Grupo **72**
89404, Alexandre Antunes Rodrigues

Língua Natural 2021/2022
98702, João Luís Ambrósio Limeiro

Introdução

Este trabalho corresponde ao projeto realizado no âmbito da cadeira de Língua Natural e tem como objetivo simular um sistema de avaliação que permite a atribuição de etiquetas (ciência, literatura, música, história e geografia) a *features*, questões e respostas. Para resolver o problema foram utilizados dois modelos, o ***CountVectorized*** concomitantemente com o ***Multinomial Naive Bayes*** e o ***Tf-idfVectorized*** simultaneamente com o ***Multinomial Naive Bayes***.

Modelos

Para ambos os modelos foi realizado um pré-processamento. Este consiste em transformar dados não tratados num formato compreensível. Para a realização do mesmo foram realizados os procedimentos presente na figura.

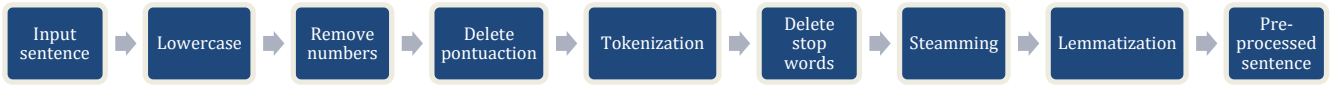

Fig.1 Pré-processamento de texto

O primeiro modelo realizado foi o ***CountVectorized*** concomitantemente com o ***Multinomial Naive Bayes***. O ***CountVectorized*** é usado para transformar texto num vetor com base na frequência de cada token. Isto é, cada posição representa o número de vezes que cada token aparece no texto.
O ***Multinomial Naive Bayes*** é um modelo de classificação estatístico baseado no teorema de Bayes. Este assume que cada *feature* (questão e resposta) é independente das restantes.
É constituído por *feature vectors* que representam as frequências com as quais certos eventos são gerados por uma distribuição multinomial $(p_i,..., p_n)$ onde p_i é a probabilidade de um evento i ocorrer ou k quando segue uma distribuição multinomial. Assim, um *feature vetor* $x = (x_1, \ldots, x_n)$ é um histograma onde x_i é o número de vezes em que um dado evento foi observado. A probabilidade de observar um histograma x é dada pela seguinte fórmula:

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} {p_{ki}}^{x_i}$$

Fig.2 Fórmula matemática Multinomial Naive Bayes

O segundo modelo, contrariamente ao primeiro, em oposição ao uso de ***CountVectorized*** usa ***Tf-idfVectorized*** juntamente com ***Multinomial Naive Bayes*****.** O ***Tf-idf*** *(term frequency–inverse document frequency)* é uma medida estatística que avalia a acuidade de uma palavra para um documento em uma coleção de documentos. Isso é feito multiplicando-se duas métricas: quantas vezes uma palavra aparece em um documento e a frequência inversa da mesma em um conjunto de documentos.

Experimental Setup

O projeto de classificação desenvolvido tem duas fases. A fase de treinamento do modelo e a fase de teste do mesmo. A performance é avaliada com base na *accuracy*, *f1-score*, precisão e *recall*.

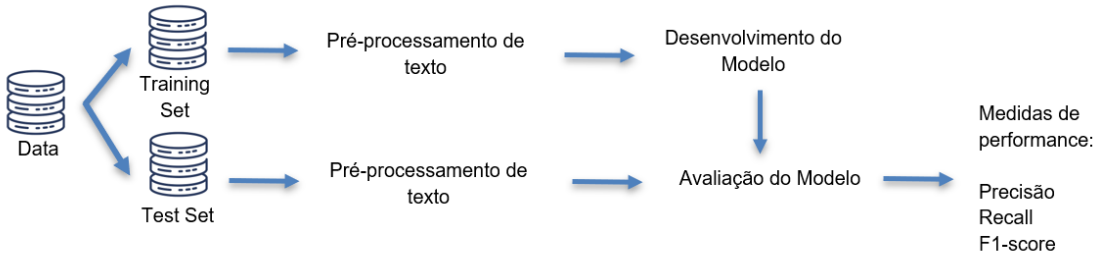

Fig.3 Experimental setup

A baseline usada tem como ponto de partida a utilização do modelo **Tf-idf Vectorized** juntamente com **Multinomial Naive Bayes**.

Resultados

Baseline (Tf-idf Vectorized juntamente com o **Multinomial Naive Bayes)**

	Precisão	Recall	f1-score
SCIENCE	0.98	0.69	0.81
GEOGRAPHY	0.83	0.96	0.89
HISTORY	0.93	0.96	0.95
LITERATURE	0.95	0.95	0.95
MUSIC	0.99	0.85	0.91

Accuracy: **83.4%**

Count Vectorized juntamente com o **Multinomial Naive Bayes**

	Precisão	Recall	f1-score
SCIENCE	0.96	0.90	0.93
GEOGRAPHY	0.92	0.92	0.92
HISTORY	0.93	0.96	0.94
LITERATURE	0.95	0.96	0.95
MUSIC	0.94	0.93	0.94

Accuracy: **87.2%**

Análise de Erros

Fig.4 Histograma dos pares de categorias confundidas

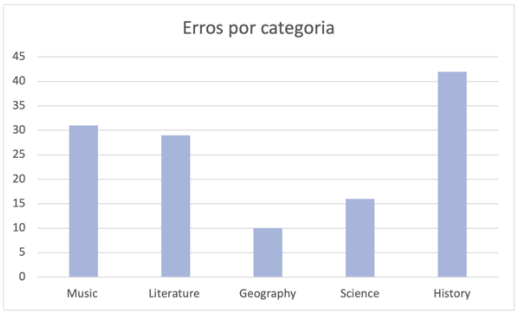

Fig.5 Histograma com o total de erros por categoria

O histograma à esquerda demonstra os pares de categorias que mais foram trocados pelo modelo, por exemplo ao indicar incorretamente *Literature* em vez de *History* ou vice-versa.
Ao analisá-lo podemos verificar que os pares *History* e *Literature* são os mais confundidos, com 17 ocorrências, por outro lado, Geography só se confunde com *History e Music, isto significa que conseguimos distinguir as categorias relativas a Geography*.
O modelo não tem em conta o significado das palavras nem o seu contexto o que dificulta a identificação correta de categorias com palavras semelhantes como *History* e *Literature*, por exemplo nesta questão *"This wild west "Belle" served time in prison for horse theft in 1883"* o nosso modelo prevê que é *Literature* pois pensamos que identificou um livro ou um filme passado no wild west porém faz parte da categoria *History*. No entanto, não houve confusão de *labels* em que as categorias eram distintas como por exemplo *Geography* e *Science o que demonstra que se contextualizar-mos as palavras podemos aumentar a precisão*.
Neste projeto os falsos negativos não são tão importantes como por exemplo num modelo em que estuda transações fraudulentas ou análises em pacientes, portanto demos mais peso à accuracy e precisão quando estávamos a estudar os melhores modelos.

Trabalho futuro

Estamos contentes por alcançar uma accuracy superior a 85%, mas sabemos que seria possível chegar aos 90% por exemplo se contextualizássemos as palavras.
O próximo passo será implementar redes neuronais aprendidas nas aulas, o que deverá aumentar bastante a accuracy e precisão da nossa categorização.

Bibliografia

https://pt.wikipedia.org/wiki/Pr%C3%A9-processamento_de_dados
https://en.wikipedia.org/wiki/Naive_Bayes_classifier
https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
https://scikit learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html