MP2 - Question Classification - Group 44

André Oliveira - 93686 Maria Beatriz Venceslau - 93734

1. Models

The pre-processing we did for each of the models we developed (similar to TF-IDF and Naive Bayes) was similar, in both we applied lemmatization, tokenization, lowercasing, removal of punctuation.

In the TF-IDF model, we removed all English stop words and all numbers; if there were two labels in one line, we added the sentence to both labels on the training set.

In the Naive-Bayes model, we removed the stop words “the”, “this”, “in” and “a”; if there were two labels in one line, we chose the one with higher frequency in the training set and lastly, the model transforms numbers with 3 or more digits into a special string “data_format”.

The first model we developed was based on the TF-IDF algorithm. We first tried using the standard TF-IDF method, but we only achieved 60.6%. Then we tried to sum all the TF-IDF values for each token by label and choosing the label with the best score achieving better performance as we can see in the Results section.

The other model we developed was a combination of CountVectorizer with Naive Bayes. We used CountVectorizer to build a matrix with all the features (words) that the sentence had. Then, we gave this matrix to the Naive Bayes to train. Finally, we built another matrix with the dev-test to predict the label for each sentence.

2. Experimental Setup

The evaluation measures we used were accuracy/recall, precision and F1 score. We used the trainWithoutDev.txt to train our models or to use it as a database and the dev_clean.txt to test the accuracy of each model. When developing our project, we started by implementing the TF-IDF algorithm and thus defining our baseline with an accuracy of 77.6%. In the Naive Bayes model, we tried a big variety of pre-processing combinations. We tried different combinations of stopwords, using or not lowercasing, the removal or not of punctuation, the use of lemmatization and our special string for dates. Besides this, the sklearn library had 4 types of Naive Bayes, we explored and found out that ComplementNB was the best.

3. Results

3.1.TF-IDF (Baseline):

In this model, we can see that the label that was identified with better accuracy was the History label, and the one with the least accuracy was the Geography label.

Metrics	General	History	Science	Music	Literature	Geography
Accuracy/ Recall	77.60% (388/500)	91.30% (126/138)	60.23% (53/88)	79.09% (87/110)	84.68% (105/124)	42.50% (17/40)
Precision	-	0.62 (126/203)	0.98 (53/54)	0.85 (87/102)	0.87 (105/121)	0.85 (17/20)
F1	-	0.74	0.75	0.82	0.86	0.57

3.2.NB Model:

In the Naive-Bayes model, we can see that the label that was identified with better accuracy was the Science label, and the one with the least accuracy was the Music label.

Metrics	General	History	Science	Music	Literature	Geography
Accuracy/ Recall	90.20% (451/500)	84.06% (116/138)	96.59% (85/88)	89.09% (98/110)	92.74% (115/124)	92.5% (37/40)
Precision	-	0.89 (116/131)	0.91 (85/93)	0.92 (98/106)	0.93 (115/124)	0.80 (37/46)
F1	-	0.86	0.94	0.91	0.93	0.86

Having suspected that our results had an accuracy that was too high we decided to perform cross-validation within the training dataset, to see if the decisions we had made during the development of our project were influenced by our objective of reaching a higher accuracy value in the development dataset. The average accuracy obtained by this method was 85%(+/-0.02).

4. Error Analysis

4.1. Naive Bayes

After analyzing the results obtained, we can verify that some questions only had 1 word that matched the ones on the training dataset.

Sentence and Answer	Correct Label	Obtained Label	Explanation
...& the Mechanics - Mike	MUSIC	LITERATURE	The matched word was “Mike” which was found on 6 LITERATURE questions and only 2 of MUSIC
...& the Blackhearts - Joan Jett	MUSIC	HISTORY	The matched word was “Joan” which was found on 10 HISTORY questions and only 4 of MUSIC

In other cases, several words matched the ones on the training dataset. Some of these words are very common on the dataset. Sometimes, just like in the example, one word may appear so much that it can minimize the importance of other words.

Sentence and Answer	Correct Label	Obtained Label	Explanation
"This first host of "The Tonight Show" emceed "Meeting of Minds" for PBS in 1977" - Steve Allen	HISTORY	MUSIC	Many of the words point towards HISTORY (first, data_format, meeting) but with a lot of matches with other labels. PBS has 2 labels for HISTORY and MUSIC each. And then “tonight” and “show” almost only appear in MUSIC.

4.2.TF-IDF

After analyzing the results obtained, we can see that since some words suffer from ambiguity depending on the context, and the model can not differentiate the context, it will choose the label where the word occurs more often.

Sentence and Answer	Correct Label	Obtained Label	Explanation
The Babylonians kept abreast of the times using a form of this instrument seen here: - Sundial	SCIENCE	MUSIC	Given the TF-IDF model calculates the importance of a word in a set of documents based on the number of documents the word appears divided by the total number of documents (with that label). The word “instrument” was found 158 in MUSIC and only 7 in SCIENCE, so it was given less priority.
"His 1543 book ""Concerning the Revolutions of the Celestial Spheres"" started an astronomical revolution" - Nicholas Copernicus	SCIENCE	LITERATURE	In this case, the word “book” was found 168 times in the LITERATURE category and only 9 in the SCIENCE category.

5. Future Work

To prevent the OverFitting situation described in the Results section, we could use ensemble models. This way, we could use different models to predict the results, avoiding the generalization error of the prediction.

Using the same type of approach to the project, the use of a larger training dataset would improve the accuracy of the models, given that they would have been exposed to a larger set of words and sentences, thus accumulating more knowledge to compare to the test dataset.

If, on the other hand, we took a different approach to the project, by developing models based on neural networks, the accuracy of the results would improve drastically, given they would have the ability to learn from the training datasets.

We can conclude that both approaches combined would improve the results even more.