ECS763U/P - Natural Processing Language

Vector Space Semantics for Similarity between Eastenders Characters

Sparsh Verma 220996233

Q1. Improve pre-processing (20 marks).

A1. The following pre-processing techniques were used to reduce the mean rank by 59.77%, and increase accuracy by 54.41%.

1. **Lower case**: Converted the full dataset to lower case to aid getting the context of the text for usage later in the development pipeline. This also helped to improve parsing.
2. **Tokenisation**: The NLTK python package was used to split the sentences into smaller segments called tokens. "Tokenization helps in interpreting the meaning of the text by analysing the sequence of the words."
3. **Remove non-alphabetical characters**: This was added to aid text parsing and allow the model to derive relationships between the different tokens.
4. **Stopword removal**: The NLTK library was used to remove all low-level words that do not have any significance in the classification of the text i.e., the, an, a, so, what, etc.
5. **Lemmatization**: "Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. For example, if confronted with the token saw, . . . lemmatization would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun."

Q2. Improve linguistic feature extraction (20 marks).

A2. The following feature extraction methods were used to improve the mean rank by 15.05%. Although the mean rank decreased, no change in the accuracy was recorded:

1. **POS-tagging**: This technique was used to group words into different lexical categories or parts-of-speech, such as nouns, verbs, and adjectives. These categories were then assigned short labels, or tags, such as NN, VB, etc.
2. **Extracting n-grams up to length 3**: The n-grams model aided the prediction of subsequent words by grouping the tokens in sets of two (trigram model), and running a probability distribution to gauge the likelihood of observing the next word in a given sequence.

Q4. Add dialogue context and scene features (20 marks).

A4. The context of the data was incorporated by using the line spoken by the characters in terms of the lines spoken by other characters in the same scene. The name of the character was used to monitor the context of the dialogue, however as observed from the output values, the effect of this addition to the features had no impact on the mean rank or accuracy. The author tried various implementations, including using the gender as contextual data, but the results were the same each time.

Q5. Improve the vectorization method (20 marks).

A5. The TF-IDF vectorisation was used to improve the mean rank and accuracy by 28.57% and 16.05%, respectively. This proves that ". . . the goal of using TF-IDF instead of the raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus."

Q6. Run on final test data (10 marks).

A6. This displays the final mean rank and accuracy values that were produced from the Dictionary and TF-IDF vectorisation methods. As can be seen, using the latter produces a perfect value for the mean rank and accuracy after applying all the new pre-processing and feature extraction techniques.