APPROACH: Similarity Measures

1. Description of our model

This project required us to build a model to classify questions according with a coarse and a fine-grained questions taxonomy, from Li and Roth. Our model is considered a hybrid, sense we decided to use **N-grams and Similarity Measures.**

To start we used some knowledge of **regex** to split the labels in coarse and fine. Then, for each question we used **tokenization** to create a list of tokens, we removed the proper names, and with **lowercasing** we made all tokens lowercase. We also used **stemming** to reduce words to their word stem.

Subsequently, we used **N-grams** to make sequences of 2 tokens in each label. Only then we removed all of the tokens that were considered **stop-words.** Regarding stop-words we created a list called **question_words** that contains words that are normally used to start questions, which we remove these words from the list of **stop_words.** We also filtered **ngrams where both tokens were stop words**, which we noticed they were very common and didn't help on the classification and could even be misleading.

After this, we had a list of questions' tokens for each label (a label is, in our project, composed by both the coarse and the fine), which we will use to compare to the questions whose labels we need to predict. To do this we decided to use **Jaccard,** which is a measure of how dissimilar two sets are. If a question in the train set has the least **jaccard distance** with a current question from the test set, the model will attribute the label of that train set question to the current question from the test set.

2. Accuracy results from evaluation our model in the development set

Fine accuracy: 73.61 %
Coarse accuracy: 81.07%

3. Short error analysis

Most of our errors are attributed to the fact that the project is not prepared to check the **semantics** of the words. For example, in question 4, "What credit card features a centurion on its face ?", our project classifies it as *LOC:other* due to the fact that the question with lowest Jaccard distance is "LOC:other What London museum features a Chamber of Horrors ?", even though it's true label is *ENTY:other*.

Another contributing factor to this error is the fact that because we remove capitalized words, some questions' tokens become sets of **Bigrams of words that normally would be considered Stop Words**, which due to the Jaccard's calculations makes these sets really likely to have the lowest distance and be considered a match. For example we have a question who after being processed becomes the following set of tokens: {"what", "what is"}, which can make a lot of questions be associated with this question's label. Perhaps if we used **feature extraction** we could try to attribute some kind of semantic recognition to our classification model.

4. Bibliography

Slides / sebentas from the course's page.
http://www.nltk.org/_modules/nltk/model/ngram.html
https://www.nltk.org/_modules/nltk/metrics/distance.html
https://www.nltk.org/_modules/nltk/stem/snowball.html

Miguel Mota Nº 90964
Rafaela Timóteo Nº 90773