Project - Sentiment Analysis of User reviews for TV shows

Background

I currently work at NBC Universal as a Business Intelligence Architect with the Reporting and Analytics group (for Digital Media). Over the past few years, the way people consume media has changed drastically. In the past, Cable TV was the primary medium for consuming TV series. Fast forward to today, and the majority of the users consume their favorite TV shows through all kinds of streaming services (more than 25 in case of NBC TV shows).

The Reporting and Analytics group at NBC historically focused on collecting and analyzing data for Cable TV usage (also referred to as Linear). With the growth of streaming services, there was a need to gather and analyze Digital data. Our group was created specifically for satisfying this purpose in the past 3 years. The goal of our group is to consolidate data provided by many of our partners (like YouTube, Hulu, Roku etc..) and provide reporting and analytics to our business users.
The data we collect and report on primarily focuses around two aspects -

- Ad Sales
- Time spent watching TV shows on various platforms

We are currently not utilizing any available data on platforms like IMDB, Rotten Tomatoes and Twitter to analyze the popularity of various TV shows.
I see this as a big potential - to collect review data from the above-mentioned websites, perform sentiment analysis over time and potentially combine this data with the data collected from streaming services. This will give the analysts a view of TV shows popularity and potentially also look at what viewers are interested in.

(I would love to be able to apply the learnings from this course to apply at my workplace, and I'm sure my employer would also much appreciate the same.)

Proposal

With the above background, here is the proposal for the project I wish to implement -

Collect user review data from IMDB and Rotten Tomatoes for all the current and previous TV shows produced by NBC Universal (or any other pre-defined list of TV shows). We will collect the following -

- Timestamp of the review
- Score provided
- Comment

We will use NLP techniques to do the following -

- Create Word cloud
- Perform sentiment analysis on review comments
- Store the sentiment scores in a database by day for integrating with existing data

We will then look to integrate this data with our existing data warehouse that collects streaming information. This will give us the potential to analyze the viewership patterns and correlate with user sentiment from the web.
(This part cannot be available as a part of my project code as this data will be internal to NBC)

Implementation

For the submission, I have gone with the following implementation –

- Using scrapy, beauifulsoup library in Python, scrape data from IMDB for TV Shows
  - The *titleid* is provided as input (this is the ID used in the URL for IMDB for TV shows/movies)
  - The list of TV Shows is provided as a text file
  - The reviews are scraped and stored in a JSON file
- Train a Machine Learning algorithm with existing review data (http://ai.stanford.edu/~amaas/data/sentiment/)
  - Use this model to give Positive/negative score to each review

TV Show -> User Reviews in JSON file -> Sentiment score for each Review in a CSV file

(Note: Due to personal emergency, I was not able to complete this coding end to end. The process is implemented for a single TV Show (title) and he model creates an output file with Sentiment score for the reviews (first 25 only).)

Installation and Execution

1. Install the following libraries –
   pip install scrapy
   pip install beautifulsoup4
   pip install sklearn
   pip insall joblib
   pip install pickle

2. The model has already been trained using sample data - http://ai.stanford.edu/~amaas/data/sentiment/

3. Use this model to give Positive/negative score

4. To Scrape user review for any movie, update the link at line # 9 in "Imdb/Imdb/Spiders/imdb_spider.py"
   (In IMDB, each Movie/TV Show has a title, which is used in the URL. The title for TV Show Manifest is tt8421350, so the link should be https://m.www.imdb.com/title/tt8421350/reviews)

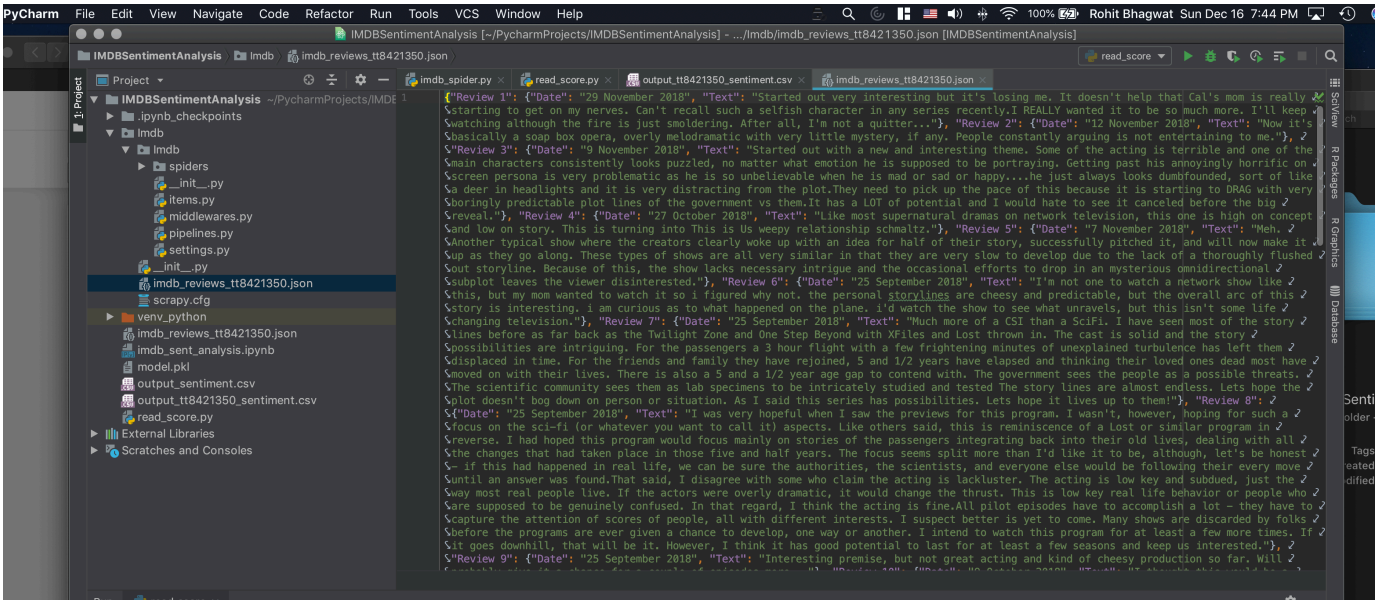
5. Run "scrapy crawl imdb" while inside the "Imdb/" directory. This will crawl and store the reviews in the folder "Imdb/Imdb/Spiders/" with ".json" extension

6. Copy the file and paste it in the main folder where the file "read_score.py" exists

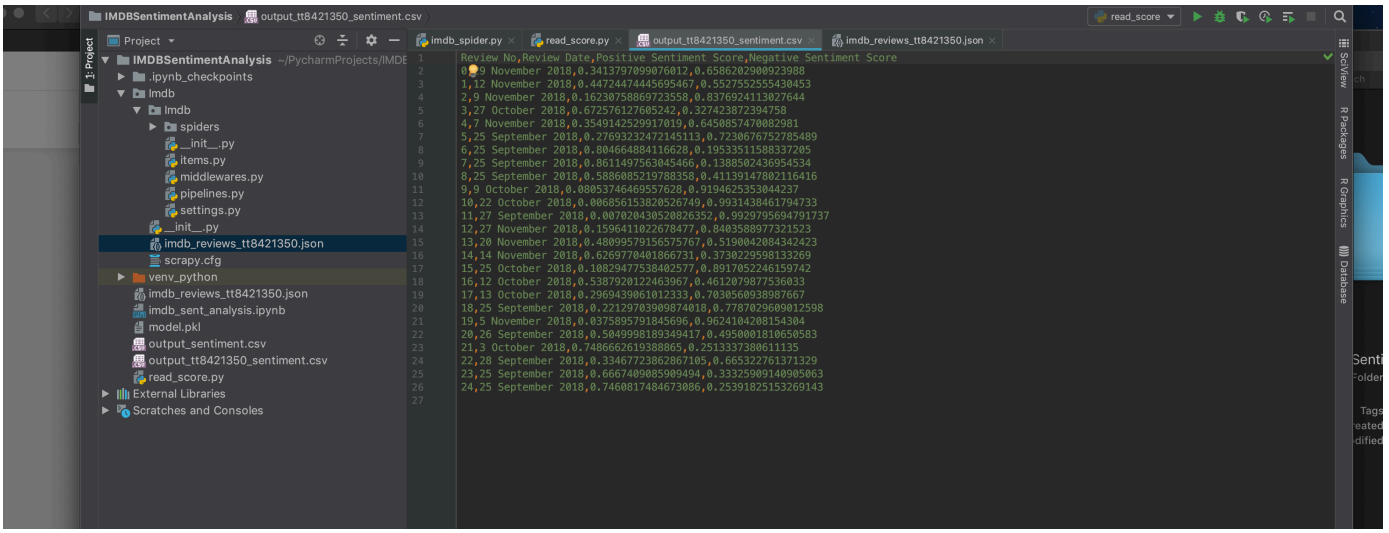
7. Run "python read_score.py" this will predict the reviews in the json file and output the scores in "output_sentiment.csv" file

Output

I ran the process for TV Show – Manifest, below is the sample JSON file with reviews



Sample Sentiment score for each review –



Future

- Parameterize Title, file generation etc..
- Create an end to end process to update the database with new process
- Join the data using Title with existing database at NBCU