Sentiment Analyses with Twitter Data

Uday Keith

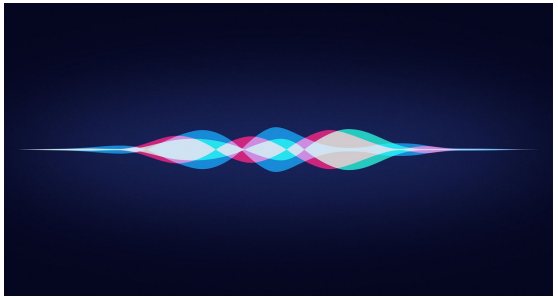
Data Science Instructor
Byte Academy

What is NLP?

Formal definition: “Natural Language Processing, or NLP, is an area of computer science that focuses on developing techniques to produce machine-driven analyses of text.”

Functional definition: branch of data science that processes, analyzes and derives information from the text data which may be used for prediction.

NLP in our daily lives: Siri, chatbots, Gmail SPAM filter



NLP: A “hard” problem

Language, any language, is inherently ambiguous. One person's interpretation of a sentence may differ from another person's. Due to this inability to consistently be clear, it's hard to have an NLP technique that works perfectly.

Example: *“Visiting aunts is a pain”*
This can be interpreted as either:

1) The act of going to meet aunts is painful
2) Aunts that come to visit is painful

Another example: *“Uday Keith is a boss”*
This can be interpreted as either:

1) Uday has a team under him
2) He is extremely cool

So...Sentiment Analyses?

Process of determining whether a piece of text is positive, negative or neutral. Also known as opinion mining, where the opinion or attitude of the text is derived.

Text_1 = “The Iphone X is an amazing phone!”

Text_2 = “The Iphone X is the absolute worst!”

Today’s Agenda

Challenge: Predict whether market sentiments on new Iphone X are positive or negative

To do:

- Using the Twitter API : Tweepy
- Getting familiar with NLTK
- Predicting using Naive Bayes Classifier

Twitter API

- API: Set of dedicated URL's that return only data responses
- Twitter API: Protocol to have JSON objects returned which contain data and metadata of Tweet (Docs: https://developer.twitter.com/en/docs)
- Twitter Account is required!!
  - Go to https://apps.twitter.com/ → Click "New App" → Fill the application details. You can leave the callback url field empty → Once the app is created, you will be redirected to the app page. → Open the 'Keys and Access Tokens' tab.
  - Copy 'Consumer Key', 'Consumer Secret', 'Access token' and 'Access Token Secret'.
  -
  -

Twitter

- Twitter is a pythonic wrapper for the Twitter API. Very easy to use **(Command: *pip install python-twitter*)**
- Tweepy is another alternative ***(Command:pip install tweepy)***
- Since we need to get tweets we will use the GET method.
- Docs: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

Lets give it a Try!

Next Step: Clean data

- First: Put data (which is a list of json objects (dicts in python) into a dataframe.
- Then: Filter for only english tweets
- Then: Use regular expression to clean for links and special characters
- Oops, we still have the retweet character, let’s deal with that next!

Ok..so how do I classify words??

- To accomplish sentiment analysis computationally, we have to use techniques that will allow us to learn from data that's already been labeled. This is called supervised learning.

- Fortunately, I have a list of positive_tweets and negative_tweets. Lets see what words are in each list. [Fork: *https://github.com/udaykeith/Sentiment-Analyses*]

Naive Bayes

- Old, reliable supervised learning technique based on Bayes' rule:

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

- Important assumption: each word/token is **independent!**

How to “fit” the model

- We now must have our Naive Bayes Classifier, learn the relative weights of our positive and negative words.
- Split both lists into training and test set.
- Train classifier on training set.
- *Lets see what our classifer has learnt!*

Sentiment Analysis: Other Use cases

Stock Market Prediction: Is a tweet regarding a stock positive (bullish) or negative(bearish)

Politics: Help predict election results (Are people speaking positively or negatively about a certain candidate or policy)

Product Reviews: Is my product viewed in a positive or a negative light?