Ekran Kartı Yorumları Üzerine Web Kazıma ve Veri İşleme Uygulaması

1.Ahmet Barış Yardımcı   
*Bilişim Sistemleri Mühendisliği*  
*Kocaeli Üniversitesi*Kocaeli, Türkiye  
[baris.yrd30@gmail.com](mailto:baris.yrd30@gmail.com)

2*.Emre Kardaş  
Bilişim Sistemleri Mühendisliği  
Kocaeli Üniversitesi  
Kocaeli,Türkiye*emrekrds0@gmail.com

*Özet: Bu çalışmada, bir Türkçe forum sitesinden elde edilen kullanıcı yorumları üzerinde çeşitli doğal dil işleme (NLP) teknikleri uygulanmıştır. Kazıma (web scraping) sürecinden başlayarak, verinin temizlenmesi, tokenizasyon, durdurma kelimeleri temizleme (stopword removal), kök bulma (stemming), ve kök çözümleme (lemmatization) aşamalarına kadar detaylı bir NLP süreci gerçekleştirilmiştir. Çalışma, Windows ortamında Python programlama dili ve çeşitli kütüphaneler (BeautifulSoup, Selenium, NLTK, vb.) kullanılarak yürütülmüştür. Ancak, bazı kütüphanelerin (spaCy, Zemberek) sınırlamaları nedeniyle alternatif teknikler tercih edilmiştir.*

# Giriş

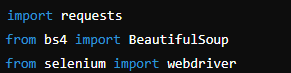
Bu bölümde çalışmanın amacı, kullanılan veriler, ve yöntemlerin genel bir tanıtımı yapılır. Örneğin: Giriş: Türkçe doğal dil işleme (NLP) üzerine yapılan bu çalışmada, bir Türkçe forum sitesinden kazınan (scraping) veriler üzerinde çeşitli ön işleme ve analiz teknikleri uygulanmıştır. Kazıma süreci ile elde edilen veriler, metin temizleme, tokenizasyon, durdurma kelimelerinin temizlenmesi, lemmatizasyon ve kök bulma gibi aşamalardan geçirilmiştir. Çalışma, NLP işlemlerinde karşılaşılan teknik sınırlamaları ele alarak, bu sınırlamaların üstesinden gelmek için alternatif çözüm yollarını araştırmıştır.

# YÖNTEM

## Web Kazıma (Scraping)

Web kazıma işlemi, Python programlama dili ile BeautifulSoup, Selenium ve diğer yardımcı kütüphaneler kullanılarak gerçekleştirilmiştir. Web kazıma sürecinde aşağıdaki adımlar izlenmiştir:

* Kod Blokları: Selenium ve BeautifulSoup kullanılarak forum sitesinden yorumlar çekilmiştir. Kod, belirli anahtar kelimeleri içeren başlıkları bulmak ve başlıklara ait yorumları toplamak üzere yapılandırılmıştır.



* Yorumların Çekilmesi: Belirli forum sayfalarından anahtar kelime içeren başlıklar ve ilgili yorumlar alınarak Scrapping.csv dosyasına kaydedilmiştir.

## Veri Temizleme

Kazıma işleminden sonra elde edilen veriler, gereksiz karakterlerden ve linklerden temizlenmiştir. Bu adımda kullanılan kod aşağıdaki gibidir:



Bu kod, tarih ve saat bilgilerini, URL’leri ve fazladan boşlukları temizleyerek Scrapping-Cleaned.csv adlı temizlenmiş bir CSV dosyası oluşturmuştur.

* 1. *Tokenizasyon*

Tokenizasyon işlemi ile, temizlenmiş yorum verisi kelime parçalarına ayrılmıştır. Bu işlem NLTK kütüphanesi kullanılarak gerçekleştirilmiştir.



* 1. *Stopwords Kaldırma*

Tokenizasyon işleminden sonra, metinde anlam taşımayan Türkçe durma kelimeleri (stopwords) çıkarılmıştır. Bu işlemde NLTK’nın Türkçe stopwords listesi kullanılmıştır.

* 1. Lemmatizasyon

Lemmatizasyon, kelimelerin kök veya temel formlarına indirgenmesi işlemidir. Bu çalışmada, spaCy kütüphanesi yerine NLTK kullanılmıştır.

* 1. Kök Bulma

Son adımda, TurkishStemmer kütüphanesi kullanılarak kök bulma işlemi gerçekleştirilmiştir.

# Sonuçlar ve Tartışma

Yapılan analizler, donanım forumlarında ekran kartları hakkında yapılan yorumların belirli başlıklar altında anlamlı bir şekilde sınıflandırılabileceğini göstermektedir. Ancak, kök bulma işlemi sırasında Türkçe karakterlerin doğru analiz edilmesi için daha gelişmiş Türkçe NLP araçlarının geliştirilmesi gerektiği gözlemlenmiştir.

Bu çalışmada, ekran kartı yorumlarının analiz edilmesi için web kazıma ve çeşitli NLP teknikleri uygulanmıştır. Elde edilen sonuçlar, donanım forumlarındaki kullanıcı geri bildirimlerinin analiz edilmesi ve anlamlandırılması için başarılı bir yöntem sunmaktadır. Gelecekte, Türkçe dil desteği daha güçlü olan lemmatizasyon ve kök bulma araçları geliştirilerek bu tür çalışmalarda doğruluğun artırılması sağlanabilir.

##### Bilgilendirme

Çalışma sırasında, bazı teknik zorluklarla karşılaşılmıştır. İlk olarak, spaCy kütüphanesinin Windows ortamında kurulumu ve çalıştırılması sırasında çeşitli uyumsuzluk sorunları yaşanmıştır. Bu nedenle, spaCy yerine NLTK kütüphanesi ile lemmatizasyon işlemi yapılmıştır. Ayrıca, Zemberek kütüphanesi ile Türkçe dilinde kök bulma işlemi yapılmak istenmiş ancak Maven kullanılarak derleme sırasında hata alınmıştır. Zemberek’in çalışmaması nedeniyle, Türkçe kök bulma için TurkishStemmer kütüphanesi tercih edilmiştir. Bu bilgilendirmeler ışığında, alternatif kütüphaneler ile süreç tamamlanmış ve istenilen sonuçlara ulaşılmıştır.

##### Kaynaklar

1. SpaCy Documentation, Explosion AI. Available: https://spacy.io/.
2. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," Trans. Assoc. for Computational Linguistics, 2017.
3. A. Dutton, Web Scraping with Python: Collecting Data from the Modern Web, 2nd Edition. O'Reilly Media, 2021.
4. A. Ahmet, "zemberek-nlp: Natural Language Processing for Turkish," GitHub repository, 2021. Available: <https://github.com/ahmetaa/zemberek-nlp>.
5. Explosion AI, "spaCy: Industrial-Strength Natural Language Processing in Python," GitHub repository, 2023. Available: <https://github.com/explosion/spaCy>.
6. NLTK Project, "NLTK: Natural Language Toolkit," GitHub repository, 2023. Available: <https://github.com/nltk/nltk>.
7. BeautifulSoup Project, "Beautiful Soup Documentation," GitHub repository, 2023. Available: <https://github.com/wention/BeautifulSoup4>.
8. T. S., "Selenium WebDriver Documentation," GitHub repository, 2023. Available: <https://github.com/SeleniumHQ/selenium>.
9. L. Richardson, "Beautiful Soup Documentation," Crummy, 2023. Available: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.