OMDENA SCHOOL: HATE_SPEECH_DETECTION

PROFESSOR: DR. MOHAMMED ZUHAIR AI TAIE

NAME: ARMIELYN C. OBINGUAR

DATE ASSIGNED: November 23,2021

DATE SUBMITTED: November 25,2021

1) How is stemming different from lemmatization and when to use each one?

First and foremost, we will only employ stemming if we are attempting to decipher the meaning of a certain word. Most of the time, this is something that must be considered while doing NLP jobs. Thus, lemmatization refers to the words that are being used in a given statement or phrase, as well as how they are being utilized in that statement or sentence in the English language.

Stemming is employed when one has to chop off the end or beginning of a word, which is known in the English Language as prefixes and suffixes, respectively, that are being added to a word. It comes from the world itself. The process of lemmatization, on the other hand, takes into account the morphological examination of the words in question. It is required to have thorough dictionaries that the algorithm may search through in order to relate the form back to its lemma in order to do this.

To summarize, creating a stemmer is significantly easier than creating a lemmatizer. To develop the dictionaries that enable the computer to find the right form of the term, considerable linguistic expertise is necessary. This will minimize noise and improve the accuracy of the information retrieval results.

2) Which one is more critical, overfitting or underfitting?

The challenge of which is more essential Overfitting vs. Underfitting happens when managing polynomial degrees. A bigger power enables the model to hit as many data points as practicable. Simulations of both situations are used to examine the underfit model, which is the quickest approach to understand the material. There is underfitting when there is a substantial inaccuracy in training. Overfitting happens when the testing error is large compared to the training error, or the gap between the two is wide.

As a conclusion, underfitting is very important and much crucial since the Learning algorithm is unable to represent training data. Model is simplistic and inflexible, making it difficult to gather data points. Model shows considerable bias.

3) Can you add some cleaning steps, other than the ones already stated in the lecture?

It is the Noise Removal which indicates that That is, everything that makes the text more than a string of words. This can take a long time depending on your data. These methods exist to replace the necessity for Regex for every HTML element. BeautifulSoup can eliminate HTML tags, and Regex substitutions can solve most of the rest:

The stemming and the normalizing are the next steps in the process of normalization. If you're going to normalize a text, you're going to have to make a decision about whether or not to remove contractions. Use lemmatization in order to collect accurate data or do complex analyses. We use nltk to import a number of sets relevant to stemming or lemmatization in order to do so.