Laboratório Nº 6: Similaridade do Cosseno

Extração Automática de Informação
2022/2023

Prof. Joaquim Filipe
Eng. Filipe Mariano

Objetivos

- Utilizar a similaridade do cosseno para determinar a que classe pertence um documento

1. Introdução

Uma abordagem comum utilizada para combinar documentos semelhantes baseia-se na contagem do número máximo de palavras comuns entre os documentos (distância euclidiana). Contudo, essa abordagem apresenta uma falha, que é à medida que o documemnto aumenta, o número de palavras comuns tende a aumentar mesmo que os documentos falem sobre tópicos diferentes. Neste sentido, a métrica de similaridade do cosseno é mais adequada para a identificação da similaridade entre documentos.

2. Métricas de Similaridade entre Documentos

No que diz respeito a um espaço vetor-modelo, criado a partir de uma matriz termo-documento, existem várias métricas que representam um valor de distância entre vetores(ou pontos) e, em alguns casos, um valor de similaridade. É de notar que nos casos em que a medida retorna um valor de distância, é possível converter esse valor para um valor de similaridade. Tipicamente, esta conversão é feita pela operação contrária, isto é, `similaridade = (1-distância)`, o que significa que quanto mais próximos os vetores(ou pontos) estejam no espaço, maior a similaridade entre os objetos a classificar.

Existem diversas medidas para o cálculo da similaridade sendo que duas das mais conhecidas são: a Distância Euclidiana e a Similaridade do Cosseno.

2.1. Distância Euclidiana

Esta medida calcula um valor de distância, não normalizado, entre dois pontos no espaço. Esta distância é equivalente ao comprimento do segmento de reta que une os respetivos pontos. É dada pela fórmula:

$$D(a,b) = \sqrt{\sum (a_i - b_i)^2}$$

A distância euclidiana tem em conta a magnitude do vetor. Em termos práticos, isto significa que, no exemplo do contexto de classificação de documentos por palavras, o valor da distância é influenciado pelo peso da palavra no documento. Esta medida é tipicamente usada em contextos em que os objetos a comparar tenham tamanhos equivalentes e os atributos dos mesmos tenham pesos equivalentes.

2.2. Similaridade do Cosseno

A similaridade do cosseno é uma métrica usada para medir a similaridade de documentos, independentemente do seu tamanho. Do ponto de vista matemático, mede o cosseno do ângulo entre dois vetores projetados num espaço multidimensional. Esta análise torna-se vantajosa porque mesmo que os dois documentos estejam distantes da distância euclidiana (devido ao tamanho do documento), é provável que possam na mesma ser similares. Quanto menor o ângulo entre vetores, maior será a similaridade do cosseno.

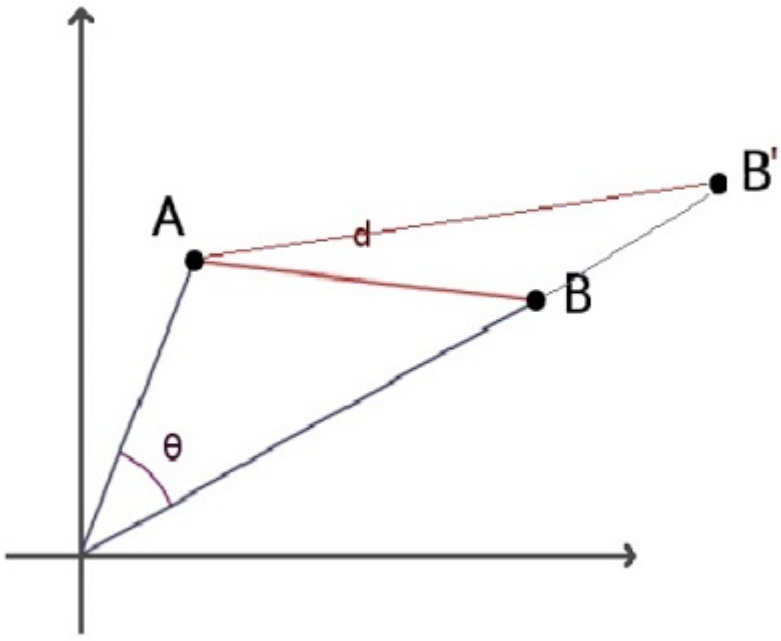
$$C(a,b) = \frac{A \cdot B}{||A||||B||}$$

ou

$$C(a,b) = \frac{\sum(a_i * b_i)}{\sqrt{\sum(a_i)^2} * \sqrt{\sum(b_i)^2}}$$

2.3. Diferença entre Distância Euclidiana e Similaridade do Cosseno

A figura abaixo demonstra a diferença entre as medidas da distância euclidiana e similaridade do cosseno num espaço bidimensional, onde d equivale ao segmento de reta traçado entre os pontos A e B e Θ igual ao valor do cosseno do ângulo entre os vetores.



Se considerarmos a distância entre A e B', observamos que, apesar de o peso dos valores ter aumentado, o cosseno do ângulo entre os vetores é constante enquanto o comprimento do segmento de reta usado no cálculo da distância euclidiana aumenta e por consequência a similaridade entre A e B diminui.

3. Exercícios

1. No módulo `featureSelection.js` crie uma função que leia os `K` melhores termos selecionados (N=1 e N=2) guardados em base de dados ou ficheiro por classe, pedido no último exercício do laboratório anterior.

Nota: De salientar que deverão ser guardados os termos e respetivo valores de `tfidf` obtidos através do somatório ou média de cada termo em todos os documentos do conjunto de treino, de uma determinada classe. Também o `idf` terá de ser guardado para todos os termos, visto que será necessário para poder calcular os valores dos termos em novos textos na fase de classificação.

Possível Exemplo:

```
[
    {
        label: "positive",
        bagofwords: [
            {
                "name": "best",
                "tfidf": 0.002,
                "idf": 0.001
            },
            (...)
        ],
    {
        label: "negative",
        bagofwords: [
            {
                "name": "worst",
                "tfidf": 0.002,
                "idf": 0.001
            },
            (...)
        ]
    }
]
```

2. Na diretoria `classification` crie um módulo `classifier.js` que exporte uma função `cosineSimilarity` que recebe como argumento um texto para ser classificado. O objetivo da função será:
a. Em primeiro lugar deverá ser feito todo o pré-processamento anteriormente programado, e que foi utilizado na fase de treino, para o texto recebido como argumento.
b. Calcular o `tf` para os termos encontrados no texto e que existem na bag of words de termos de cada classe.
c. Utilizar o `idf` de cada termo encontrado no texto e que existe na bag of words de termos de cada classe.
d. Calcular o `tfidf` de cada termo encontrado no texto e que existe na bag of words de termos de cada classe.

e. Pretende-se com esta operação utilizar os termos encontrados no texto processado e que existem na bag of words calculada durante o treino, para cada classe (`positive` e `negative`), para realizar o cálculo do `tfidf` de cada um desses termos. Ou seja, o `tf` é utilizado o que foi obtido para o texto agora processado e o `idf` o que foi guardado previamente.
f. Calcular a similaridade do cosseno para a classe `positive` e para a classe `negative`, utilizando como vetor A os valores obtidos no treino e como vetor B os valores calculados para o texto recebido como argumento. A classe cujo resultado da similaridade for maior será a que deverá ser retornada pela função, com o respetivo valor de similaridade.

3. No módulo de testes criado anteriormente na diretoria `test`, crie uma função para testar um conjunto de textos da classe `positive` e da classe `negative` para classificá-los utilizando a função criada no exercicio anterior. Deverá utilizar textos diferentes daqueles que utilizou para o conjunto de treino.