Question Classification Report – NL MP1

Joana Coutinho – nº 87666
Maria Inês Morais – nº 83609
Group 19

Approach: Hybrid (Stochastic Gradient Descent (SGD) classifier and Rule-based)

1. Description of the model

Our initial approach was to try different combinations of the simpler types of methods mentioned throughout the course, such as Jaccard distances, MED and Rule-based, to use as a baseline for the project, which eventually led to the following model:

Initially the questions from the training and development set are pre-processed: the appropriate stop-words are removed, WordNetLemmatizer is applied and every remaining word is lowercased. The pre-processed questions from the training set are then transformed into a matrix of token counts. With the help of a Grid-Search, we are able to find the optimal parameters for the developed model, among these are the option to use or not use tf-Idf, and which range of N-grams to apply. Subsequently the SGD classifier is trained with the pre-processed questions and labels from the training set. Finally, if the questions do not match a pattern in the Rule-based classifier, they are pre-processed and fed to the SGD classifier.

2. Accuracy results

Model	Jaccard and MED	Jaccard, Med and Rule Based	Hybrid
Coarse-grained	71.5%	75.18%	86.96%
Fine-grained	59.41%	62.48%	81.91%

3. Error Analysis

When performing an analysis to the questions that were incorrectly classified by our model, we come to the conclusion that it becomes harder to correctly classify questions that are very specific and use terms that are rarely seen in the training set. Examples are: “What shampoo prevents eczema, seborrhea, and psoriasis?” or “What canal does the Thatcher Ferry Bridge span? ”. Since the classifier uses a set of weights to represent the terms, it becomes harder for the model to properly identify questions that vary from what it was trained with.

In conclusion, even though the developed model classifies some questions incorrectly, it was the approach that allowed us to achieve the best results, after trying out different solutions. We also recognize that this remains a Natural Language problem without a precise solution and hence an error-free model.