ulaşmayı hedefler [2]. Metin madenciliği çalışmaları birçok konuda yapılmaktadır. Twitter, Facebook, YouTube, LinkedIn, haber siteleri, hava durum siteleri gibi birçok örnek verilebilir. Bu veriler toplanarak analizlerden geçirilir. Analizler ile kullanıcıların birçok durum ve bilgileri tespit edilir. Sosyal medyada yaygın olarak tartışılan, kutuplaştırıcı bir konu haline gelen Suriye mülteci krizine yönelik kamuoyunun görüş ve duyguları araştırılmıştır. Twitter'da konuyla ilgili kamuoyu görüşlerini analiz etmek için Türkçe ve İngilizce olmak üzere iki dilde toplam 2.381.297 ilgili tweet toplanmıştır [3]. Diğer örnek ise YouTube içeriği ile ilgilidir. Çeşitli alanlardaki ilgi ve farkındalıklarla ilgili mevcut eğilimleri anlamak için, hemşirelikte trend kelimeleri keşfetmek için YouTube içeriğini kullanılmıştır. İlgili YouTube içerikleri toplanarak analiz işlemleri gerçekleştirilmiştir. En sık kullanılan anahtar sözcükler olarak "alan", "iş", "yeterlilik", "fikir" ve "başarı" yı tanımlamıştır [4]. Facebook'ta dünya çapında en popüler on altı haber kanalı sitesinden yaklaşık 37551 gönderi metin madenciliği işlemleri tarafından toplanmış ve gerekli işlemlerden geçirildikten sonra analiz edilmiştir. Sonuçlar, toplanan verilerin genel olarak 3 ana başlıkta yoğunlaştığını göstermiştir: "Rio de Janeiro", "ABD seçimleri" ve "İngiltere, Avrupa Birliği'nden ayrılıyor". Bu üç ana konu, tüm haber kanallarında tartışılan gündemdeki konular olarak kabul edilmektedir [5]. Diğer bir makale, satın alınan ürünlerle ilgili tüketici geri bildirimlerini anlamak için metin madenciliği uygulama yöntemini özetlemektedir. Amazon tüketici yorumları metin madenciliği yöntemleri ile toplanarak analiz edilmiştir. Tüketicilerden bu bilgilere ihtiyaç duyanlara, ürün geliştiricilere vb. raporlar halinde sunulabileceği ve daha iyi bir ürün teslimatı ile sonuçlanabileceği analizler sonucu tespit edilmiştir [6].

Bu çalışmanın amacı, Wikipedia üzerinde metin madenciliği konusunda alt yapı oluşturmak ve veri görselleştirme yöntemleri ile verileri kolay anlaşılır hale getirmektir. Wikipedia'da yer alan birçok veri için kullanılabilmektedir.

2. MATERYAL VE METOT

2.1. Projede Kullanılan Araçlar

Bu projede kullanılan araçlar;

- Wikipedia
- Python
- Anaconda Navigator
- JupyterLab

Wikipedia, kullanıcılar tarafından ortaklaşa birçok dilde hazırlanan bir internet ansiklopedisidir.

Python, nesne yönelimli, yorumlamalı, birimsel (modüler) ve etkileşimli yüksek seviyeli bir programlama dilidir. Modüler yapısı, sınıf dizgesini (sistem) ve her türlü veri alanı girişini destekler. Hemen hemen her türlü platformda çalışabilir. Python ile sistem programlama, kullanıcı arabirimi programlama, ağ programlama, web programlama, uygulama ve veritabanı yazılımı programlama gibi birçok alanda yazılım geliştirebilirsiniz [7].

Anaconda Navigator, Anaconda dağıtımında bulunan ve kullanıcıların komut satırı komutlarını kullanmadan uygulamaları başlatmasını ve conda paketlerini, ortamları ve kanalları yönetmesini sağlayan bir masaüstü grafik kullanıcı arabirimidir. Navigator, paketleri Anaconda Cloud'da veya yerel bir Anaconda Deposunda arayabilir, bir ortama kurabilir, paketleri çalıştırabilir ve güncelleyebilir [8].

JupyterLab, veri formatlarını görüntülemek ve işlemek için birleşik bir model sunar. JupyterLab, birçok dosya formatını (resimler, CSV, JSON, Markdown, PDF, Vega, Vega-Lite, vb.) desteklemektedir. Bu formatlarda zengin çekirdek çıktılarını da görüntüler. [9].

Kullanılan kütüphaneler;

- Numpy (Numerical Python)
- Pandas
- Matplotlib
- LXML

Numpy, bilimsel hesaplamaların hızlı bir şekilde yapılmasını sağlamaktadır. Dizinler üzerinde matematiksel, cebirsel ve istatistiksel operasyonlar uygulamak için kullanılır.

Pandas, "ilişkisel" ve "etiketli" verilerle çalışmayı kolay ve sezgisel hale getirmek için tasarlanmış hızlı, esnek ve etkileyici veri yapıları sağlayan bir Python paketidir. Her dilde mevcut olan en güçlü ve esnek açık kaynak veri analizi / manipülasyon aracı olmak gibi daha geniş bir amaca sahiptir [10].

Matplotlib, grafik çizim paketi Python'la bilimsel programlamanın en önemli araçlarından birisidir. Matplotlib ile verileri etkileşimli olarak görselleştirme, yayınlamaya uygun yüksek kalitede çıktılar hazırlama işlemleri yapılmaktadır. Hem iki boyutlu hem de üç boyutlu grafikler üretilebilir [11].

LXML, Python'un kendi içerisinde halihazırda bulunan bir XML kütüphanesi mevcuttur. Ancak XML yapılarının (etiketlerin isimleri, özellikleri ve hangi etiketin hangi etiket içerisinde yer aldığı) tanımlarını bilgisayarın ana hafızasında açamaz ve bu tanımları kullanarak XML dosyaları doğrulanamaz [12].

2.2. Veri Seti

Projede kullanılan veri seti, Wikipedia'da HTML çözümleme yöntemi ile elde edilmiştir. HTML çözümleme yöntemi ile yer alan bütün tablo bilgileri arasından nobel ödülü sahipleri tablosu ayrıştırılmıştır.

2.3. Metin Ön İşleme

Veri seti üzerinde analiz ve görselleştirme işlemlerinin yapılması için ön işlemlerden geçirilmesi gerekmektedir. Bunlar; tokenizasyon, kullanılmayan kelime ve karakterleri kaldırma, kelimeleri normalleştirme, büyük ve küçük harf farklılıklarını kaldırma, kelimenin sağında ve solunda yer alan boşlukları temizleme işlemleridir.

Tokenizasyon, metnin boşluk, — , vb. istenilen özelliklere göre parçalara ayrılmasıdır. Bu parçalar üzerinden token array içerisine atılmasıdır. [13].Veri kazıma (Web Scraping) yöntemleri ile Wikipedia verilerinin ön işlenmesi gerçekleştirilmiştir. Veri kazıma diğer bir adıyla Data Scraping, web sitelerinden veri

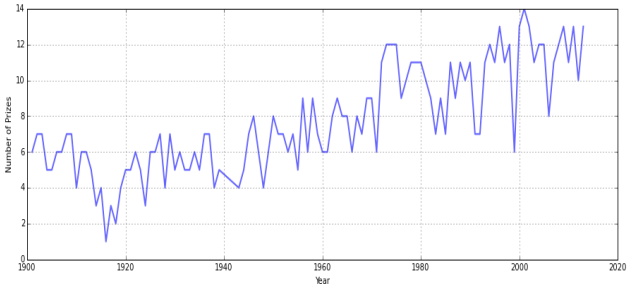
çıkarma işlemidir. Web üzerinde birçok veri bulunmaktadır. Veri kazıma'da bu veri toplama işlemini otomatikleştirmektedir. Dağıtık verileri daha düzgün şekilde sunmaktadır [14]. Veri kazıma yöntemleri ile elde edilen verilerde harf olmayan (link, hashtag vb.) birçok karakter mevcuttur. Bu karakterlerin temizlenmesi sağlanmıştır. Ayrıca trim işlemi ile kelimenin sağında ve solunda yer alan boşlukların da silinmesi sağlanmıştır.Ön işleme sonucunda konular, url bilgileri ve yıllara göre dağılımları tablo halinde görüntülenmiştir.

2.4. Veri Görselleştirme

Veri görselleştirme, karmaşık ve dağınık verileri düzenleyerek kolay anlaşılabilir, yorumlanabilir hale getirmektir. Bir başka ifade ile veri görselleştirme, soyut verileri görsel hale getirerek somutlaştırmak ve bundan bir ön bilgi elde edilmesidir [15]. Wikipedia üzerinden getirilen nobel ödülü sahipleri listesinde veri görselleştirme işlemleri uygulanmıştır. Bu şekilde sunulan bilgilerin daha kolay anlaşılması ve yorumlanması sağlanmıştır. Python veri görselleştirme kütüphanelerinden Matplotlib, Seaborn, Plotly en popüler olanlardandır. Kullanım kolaylığı ve kolay öğrenilmesi sayesinde avantaj sunduğu için Matplotlib tercih edilmiştir. Matplotlib ile yıllık ödül sayısı ve verilen nobel ödüllerinin kümülatif toplamlarının veri görselleştirme işlemi uygulanmıştır.

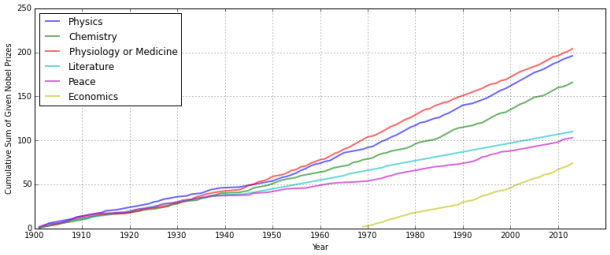
3. BULGULAR

Wikipedia üzerinden, nobel ödülü sahipleri verilerine ulaşılmıştır. Veriler ön işlemden geçerek analiz aşamasına alınmıştır. Pandas kütüphanesi ile veri analizi işleminde seriler ve dataframe'ler kullanılmıştır.



Şekil 1. Yıllara ait ödül sayısı sonucu

Şekil 1'de nobel ödül listesindeki toplam ödül bilgisinin yıllara göre dağılımı verilmiştir. Örneğin; 1901 yılında toplam ödül sayısı 6, 1914 yılında toplam ödül sayısı 3 olarak grafikte gösterimi sağlanmıştır. Şekil 1'deki grafikte 1900 ile 2013 yılları arasındaki ödül sayıları verilmiştir. 1914 ile 1918 yılları arası 1. Dünya Savaşı ve 1939 ile 1945 yılları arası 2.Dünya Savaşı nedeniyle nobel ödül sayısındaki düşüş Şekil 1'de görülmektedir.



Şekil 2. Nobel ödüllerin kümülatif toplamı

Şekil 2'de nobel ödül listesinin kategorilere göre dağılımı ve kümülatif toplamı verilmiştir. Şekilde görüldüğü gibi ekonomi alanında nobel ödül sayısı 1969 yılından itibaren artış göstermiştir. 1900 ile 2020 yılları arasında kümülatif toplam alınmıştır. Kümülatif toplam için Matplotlib kütüphanesi kullanılmıştır.

Tablo 1'de Pandas kütüphanesi ile nobel ödülü kazanan kişi, konusu ve kazandığı yıl bilgileri Wikipedia üzerinden getirilmiştir. Tablo 1'de ilk 5 sonuca yer verilmiştir.

Tablo 1. Pandas ön işleme sonucu

Winner name	Subject	Year
Wilhelm Röntgen	Physics	1901
Jacobus Henricus van 't Hoff	Chemistry	1901
Emil Adolf von Behring	Physiology or Medicine	1901
Sully Prudhomme	Literature	1901
Henry Dunant	Bilgi girdisi	1901
	Peace	1901

Tablo 2. Yıllara göre nobel ödül kazanan toplam kişi sayısı

Year	Count
1901	6
1902	7
…	…
2011	13
2012	10
2013	13
…	…
2020	12

Pandas kütüphanesi ile dataframe kullanılarak Tablo 2 oluşturulmuştur.

4. TARTIŞMA VE SONUÇ

Bu çalışmada Wikipedia'dan alınan veriler ön işlemden geçirildikten sonra analizler gerçekleştirilmiştir. Numpy, Pandas, Matplotlib, LXML kütüphaneleri kullanılarak ön işlem ve analiz işlemleri yapılarak metin madenciliği çalışması gerçekleştirilmiştir. Her dilin farklı kuralları olduğu için metin madenciliğinde zorlukları beraberinde getirmektedir. Veri setinin düzgün bir şekilde oluşturulmasına engel olabilmektedir. Fakat birçok metin ön işleme yöntemi ile veri seti anlaşılabilir ve analiz edilebilir hale getirilmektedir. Tokenizasyon, lemmatizasyon, sık kullanılan ve bir anlam ifade etmeyen durma kelimelerinin çıkarılması, tüm karakterlerin küçük harfe çevrilmesi, gereksiz karakterlerin metinden ayıklanması, her türlü noktalama ve özel karakterlerin silinmesi gibi işlemlerden geçirilerek daha iyi bir veri seti elde etmek mümkündür.

Literatürde metin madenciliği konusunda Türkçe kaynak yeterince bulunmadığı araştırmalar sonucunda görülmüştür. Wikipedia üzerinde metin madenciliği çalışmaları yok denecek kadar azdır. Bu motivasyon ile çalışma gerçekleştirilmiş olup konu hakkında detaylı bilgiler verilmiştir.

KAYNAKÇA

[1] İnternet: Vikipedi, https://tr.wikipedia.org/wiki/Vikipedi, (Erişim Tarihi: 28.12.2020).

[2] İnternet: Metin Madenciliği Nedir?, http://www.metinmadenciligi.com, (Erişim Tarihi: 28.12.2020).

[3] Nazan, Ö., and Ayvaz, S. Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. (2018) Telematics and Informatics 35.1 (2018): 136-147. DOI:10.1016/j .tele.2017.10.006

[4] Lim, Ji Young, et al. Identifying trends in nursing start-ups using text mining of YouTube content. Plos one15.2 (2020): e0226329. DOI:10. 1371/journal.pone.0226329

[5] Salloum, Said A., Mostafa Al-Emran, and Khaled Shaalan. Mining text in news channels: a case study from Facebook. International Journal of Information Technology and Language Studies 1.1 (2017): 1-9.

[6] Jack, Lleyana, and Y. D. Tsai. Using text mining of amazon reviews to explore user-defined product highlights and issues. Proceedings of the International Conference on Data Science (ICDATA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), (2015).

[7] İnternet: Python (programlama dili), https://tr.wikipedia.org/wiki/Python_(programlama _dili), (Erişim Tarihi: 28.12.2020).

[8] İnternet: Anaconda (Python dağıtımı), https://tr.wikipedia.org/wiki/Anaconda_(Python_da ğıtımı), (Erişim Tarihi: 28.12.2020).

[9] İnternet: JupyterLab Overview, https://jupyterlab.readthedocs.io/en/stable/getting_s tarted/overview.html, (Erişim Tarihi: 28.12.2020).

[10] İnternet: Pandas Nedir? Nasıl Kullanılır?- Python Kütüphanesi, https://teknoloji.org/pandas-nedir-nasil-kullanilir-python-kutuphanesi/, (Erişim Tarihi: 28.12.2020).

[11] İnternet: Python Bilim Matplotlib, https://pybilim.wordpress.com/2014/01/01/matplotl ib-1-temel-grafikler/, (Erişim Tarihi: 28.12.2020).

[12] İnternet: Python ve XML, https://ozymaxx.github.io/blog/2017/09/12/python-xml/, (Erişim Tarihi: 28.12.2020).

[13] İnternet: Metin Madenciliği'nde (Text Mining) Kavramlar-1, https://medium.com/algorithms-data-structures/metin-madenciliğinde-text-mining-kavramlar-1-e11b87b28847, (Erişim Tarihi: 28.12.2020).

[14] İnternet: Python ile Veri Kazıma (Web Scraping) Çalışması, https://medium.com/kaveai/web-scraping-453e96a86195, (Erişim Tarihi: 28.12.2020).

[15] İnternet: Python ile Veri Görselleştirme: Matplotlib Kütüphanesi-1, https://medium.com/datarunner/matplotlibkutuphanesi-1-99087692102b, (Erişim Tarihi: 28.12.2020).