Paraphrase Identification System Description

Eric Nguyen

October 27, 2022

Features

For this project, I searched for any features that seemed applicable to the string/sentence similarity problem. I used trial-and-error to test different features and their impact on the accuracy. I removed features that did not significantly impact test accuracy. The following features are used:

- Word Mover Distance (Package: Gensim)
- BLEU Evaluation Metric (Package: NLTK)
- METEOR Evaluation Metric (Package: NLTK)
- Jaccard Similarity (Package: TextDistance)
- Damerau Levenshtein Similarity (Package: RapidFuzz)

Data Preprocessing and Feature Preprocessing

For data preprocessing, the punctuation was corrected to follow standard English and each sentence was set to lowercase. I found that these preprocessing steps improved the accuracy through trial-and-error.

For feature preprocessing, libraries were used to create the features so they are already preprocessed.

Algorithms and Libraries Used

The algorithm I used for this project is logistic regression. Compared to SVM, through trial-and-error, I found logistic regression had the better accuracy.

Many different libraries were used in this project, including: Pandas, scikit-learn, Gensim, NLTK, TextDistance, RapidFuzz.

Pandas is used to store the data in DataFrames. scikit-learn is used for its logistic regression model. The rest of the libraries, Gensim, NLTK, TextDistance, and RapidFuzz are used to calculate the features. For the word mover distance, I used the `glove-wiki-gigaword-50` corpus because it had the least number of records which would make it faster to load.

Lessons Learned

Through this project, I learned...

- about the various metrics used in NLP such as word mover distance and the BLEU evaluation metric
- that adding more features does not make your classifier more accurate
- how classification works using logistic regression or SVM on a set of features