Project - Sentiment analysis of TV shows

Sunday, October 21, 2018   09:17

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

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)

Technologies -
- Python BeautifulSoup for scraping data
- NLTK / spaCy (any other?)
- MySQL to store data
- Tableau to visualize data

What is the function of the tool?
- Perform sentiment analysis of user reviews from public websites on TV shows
   Open up various possibilities by helping analyze this data alongside existing streaming data

Who will benefit from such a tool?
- Analysts and Production teams at NBC Universal

Does this kind of tool already exist? If so, how is your tool different?
- Similar studies have been conducted. This tool will specifically look to target TV shows and store the results in a database to enable historical analysis

What existing resources can you use?

What techniques/algorithms will you use to develop the tool?
- Natural Language processing libraries - NLTK / spaCy / others

How will you demonstrate the usefulness of your tool?
-

A very rough timeline to show when you expect to finish what.

What techniques/algorithms will you use to develop the tool?
- Natural Language processing libraries - NLTK / spaCy / others

How will you demonstrate the usefulness of your tool?
-

A very rough timeline to show when you expect to finish what.
- Data extraction (sample set of TV shows from IMDB and Rotten Tomatoes) - November 15th
- Sentiment Analysis - December 1
- The process to load data on a daily basis - December 1
- Word Cloud and course submission - December 9

Future enhancements
- Collect data from websites like Twitter, Facebook
- Collect data for each episode if available