Group 22
João Antunes – nº 87668

Question Classification Report – Natural Language MP2

1. Models

In this project were implemented 3 different models. Before all the models receive the questions and answers from the train-file and the test-file, every question and answer are inserted into the same string and are then pre-processed by removing stop-words, eliminating all non-alpha-numeric characters (using the regular expressions python package "re"), lowercasing, and lemmatizing every remaining word of a question. After all questions and answers have been pre-processed, every question and answer from the train-file and test-file and every topic from the train-file are then inserted into 3 respective lists. All models then receive these lists as inputs.

The first model is based on the Jaccard Similarity Distances algorithm. each test-file question is compared to all train-file questions and answers. Each comparation returns a Jaccard Similarity score and when the highest score is found the most likely topic will be printed to the output file, which is the topic associated with the train-file question and answer. This model was used as the baseline for this project.

The second model is based on a Naïve Bayes Classifier, a CountVectorizer and a TF-IDF transformer that are inserted as parameters into a Pipeline. The Pipeline is then trained with all questions and answers and topics from the train-file and the return of this training is then used to predict the topics of all questions and answers from the test-file and print them to the output file.

The third model is an extension of the second model, the only difference is that after the Pipeline is created, it is performed a Grid-Search on the Pipeline which makes it possible to find the most optimal parameters: Which n-gram range to use, whether the TF-IDF transformer should be used or not and what classifier alpha value should be used. The return of this Grid-Search is then trained with all questions and answers and topics from the train-file and the return of this training is later used to predict the topics of all questions and answers from the test-file and print them to the output file, as referenced before. This model is the final model of this project

2. Experimental Setup

The evaluation measure used to compare the previous models was the accuracy of correct topics. For this it was considered the first model (Jaccard) as the baseline. It was also considered the accuracy percentages marks to beat that were given in the project description statement (the 68% mark and the 80% mark).

3. Results

Questions and answers	Jaccard (Baseline)	Naïve Bayes Classifier with CountVectorizer and TF-IDF	Naïve Bayes Classifier with CountVectorizer, TF-IDF and Grid-Search (Final Model)
Geography	75.00%	52.50%	90.00%
Music	78.18%	86.36%	91.82%
Literature	81.45%	86.29%	88.71%
History	75.36%	92.03%	88.41%
Science	78.41%	72.73%	89.77%
General	78.00%	82.80%	89.60%

From the previous results it is possible to conclude that the baseline was able to beat the 68% accuracy mark, but it did not beat the 80% mark. The Naïve Bayes Classifier with CountVectorizer and TF-IDF model showed better results than the baseline and it did beat the 68% accuracy mark and the 80% accuracy mark. Finally, the Naïve Bayes Classifier with CountVectorizer, TF-IDF and Grid-Search model was able to beat both accuracy marks and showed the best accuracy results of all three models.

4. Error Analysis

After carefully analyzing the questions in which a wrong topic was attributed, I reach to the conclusion that one of the most common errors is when some questions and answers have more specific words that are not as frequent in other questions. Since the final model takes into consideration weights on how frequently a word appears, this makes it more difficult to attribute the correct topic to the provided questions and answers because it varies from that the model was trained with. One example of this is the music question "...& the Blackhearts" with the answer "Joan Jett". The final model classified this question and answer as a history question. In this case "...& the Blackhearts" is part of a name of the band "Joan Jett & the Blackhearts" and these words are very specific to the music topic and do not appear in any other question, leading to the model not having enough data to classify this question correctly. Another common error is due to some questions and answers having words that are very frequent in questions and answers of other topics, and it becomes ambiguous to the classifier to predict the correct topic. For instance, history questions and answers frequently have years, dates, or locations, and it is more likely to the model to classify questions and answers that contain these words as a history question and answer. This might lead to wrong topics classifications. One example of this is the music question "December 8, 1980 in New York City" with the answer is "John Lennon" in which the classifier incorrectly identifies as a history question and answer.

5. Future work

If I was given more time to improve my final model, I would try to extend it by adding a Rule-Based approach and print its results. When this approach would not be able to identify the topic, I would then print the result returned by the Naïve Bayes Classifier with CountVectorizer, TF-IDF and Grid-Search classifier.

6. Bibliography

https://www.nltk.org/py-modindex.html
https://scikit-learn.org/stable/user_guide.html