Laboratório Nº 5: Text Mining - Bag of Words Model: Tops e Visualização

Extração Automática de Informação
2021/2022

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

1. Exercícios

1. Na diretoria `features` criar um 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.

2. Aplicar a função anterior ao processamento do conjunto treino que está a ser realizado de modo a obter as melhores `K` *features* para as 4 métricas referidas anteriormente. Esses resultados deverão ser guardados em base de dados.

3. Crie uma página que permita visualizar as melhores `K` *features* para as 4 métricas referidas anteriormente, tanto para bigramas como para unigramas, para as classes `happy` e `not happy`. Deverá ser facilmente visível qual o termo, a posição (dentro do top `K`), o valor encontrado e o tipo de métrica (binário de unigramas, nº de ocorrência de unigramas, etc). Pode optar por apresentar numa tabela ou num simples gráfico de barras utilizando alguma biblioteca standard.

4. De modo a tornar a seleção de *features* mais dinâmica e flexível poderá dar a opção ao utilizador para selecionar o `K` para os melhores unigramas, assim como o `K` para os melhores bigramas. No final os melhores unigramas e bigramas selecionados e guardados em base de dados, deverão ser escolhidos os respetivos tf-idf para serem devolvidos num array de `Term` dos selecionados como *features* selecionadas do conjunto de treino.