Laboratório Nº 5: Text Mining - Bag of Words Model: Seleção de Características

Extração Automática de Informação
2022/2023

Prof. Joaquim Filipe
Eng. Filipe Mariano

Objetivos

- Seleção de *features*
- Identificação dos Top termos para cada métrica estudada
- Visualização desses Top termos através de tabelas ou gráficos

Exercícios

1. Crie uma função `sumVector` no módulo `bagOfWords.js` que recebe como parâmetro de entrada um array de objetos do tipo `Term` em que o `name` do termo é o mesmo em todo o array e devolve o somatório de todos os elementos das várias propriedades mencionadas anteriormente, excepto do `idf` (que se manterá o mesmo valor) e do `tfidf` (que será a multiplicação da soma dos `tf` pelo `idf`).

2. Crie uma função `avgVector` no módulo `bagOfWords.js` que recebe como parâmetro de entrada um array de objetos do tipo `Term` em que o `name` do termo é o mesmo em todo o array e devolve a média de todos os elementos das várias propriedades mencionadas anteriormente, excepto do `idf` (que se manterá o mesmo valor) e do `tfidf` (que será a multiplicação da soma dos `tf` pelo `idf`).

3. Exporte as funções `sumVector` e `avgVector` do módulo `bagOfWords.js` e utilize-as no módulo `train.js` na função `process` para poder chamá-la para cada termo da bag of words de textos positivos e negativos.

4. Repita todo o processo realizados neste laboratório, mas agora para bigramas de palavras.

5. Na diretoria `classification` criar um novo módulo `featureSelection.js` que deverá exportar uma função `selectKBest` que recebe como parâmetro de entrada (1) um array de objetos do tipo `Term`, (2) um número inteiro `K`, (3) a métrica (binário, número de ocorrências, tf ou tf-idf) e (4) se utiliza o vetor de somatório (utilizar este por *default*) ou médias, e devolve num array dos `K` melhores `Term` existentes nesse array de objetos passado como primeiro argumento.

6. Aplicar a função anterior ao processamento do conjunto treino que está a ser realizado de modo a obter as melhores `K` *features* para o `tfidf`. Por *default* vamos considerar o `K` como 10% do número total de termos únicos.

7. Gravar em base de dados ou em ficheiro os resultados dos termos e métricas calculadas para o conjunto de treino. Assim como, dos termos e métricas referentes aos selecionados pela função `selectKBest`. Sempre que o processamento do conjunto de treino é feito, estes resultados são recriados na base de dados (ou em ficheiro).