Diego Valdes

IST 736

Home Work 1

10/11/2019

**Introduction**

If one were to consider for a moment, the amount of data they generate in a day, they may be surprised at the footprint they are leaving without even knowing. Consider, for a moment, all of the internet searches a person does in a day for shopping, dining, and random questions. Then add the social media posts. Top that off with any pictures taken that day. Lastly, lets add any third party indirectly collecting data, like traffic lights as an example. Repeat that for everyone in the developed world, and suddenly that’s a whole lot of information to comb through. Now imagine, if through all of that data, a company was tasked with finding a particular piece of information to assist them is improving sales. The analogy ‘needle in a haystack’ may seem appropriate.

That is where artificial intelligence comes into play and why it’s one of the buzz words in every industry today. It is a possible solution to problems that didn’t exist in the past, as well as for situations where humans are simply not the right tool to accomplish the task. In most of these situations, humans are the best suited, but the volume of the work and the time constraints make it impossible to rely on humans. Think of the man power required to read all the tweets posted in one day to search for something inappropriate; there may not be enough humans on the planet to accomplish that task in a timely manner.

In the past few years, social media corporations have been at the forefront of critisicism for their inability (or lack of effort) in curbing hate speech, or other forms of speech that can be deemed harmful. While there are many avenues to examine this issue, the one at heart is if these corporations can use A.I. to target it’s customers for advertisements, then the same technologies need to be used to analyze user produced content to find anything that is inappropriate.

**Analysis**

**About the Data**

The dataset is a collection of movie reviews provided by the course. The intended format is a complete review in one column and the sentiment of the review in the second column. Each review is riddeled with punctuation marks and special characters that make it impossible to bring in for analysis immediately. As part of the cleaning process, the dataset was read in and the last word (the sentiment) was removed and placed in it’s own column. The remaining text on each line was stripped of all punctuation and special characters. The data was then reconstructed in a new csv file with a header, and two columns.

Separate smaller datasets where created based on the specifications of the sentiment analysis packages (NLTK and Vader). NLTK required that the reviews be a list of words with the sentiment separated. Vader only required the review to be a list of words without the sentiment classification.

**Models**

The NLTK sentiment analyzer was used to create the first model. The data was split into the required format, a list of the words in the review and the sentiment. Words were marked as negative as part of the process and a model was created using the training set and negative defined words. The test set was used to test the accuracy of the model.

The second model was built using the NLTK Vader package. Vader didn’t require training, it took in the reviews and assigned each one a percentage as either negative, positive, or neutral.

**Results**

NLTK Sentiment Analysis functions more like a traditional machine learning model, where it is trained and tested, and the test results in an accuracy. NLTK performed well, predicting the sentiment of reviews 84% of the time. The model did better predicting negative reviews but not by a significant margin, ~ 3%.

Vader wasn’t trained and assessed the reviews ‘right out of the box’. It assigned a value to negative, positive, neutral for each review. Reviews in the dataset were in order with negative reviews being the first group. As with NLTK, these reviews were shuffled before being processed. In order to make a comparison between Vader’s results and the sentiment, the resulting percentage sentiment would need to be compared with the sentiment of the raw data. This could be done by creating a threshold in Vader’s results to make the results binary.

Unfortunately, due to lack of experience in the topic at hand, and the with the package at hand, and the lack of knowledge on how the sentiment was assigned, this course of action will not be taken. It is imperitiave to be aware of when an analysis has reached a point where more knowledge is required to be able to move the process forward.

**Conclusion**

Both packages have advantages/disadvantages over the other. The biggest downfall of Vader, is a defined way to determine if the results have any validity. The package functions as a black box, with results being generated, but no way, currently known, of verifying the results. This may be of concern to an implementer of the package, as there are times language isn’t black and white, or positive and negative as it may seem to clearly be.

NLTK has the advantage of functiononing as a traditional machine learning algorithim. This is valuable because it can be trained, tested, and placed up against the scrutiny that any A.I. should continually have to undergo. The results of this model were also extremely promising, performing at a rate worthy of a ‘B’.

In conclusion, NLTK would be the better selection of these two packages. It allows for the implementer to dictate the training methods and standards better than Vader did in this analysis. If Vader would need to be considered for future projects, it would need to under go more scrutiny and a method to validate results be created. NLTK already has much of that available to it and because of the training methods, can be altered to meet the changing standards of language without waiting for a 3rd party to update a package.