EECS763P Assignment 2 Report

Name: Bheki Maenetja
Student No.: 230382466

Question 1

Six pre-processing techniques were examined (see code block 6 in the notebook). The first is the basic whitespace tokenisation provided by default. The next technique involves whitespace tokenisation as well as the removal of stopwords and punctuation. Building on this technique, lemmatisation and stemming were also examined. Finally, the technique of whitespace tokenisation was re-examined, this time with all tokens either uppercased or lowercased. The results (see table 1 or code block 19 in the notebook) show that lemmatisation and stemming yield the best performance in terms of mean rank whilst the lowercasing and uppercasing techniques yield better results in terms of accuracy. All the additional techniques lead to an improvement in mean rank and accuracy on the validation set in comparison to the default whitespace tokenisation technique provided.

	Preprocessing Technique	Mean Rank	Mean Cosine Similarity	Accuracy
0	Basic Whitespace Tokenisation	4.6	0.970354	0.4
1	Remove stopwords and punctuation	3.9	0.979640	0.3
2	Stem of words	3.8	0.978641	0.3
3	Lemma of words	3.8	0.979483	0.3
4	Lowercasing	4.1	0.970827	0.5
5	Uppercasing	4.1	0.970827	0.5

Table 1 Results of different pre-processing techniques with default technique for feature extraction.

Question 2

Four feature extraction techniques were examined (see code blocks 8 and 9 in the notebook). The first is the simple token count (or term frequency) feature provided by default. The next technique involves term frequency inverse document frequency (TF-IDF). N-gram (specifically bi-gram) features are also examined. Finally, parts-of-speech (POS) tagging is also looked at. The results (see table 2 or code block 20 in the notebook) show that TF-IDF performs slightly better than the rest both in terms of mean rank and accuracy.

	Feature Type	Mean Rank	Mean Cosine Similarity	Accuracy
0	Simple word count	4.6	0.970354	0.4
1	TF-IDF	4.4	0.950503	0.4
2	N-Grams	4.8	0.531516	0.1
3	POS Tagging	5.4	0.951395	0.1

Table 2 Results of different feature extraction techniques with default pre-processing technique.

Question 3

The function create_character_document_from_dataframe has been modified (see code block 4) to incorporate additional context. The function now looks at every episode and scene and incorporates the lines spoken in those events into the feature set.

Question 4

The grid search for the best parameters (pre-processing and feature extraction techniques) involved looking at every combination of techniques (a total of 24 combinations). The results (see table 3 or code block 21 in the notebook) that simple token count features consistently yield the best results in terms of mean rank and accuracy, regardless of the pre-processing technique used; although this feature does work best when lemmatisation or stemming are used in pre-processing.

	Preprocessing Technique	Feature Type	Mean Rank	Mean Cosine Similarity	Accuracy
0	Basic Whitespace Tokenisation	Simple word count	4.6	0.970354	0.4
1	Basic Whitespace Tokenisation	TF-IDF	4.4	0.950503	0.4
2	Basic Whitespace Tokenisation	N-Grams	4.8	0.531516	0.1
3	Basic Whitespace Tokenisation	POS Tagging	5.4	0.951395	0.1
4	Remove stopwords and punctuation	Simple word count	3.9	0.979640	0.3
5	Remove stopwords and punctuation	TF-IDF	4.2	0.954173	0.2
6	Remove stopwords and punctuation	N-Grams	4.9	0.539444	0.1
7	Remove stopwords and punctuation	POS Tagging	4.8	0.975152	0.2
8	Stem of words	Simple word count	3.8	0.978641	0.3
9	Stem of words	TF-IDF	4.1	0.955509	0.3
10	Stem of words	N-Grams	4.9	0.541228	0.1
11	Stem of words	POS Tagging	5.5	0.975783	0.1
12	Lemma of words	Simple word count	3.8	0.979483	0.3
13	Lemma of words	TF-IDF	4.3	0.954065	0.2
14	Lemma of words	N-Grams	4.9	0.536428	0.1
15	Lemma of words	POS Tagging	4.7	0.976095	0.2
16	Lowercasing	Simple word count	4.1	0.970827	0.5
17	Lowercasing	TF-IDF	4.2	0.953874	0.4
18	Lowercasing	N-Grams	5.0	0.541618	0.1
19	Lowercasing	POS Tagging	4.7	0.948832	0.3
20	Uppercasing	Simple word count	4.1	0.970827	0.5
21	Uppercasing	TF-IDF	4.2	0.953874	0.4
22	Uppercasing	N-Grams	5.0	0.541618	0.1
23	Uppercasing	POS Tagging	4.4	0.967708	0.3

Table 3 Results of every combination of pre-processing and feature extraction technique.

Question 6

Running the system on the best parameters (lemmatisation for pre-processing and simple token counts for features) yields a slightly worse performance on the test set than the default parameters. For the test set the mean rank is 4.1 with an accuracy of 0.2 (see code blocks 22 and 23 in the notebook). This suggests that while these parameters may have worked best on the validation set, they fail to capture some intrinsic properties of the test set and may in fact lead to overfitting on the training and validation sets.