WEEK 9 – Final Project

Jaques D’Erasmo Santos Fernando

Abstract

This project is a sentiment analysis about Netflix. The dataset used was a 2000 tweets requested from Twitter, and the Nick Snader’s tweet's corpus, that is a collection of nearly 5000 classified tweets labelled as positive, negative, neutral or irrelevant. The research questions was: What is the sentiment about the Netflix; what is the most frequent words on those 2000 tweets; what is the word frequency and the rank of words in a specific frequency interval. In this work, the methodology used was the sentiment analysis using Twitter API and NLTK Naive Bayes Classifier, and the findings are the sentiment about the Netflix and few different perspectives about the word frequency, including a word cloud model, based on the 2000 tweets.

Motivation

The actual project proposes a sentiment analysis about the Netflix platform. This event is an important evaluation to the Netflix company decision makers since the opinion from the users can be considered directed related with its possibility to keep growing their business or, if necessary, review their services to better attend their audience.

To proper address this problem, it will be used the Twitter API to obtain a certain amount of tweets, and the NLTK Naive Bayes Classifier in order to train the Machine Learning Model. It is important to mention that the train dataset is called Nick Sander's corpus, created by Nick Sander, and that contain approximately 5000 of tweets properly labelled. A particularity about this corpus is that it was initially created to be used to train model specifically related with high technological companies as Google, Amazon, Apple, Windows, YouTube and Netflix, for example.

Word frequencies, including a amazing WordCloud view will be presented, as well.

Dataset(s)

Two datasets will be used in this project

1 – A collection of 2000 tweets collected from Twitter platform through the Twitter API system;

2 – Nick Sander's tweets corpus with approximately 5000 classified tweets labelled as positive, negative, neutral or irrelevant. Link to download the corpus https://github.com/zfz/twitter_corpus

Data Preparation and Cleaning

First, it was necessary to remove the duplicates. The Twitter always duplicates few tweets. Also, it was necessary to remove stopwords (words that is undesirable to the analysis, replace and remove URL links link ‘www’, http or https, replace and remove @ and # symbols, remove emotion icons and tokenize words.

Research Question(s)

What is the sentiment analysis about the Netflix platform services based on its user's opinions on Twitter?

Additionally to this main research question few other secondary questions will also be answered, as for example: What are the top 20 words in the vocabulary from the tweets used on the sentiment analysis? Is the vocabulary of words large or it is a short vocabulary? How a Word Cloud from the tweets looks like, and can this visualization be helpful?

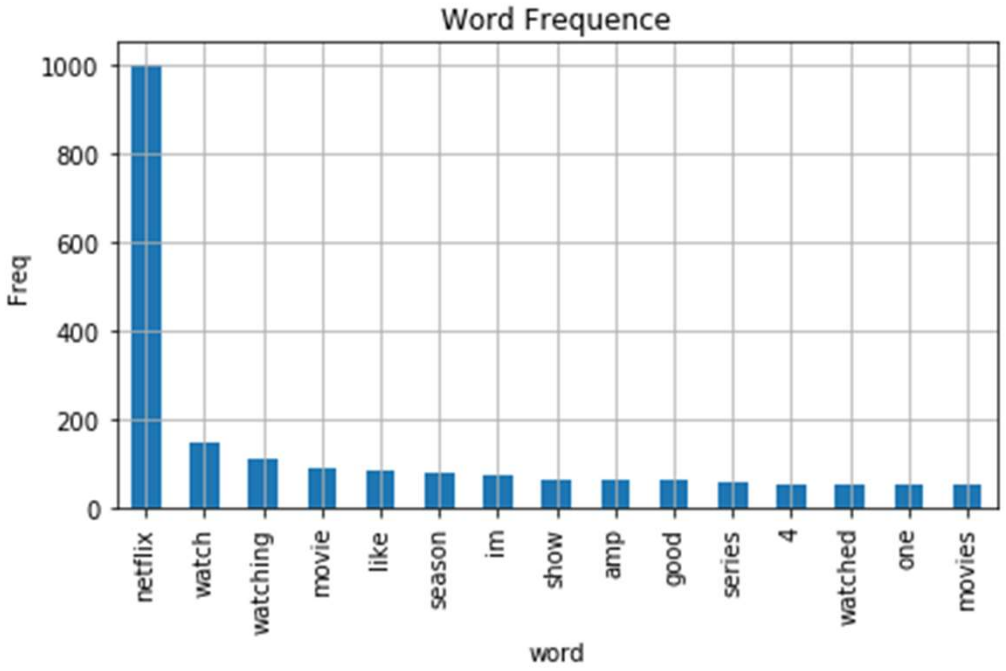
Methods

In order to perform a sentiment analysis about Netflix, it will be used the Twitter API System to obtain the tweets and the NLTK Naive Bayes Classifier to train a Machine Learning Classifier model.

Findings

Word frequency – Top 20 words

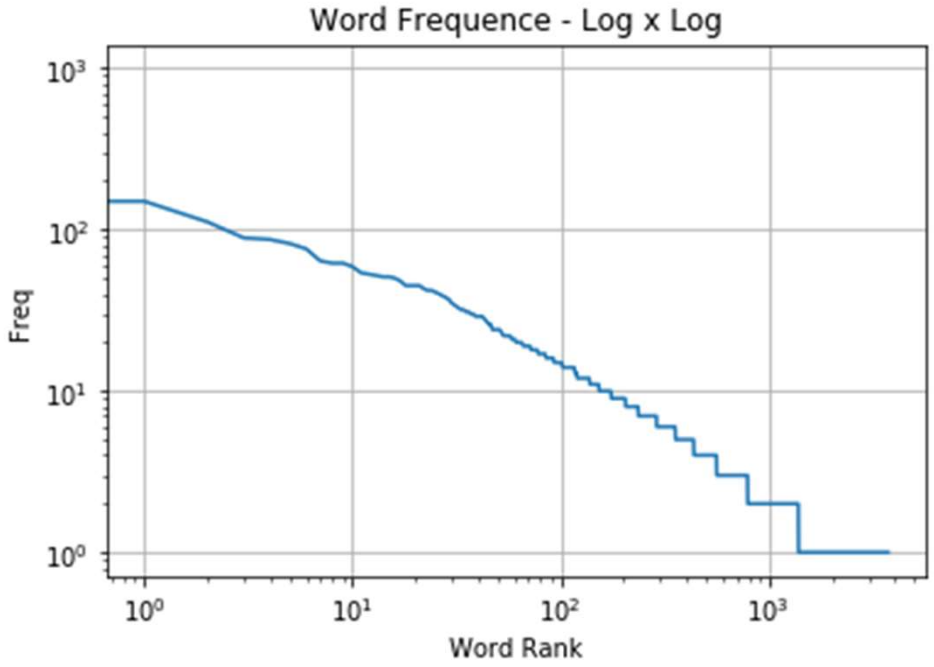
The graph in the right provides a quick and clear visualization about the top 20 most frequent words in the 2000 tweets obtained from the Twitter platform related with Netflix service.



Findings

Word frequency – Log x Log

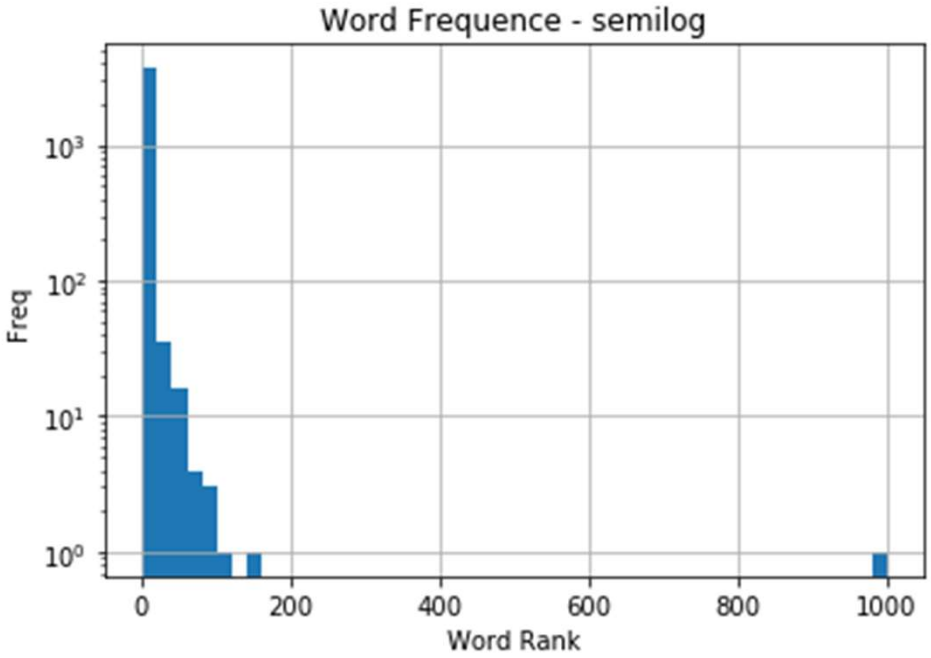
The flat shape of the log x log word frequency graph indicates that there is a large range of words in the tweets, or simply, it has a large vocabulary.



Findings

Word frequency – semiLog

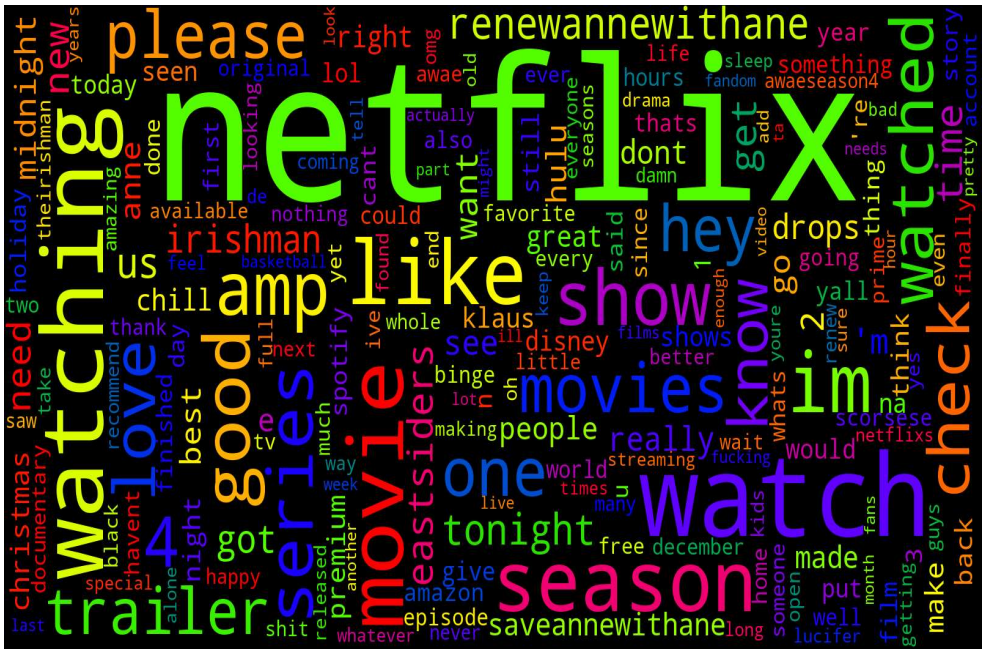
The *semilog* word frequency graph allows us to display how many words have a count in a specific range. In this specific analysis, our range is between 0 and 200 words were is located the most frequent words in the vocabulary.



Findings

Word frequency – Word Cloud

The *Word Cloud* is, in my opinion, one of the most sophisticated and easy to understand, visualization methods for a sentiment analysis. Even not providing the final sentiment result it easily leads to a major comprehension about the analyzed topic.



Findings – Sentiment Analysis Result

```
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
                SENTIMENT RESULTS USING NAIVE BAYES CLASSIFIER
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=

Positive Sentiment = 0.15%
Negative Sentiment = 0.89%
Neutral Sentiment = 95.57%
Irrelevant Sentiment = 3.40%
```

First 6 tweets

[{'text': "RT @daniballada: Each episode title of #AnneWithAnE is a quote from works of Anne's favorite authors. The seasons are dedicated to Charlott…",
  'label': None},
 {'text': '@Granddaddy_Zach On Netflix ?', 'label': None},
 {'text': '@CarrieCnh12 Yes she was distracted by her boo #DragonPrince #SaturdayNightSciFi #Netflix',
  'label': None},
 {'text': "Daybreak on Netflix is my type of show.\nIt's a shame it's only 10 episodes and I'm half way through.",
  'label': None},
 {'text': 'RT @6ODWG: watching the blacklist on Netflix https://t.co/1nMcJ5OSB8',
  'label': None},
 {'text': "RT @MillennialOfMNL: i recommend watching carole & tuesday on netflix!!! it's about two girls wanting to make music in a futuristic world w…",

Findings – Sentiment Analysis Result

Naive Bayes Classifier

According to the preview slide the overall sentiment about the Netflix service can be considered **Neutral** (95.57%). A quickly evaluation about the first 6 tweets, also presented above, allows us to understand that, in general, the tweets are expressing a feeling about a movie, or series, or even just doing a comment about their personal mood or what they are watching at that time, what could justify the overall neutral evaluation about specifically Netflix services.

It is important to remember that the model was trained with a corpus specifically created for this purpose, identify the sentiment about the brand and its services. This point will be very important to understand the next extra analysis.

Findings – Sentiment Analysis Result

TextBlob Library

During the project development, I had the opportunity to discover and learn about a great library called TextBlob. TextBlob is an object-oriented NLP text-processing library that is built on the NLTK and pattern NLP libraries and simplifies many of their capabilities (Deitel, 2019).

This section has the object to provide a simplified approach for sentiment analysis but also clarify few particularities that may limit the TextBlob usability.

Findings – Sentiment Analysis Result

TextBlob Library

```
-=-=-= PERCENTAGE -=-=-=
42 % Positives

13 % Negatives

44 % Neturals
```

The result of the Sentiment Analysis using the TextBlob library is presented in the left box, where, according it, 42% of the tweets are positive, 13% are negative and 44% neutral.

In fact, this evaluation can be considered correct, but it is important to clarify that "**libraries like TextBlob have pretrained machine learning models for performing sentiment analysis**" (Deital, 2019).

Another important point is that, by default, "TextBlob use a PatternAnalyzer, which uses the same sentiment analysis techniques as in the Pattern library" (Deitel, 2019). That is different to the Naive Bayes Classifier used previously.

Considering those points, it is possible infer that the result from TextBlob might correct about its perception but possible about a general evaluation, not specifically focused on the Netflix services user's sentiment.

Limitations

The main limitation encountered was that the model is limited to analysis sentiment about high tech companies (Amazon, Google, Apple, Windows, YouTube, Netflix, etc) user’s satisfaction about those companies' products and service's quality. If used in another context may end up leading to misguided results.

The model can be also limited about the language, since it is designed to capture tweets only in English. An application to translate the tweets if it was not originally in English was tried without success, what can be a point for future improvements.

Conclusions

The Sentiment Analysis using the Naive Bayes Classifier provided a conclusion that 95.57% of the Netflix user's sentiments is neutral. According to this same analysis, 0.15% can be considered positive, 0.89% negative, and 3.40% irrelevant. This result is correlated to the fact that the most part of the tweets was not focused on the Netflix platform services but in general topics, that was identified by the model as neutral, in most of the cases, since the training data was specific related to the products and services.

An additional sentiment evaluation using the TextBlob library provided an entirely different result with 42% of the tweets positive, 13% negative, and 44% neutral. This result can be classified as a misleading result since the TextBlob library was trained with a generic corpus and use a different sentiment analysis technique.

Conclusions

The word frequency analysis provided valuable information and comprehension about the dataset. Based on these analysis, it is possible to infer that there is a large range of words on the dataset and that there are around 200 words concentrating the most frequent words on the entire dataset vocabulary.

Additionally, it is important to mention how helpful was the Word Cloud visualization, creating a quick and easy understand result about the dataset and its word varieties.

Acknowledgements

Thank you Nick Sanders, through Sanders Analytics, for providing an amazing corpus to be used as a training data to the sentiment analysis model.

A special thank you to my permanent Data Science study group composed by old friends who decided to go through this path together. Felipe Solares, Eduardo Passos and Felipe Brandão, that provided valuable feedbacks and insights to this project.

References

Deitel, H., Deitel, P., & Deitel, P. J. (2019). Python for Programmers, First Edition. Retrieved from: https://learning.oreilly.com/library/view/python-for-programmers/9780135231364/ch11.xhtml

Total Training. (2017). Machine Learning - Twitter Sentiment Analysis in Python. Retrieved from: https://learning.oreilly.com/videos/machine-learning/10000LCTWITTE

Edx Data Science MicroMaster Python for Data Science class notes. (n.d.). Retrieved from: https://courses.edx.org/courses/course-v1:UCSanDiegoX+DSE200x+3T2019/course/