ECS763P/U Assignment 2: Distributional Semantic Similarity Based IR within EastEnders Characters (40%)- Markscheme

Please refer to the example file `Assignment_2_distributional_semantics_solutions` for an example of a good piece of coursework showing evidence of all the below.

Evidence of completion will be primarily taken from the notebook, however, where feasible the accompanying 2-page report will provide evidence, particularly for the marks which include analysis or observation.

1 Improve pre-processing (10 marks)

- 1 mark each for evidence of having tried any of the following (or other good pre-processing technique): removing punctuation/other characters, converting to lowercase, removing stopwords, applying lemmatization/stemming **(4 marks max)**
- Clear showing/description of results on validation data from more than one setting **(3 marks max**) (1 mark for one setting))
- Systematic exploration showing selection of best set of techniques, showing systematic improvement **(3 marks max)**

2 Improve Linguistic Feature Extraction (15 marks)

- 2 marks each for evidence of having tried any of the following (or other good feature extraction): minimum document frequencies (mdf) threshold for words with different values for mdf, different n-gram feature extractions, POS tags, language model features, dependency/constituency parse features, Feature selection/reduction, gender/sentiment classification **(8 marks max)**
- Systematic exploration showing selection of best set of techniques leading to improvements over the baseline (**3 marks max)**
- Evidence of feature analysis of some kind, through feature ranking or other method with some discussion **(4 marks max)**

3 Incorporating context features (15 marks)

- Successful incorporation after appropriate extra information from the dataframe/file into the pre-processed text appropriately other than the line itself (e.g. previous line, next line and/or scene information) **(5 marks max)**
- Incorporation of the extra features into the feature set appropriately, with some systematic selection of the features, possibly reselecting all hyperparams again **(5 marks max)**
- Evidence of feature analysis of some kind, through feature ranking or other method with some discussion of the utility of the features with observations **(5 marks max)**

Mean rank range	points
1.0 <=1.1875	3
1.1875<=1.375	2.5
1.375<= 1.5	2
1.5<=1.625	1.5
1.625<=1.75	1
1.75<=2.0	0.5

Table 1: Performance points for Q4 for validation and test tests.

4 Improve the vectorization method (10 marks)

- Successful implementation of TF-IDF or other vectorization method **(6 marks**) (3 marks for partial success)
- Note/analysis of the improvements/effect on the classification task on validation data with clear demonstration of improvement over Q3-4 **(4 marks)**

5 Select and test the best vector representation method (10 marks)

- Final setting clear and runnable with final results on test data presented at the bottom, showing the mean rank (**2 marks** (1 mark if not clear how the best setting used on the test data was arrived at))
- Further exploration to improve the method, for example, re-running feature selection and/or pre-processing in combination with the vectorization technique (**2 marks max**)
- Performance points of the best selected setting on validation set as per Table 1. (**3 marks max**)
- Performance points of the best selected setting on test set as per Table 1. (**3 marks max**)