MAKİNE ÖĞRENMESİ

Enes GELMEZ

Bilişim Sistemleri Mühendisliği Bölümü

Kocaeli Üniversitesi

171307066@kocaeli.edu.tr

Özet

*Bu projede IMBD sitesinden belirli film türlerinin (Aksiyon,Suç,Drama) duygu analizinin yapılması için her türden yaklaşık olarak 30000 adet yorum çekilmiştir. Bu çekilen yorumlar bir diğer projede olumlu veya olumsuz olacak şekilde ayrılıp hangi tür bu internet sitesinde hangi tür filmlerin daha olumlu yorum aldığı hesaplanacaktır.*

# Giriş

import requests

from bs4 import BeautifulSoup

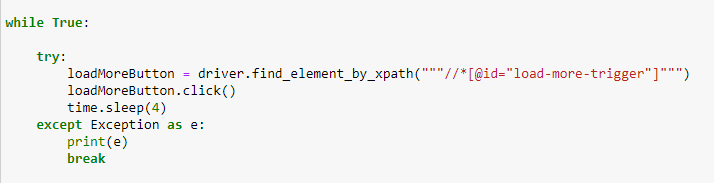
from selenium import webdriver

import time

import pdb

İlk olarak bu kütüphaneleri kullandım. BeautifulSoup kütüphanesi, HTML veya XML dosyalarını işlemek için kullanılan bir kütüphanedir. Bu yüzden IMDB sitesinin kaynak kodlarının arasından veri işlemek için bu kütüphaneyi kullandım.

[1] Selenium webdriver, karşılaştığım bir sorundan dolayı araştırmalarım sonucu bu kütüphanenin işime yarayacağını buldum. Bu sorun IMBD sitesinde bir filmin yorumları bir sayfada maksimum 25 adet gözükmektedir. Daha fazla yorum yüklemek için “Load More” butonuna tıklanması gerekmektedir. Selenium web uygulamalarını test etmek için kullanılan bir kütüphanedir. Bu kütüphaneyi projeme dahil ettiğim zaman otomatik olarak her 4 saniyede bir “Load More” butonuna tıklamasını sağladım ve böylelikle istenilen filmin tüm yorumları yüklenene kadar devam edecektir.

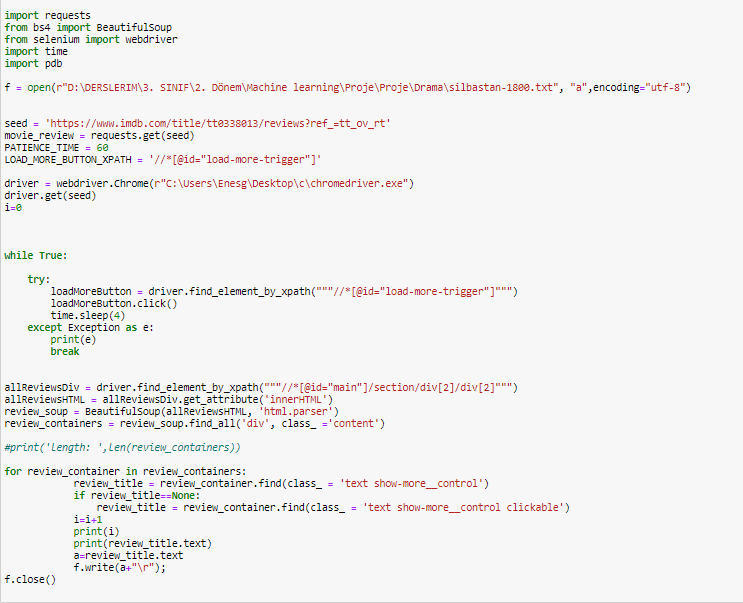


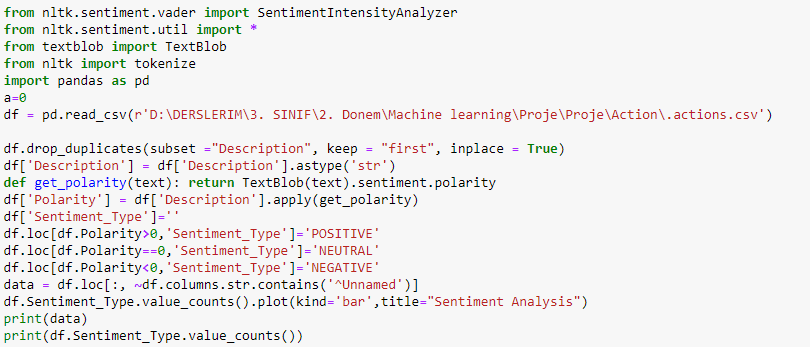
# Temel Bilgiler

Proje JupyterLab üzerinde yazılmıştır.

# Proje Kodları

[2][3] Kırmızı ile işaretlenen bölümde daha önceki kısımlarda container içerisinde tutulan yorumların üst divlerini işleyerek sadece yorumların çekilme işlemi yapılmıştır. Bu yorumları da üst kısımda yeşil ile işaretlenen yoldaki “.txt” uzantılı dosyaya yazmaktadır.





1. Resimdeki kısımda ise verilerimin temizlemeden önceki halinde olumlu veya olumsuz olarak sınıflandırılmasını sağlayan bir kod.

# Donanım

Bu uygulamayı kendi bilgisayarımı kullanarak hazırladım.

Özellikler

İşlemci: Intel(R) Core(TM) i7-8750 CPU @ 2.20GHz(12 CPu)

Bellek: 16GB

Ekran Kartı: Nvidia GTX 1050Ti

İşletim Sistemi: Windows

# Sonuç

5000 yorumu olan bir filmin yorumlarını çekmek için geçen süre yaklaşık 10-15 dakika

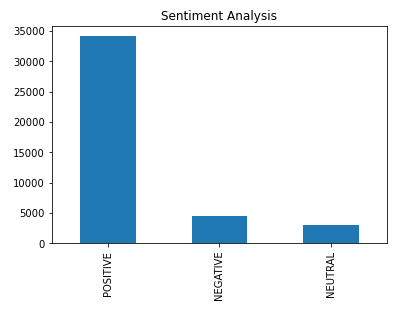
Verilerin sayısı hedeflediğim sayı: 60000

Şu anki veri sayısı: ~100000

Aksiyon filmlerindeki;

Olumlu yorum :34152

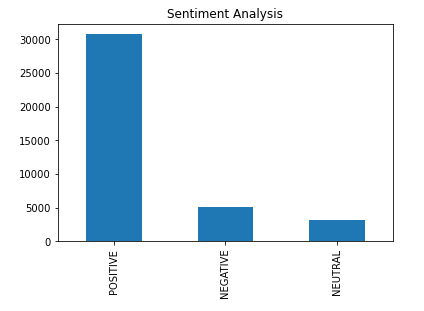
Olumsuz yorum:4441

Nötr:2951

Suç filmlerindeki;

Olumlu yorum :30763

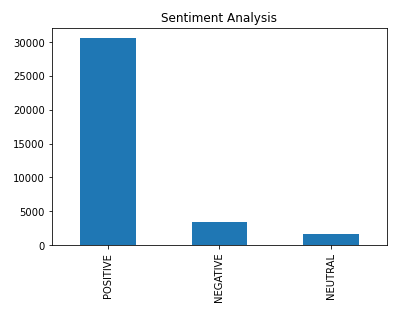
Olumsuz yorum:5089

Nötr:3091

Drama filmlerindeki;

Olumlu yorum :30591

Olumsuz yorum:3414

Nötr:1641

# Kaynaklar

1. <https://stackoverflow.com/questions/55527423/why-do-i-only-get-first-page-data-when-using-selenium>
2. <https://www.youtube.com/watch?v=00bc9BdUPSw>
3. <https://shravan-kuchkula.github.io/scrape_imdb_movie_reviews/#step-3-for-each-movie-extract-the-movies-user-reviews-link>