Grupo 16
92513 - Mafalda Ferreira | 92546 - Rita Oliveira

1. Descrição dos modelos

O primeiro modelo baseia-se no *Jaccard*, no *Stemming* e no filtro das *Stop Words* e sinais de pontuação. Os *tokens* são retirados das perguntas e são removidas as *Stop Words* e os sinais de pontuação não necessários nas perguntas e respostas. De seguida, realiza-se o *stemming* nas palavras. Por fim, usa-se o *Jaccard* para encontrar a pergunta no treino mais semelhante à pergunta no teste. O modelo previne que a pergunta do teste fique com a etiqueta da pergunta do treino.

O segundo modelo tira proveito do pré-processamento do filtro das *Stop Words* e sinais de pontuação já desenvolvidos. Este modelo baseia-se na combinação dos dois conceitos *Linear Support Vector Classifier* e *Count Vectorizer*, sendo que este último recorre a uma escala de *ngram's* entre um e dois, considerando, por isso, *unigram's* e *bigram's*.

2. Experimental Setup

Para este projecto têm-se em conta duas *baselines*: uma é o primeiro modelo que foi descrito na secção anterior e a outra é baseada no *Count Vectorizer* e no *Support Vector Classifier.* O segundo modelo vai adotar algumas técnicas que foram utilizadas nas baselines. Para se avaliar e comparar os modelos utilizou-se o seguinte *evaluation measure: accuracy*. A *accuracy* foi calculada através da biblioteca *sklearn.*

O primeiro modelo foi desenvolvido utilizando-se o Jaccard, *Stemming* e o filtro apenas das *Stop Words*. Verificou-se que havia sinais de pontuação que não eram importantes para a classificação das perguntas. De seguida realizou-se também o filtro dos sinais de pontuação. Com este novo filtro, o modelo ficou com 79,2% de *accuracy*. Tendo uma boa *accuracy*, decidiu-se adaptar o *stemming* e o filtro para o segundo modelo.

Posteriormente, foi desenvolvido um segundo modelo. Optou-se por recorrer ao *Count Vectorizer*, responsável pela criação de uma matriz de frequências, em que cada coluna representa a contagem do número de ocorrências de cada *token* para cada frase do conjunto de treino atribuída por linha. Posteriormente, era pretendido um algoritmo capaz de suportar a classificação de múltiplos rótulos, pelo que se optou pelo *Support Vector Classifier*. Este algoritmo permite através de um conjunto de treino maximizar a separação entre os vários conjuntos de dados estabelecendo vários planos num espaço N-dimensional, em que N corresponde ao número de rótulos de classificação.

Considerando o modelo anterior como *baseline*, foi aplicado um *tokenizer* responsável pela geração e processamento de *tokens*, adotando *stemming* e o filtro de *Stop Words* e sinais de pontuação do primeiro modelo. A definição de uma escala de *ngram's* entre um e dois permitiu considerar sequências de uma palavra (*unigram's*) e duas palavras (*bigram's*). Por último, a utilização de um *Linear Support Vector Classifier* permitiu uma melhor maximização da separação dos conjuntos de dados.

3. Resultados

De seguida, apresenta-se a tabela relativa aos resultados dos modelos desenvolvidos:

- Primeira *Baseline* (Primeiro modelo): *Jaccard* + *Stemming* + Pré-processamento;
- Segunda *Baseline* (Segundo modelo): *Counter Vectorizer* + *Support Vector Classifier;*
- Modelo Final (Baseado no segundo modelo): *Count Vectorizer* + *Linear Support Vector Classifier* + *ngrams* + pré-processamento.

Classificador / Resultados	***Accuracy* Global (%)**	***Accuracy* por Rótulo (%)**				
		GEOGRAPHY	***MUSIC***	***LITERATURE***	***HISTORY***	***SCIENCE***
Primeira *Baseline*	79.2	87.5	82.7	84.7	71.0	76.1
Segunda *Baseline*	81.6	77.5	78.2	81.5	87.7	78.4
Modelo Final	89.2	77.5	92.7	92.7	84.8	92.0

4. Análise de erro

A partir da tabela, é possível inferir que o modelo final apresenta melhores resultados. No entanto, não é possível obter uma classificação completamente correta, pois a separação de conjuntos a partir de vários planos não é ideal. Adicionalmente, observa-se que os rótulos ***HISTORY*** e ***GEOGRAPHY*** revelam valores de *accuracy* inferiores a 90%.

Em relação ao rótulo ***GEOGRAPHY***, verifica-se que este apresenta a menor *accuracy*. Este resultado deve-se ao facto de o modelo classificar a maioria das frases incorretamente como ***HISTORY*** nas perguntas 22, 25, 64 (*There are first-person accounts of this volcano's eruption **in 1767** (as there had been **in 79** A.D.) Vesuvius*), 184, 195, 266 e 316. Na frase 64, a referência temporal é responsável pela classificação incorreta, pois no conjunto de treino, estes *bigrams* foram referenciados em frases relativas a ***HISTORY***. Por outro lado, também se verificou a classificação incorreta inversa nas perguntas 42, 84, 300, 353 e 375.

Relativamente ao rótulo ***HISTORY***, verifica-se que a maioria das perguntas classificadas incorretamente têm este rótulo. Após várias análises, conclui-se que a razão pela qual o modelo classifica incorretamente as perguntas como ***HISTORY*** foi que durante a fase de treino do modelo este indicou uma elevada frequência de algumas palavras e sequência de duas destas nas perguntas classificadas com ***HISTORY***. Por exemplo, a pergunta 184 (***Alexander the Great** led his troops as far east as the Hyphasis, now the Beas, river in this country*) contém três palavras referentes à figura histórica ***Alexander the Great***. Verifica-se no conjunto dos treinos que estas palavras e a sua sequência se encontram na maioria das perguntas de história. Assim, durante a fase de teste, na qual o modelo encontrou estas palavras nas perguntas, classificou-as como ***HISTORY.***

Em suma, devido à avaliação das palavras nas perguntas durante o treino do modelo, o surgimento de determinadas palavras e a sua sequência na fase dos testes originaram uma classificação errada das perguntas.

5. Trabalho futuro

Como descrito na secção anterior, um conjunto de palavras e a sequência destas influencia a classificação das perguntas. Deste modo, existem muitas perguntas incorretamente classificadas como ***HISTORY*** e outras incorretamente não classificadas como ***GEOGRAPHY***, pelo que contribuem para a diminuição da *accuracy* global. Se existisse mais tempo, ir-se-ia investigar como aumentar a *accuracy* destes rótulos, para além de diminuir as perguntas classificadas incorretas como ***HISTORY*** e aumentar a classificação correta do rótulo ***GEOGRAPHY.***

6. Bibliografia

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
https://realpython.com/nltk-nlp-python/
https://www.statology.org/jaccard-similarity-python/