CS 410: TECHNOLOGY REVIEW

Sentiment Analysis with NLTK and VADER

Rudy Rath

GitHub Link: https://github.com/rudyrath/tech_review

Introduction

In this article I will provide a brief introduction to Sentiment Analysis, followed by an overview of the NLTK and VADER tools, and then briefly describe the commands for doing Sentiment Analysis with NLTK and VADER.

Sentiment Analysis

Wikipedia defines Sentiment analysis as –

"Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information."

Basically, it is the process of computing whether a piece of writing is positive, negative, or neutral.

It has many applications such as-

- Businesses can discover insights about their products and services, such as what people like or dislike about the product or service, how is the response to branding, marketing campaigns, product launches etc. and then use the information to make better business decisions.
- In Politics, it can be used to understand public view about actions, proposed legislations of the government, popularity of electoral candidates etc.
- In Society, it can be used to understand the general mood of the population in news, social media, identify potentially dangerous situations like civil unrest, riots etc.

NLTK

Steven Bird and Edward Loper, from the Department of Computer and Information Science at the University of Pennsylvania, created **Natural Language Toolkit (NLTK),** a platform for building NLP applications in Python.
NLTK is described as a platform rather than just another Python library because, in addition to a collection of modules, it includes several contributed datasets. These datasets (corpus), can be used to tokenize and tag text, identify named entities, display a perse tree etc.

NLTK requires Python versions 3.6 and above and it comes with default installation with **Anaconda** or can be installed using **pip**.

NLTK Data (corpora, grammars, trained models) can be downloaded using the **download()** method

NLTK Documentation at https://www.nltk.org provides API References and Example usage.

There is also a free book *"Natural Language Processing with Python"* available at https://www.nltk.org/book which was published by O'Reilly.

VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as either positive or negative.

VADER not only about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.

It uses a dictionary of terms that it can evaluate and the GitHub repository at https://github.com/cjhutto/vaderSentiment includes examples for-

- Negations - a modifier that reverses the meaning of a phrase ("not great").
- Contractions - negations, but more complex ("wasn't great").
- Punctuation - increased intensity ("It's great!!!").
- Slang - variations of slang words such as "kinda", "sorta", "Hella" etc.

Sentiment Analysis with NLTK and VADER

To do Sentiment Analysis with NLTK and VADER one should extract the data from social media to prepare the data set.

After that one can–

- Download the VADER lexicon using

      nltk.download('vader_lexicon')

- Import the VADER Sentiment Analyzer using

      from nltk.sentiment.vader import SentimentIntensityAnalyzer

- And then score a particular comment from the collection of comments using

      analyzer.polarity_scores()

The scoring is a ratio of the proportion for text that falls into each category.

Conclusion

In this review we see that NLTK, and VADER can be used easily for performing Sentiment Analysis. However, there are some limitations to this language specific method like - the grammatical mistakes and misspellings may not be handled correctly, emotions like sarcasm may be misinterpreted, domain-specific contexts may not be part of the lexicon etc.
One would need advanced Machine Learning and Deep Learning techniques to handle such scenarios.

Bibliography

https://en.wikipedia.org/wiki/Sentiment_analysis

https://www.nltk.org

https://github.com/cjhutto/vaderSentiment

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.