

Manual de Utilizador - Classificador de Texto

Instituto Politécnico de Setúbal - ESTS

Metrado Engenharia Informática - Extração Automática de Informação

António Carlos Ferreira Pinto

Diogo Costa

Guilherme Malhado



**POLITECNICO
SETUBAL**

Índice

1. [Introdução](#)
2. [Instalação](#)
3. [Guia de Utilização](#)
4. [Anexos](#)

Introdução

Uma aplicação web simples, maioritariamente de página única. É um classificador de textos, mais em específico reviews de comida, tendo duas classes para as reviews, reviews Positivas e reviews Negativas. Feito usando Node.js com Express.js, e uma base de dados em mysql

Instalação

Requisitos:

- Nodejs versão 18+

Correr a aplicação pela primeira vez:

1. Antes de correr o programa localmente criar um ficheiro .env usando o sample.env como base.
2. Num terminal na diretoria da aplicação correr:

```
npm i
```

3. No mesmo terminal correr para inicializar a aplicação em localhost, correr:

```
npm start
```

Guia de Utilização

Para inicializar inicializar o a aplicação correr:

```
EAI-Labs$ npm start
```

Estrutura da aplicação web:

A aplicação funciona maioritariamente numa única página

Food Reviews

Welcome to Food Reviews

Selects	Process	Class Identifier (Cosine Similarity)	Class Identifier (Naive Bayes)
GET /reviewFood/:id <input type="text"/> Submit	Process all Terms into BD. Already stored: 130996 Submit	Phrase Classification <input type="text"/> Submit	Phrase Classification <input type="text"/> Submit
GET /positiveReviews/:limit <input type="text"/> Submit	Process all Terms Statistics into BD. Already stored: 1365 Submit	Result	Result
GET /negativeReviews/:limit <input type="text"/> Submit		Matriz de confusão Generate	Matriz de confusão Generate

1

2

3

4

5



Composta por um titulo e 5 regiões distintas (enumeradas na image)

1. Selects - Selecionar dados do nosso dataset

Selects

GET /reviewFood/:id

Submit

GET /positiveReviews/:limit

Submit

GET /negativeReviews/:limit

Submit

Neste componente o utilizador pode facilmente fazer queries à base de dados para ver diferentes entradas do dataset utilizado para o nosso modelo.

Existem 3 alneas nesta região:

- GET /review/:id -> Endpoint que retorna a review com o id introduzido na textbox;
- GET /positiveReviews/:limit -> Endpoint que retorna X reviews classificadas como positivas (Score de 4 ou 5), X sendo o valor introduzido na textbox;
- GET /negativeReviews/:limit -> Endpoint semelhante ao anterior, mas para reviews negativas (Score de 1 ou 2).

2. Process

Esta região da aplicação é desaconselhável mexer, uma vez que qualquer um destes processos é inicializado irá demorar uma quantidade de tempo considerável.

Process
Process all Terms into BD. Already stored: 130996

Process all Terms Statistics into BD. Already stored: 1365

Nesta região temos dois botões:

- O primeiro para processar os todos os dados contidos no nosso training set, de forma a obter os termos para os nossos classificadores;
- O segundo para depois agregar e processar as componentes desses termos.

Podemos ver nesta região a quantidade de entradas criadas na nossa tabela de termos, sendo cada entrada a presença de um termo num documento; podemos também ver o numero de entradas na nossa table de estatísticas dos termos onde estão os nossos melhores termos agregados.

3. Class Identifier - Similaridade de Cosseno

O primeiro dos nossos dois algoritmos de classificação, podemos testar frases/reviews originais, ou ver os resultados dos testes do nosso classificador.

Class Identifier (Cosine Similarity)

Phrase Classification

Submit

Result

Matriz de confusão

Generate

Secção 1 - Testar Frases/Reviews originais

Nesta secção o utilizador pode testar uma review de comida original para ver se é identificada com positiva ou negativa. É só escrever uma frase na textbox e clicar no botão de submit ao lado. O resultado será a resposta se é classificado como Positivo ou Negativo e os valores ponderados (Similarity values) respetivos a cada Classe (por ordem Positiva e Negativa) onde podemos observar que o valor mais elevado é aquele que determina a classe estimada da frase.

Phrase Classification

Submit

Result

Negative

Similarity values: 0.37899712541652353,
0.40764192771392055

Secção 2 - Matriz de confusão - Testar o classificador

Nesta secção inicialmente temos só um botão para gerar os nossos testes, este quando primido gera a Matriz de confusão e as Métricas Prec, Rec e F1 (Ver anexo I para explicação de o que significa cada um) **A primeira vez que este processo é efetuado poderá demorar alguns minutos a calcular.**

Matriz de confusão

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP) 92	False Negative (FN) 108
Actual Negative	False Positive (FP) 88	True Negative (TN) 112

Métricas

Prec:

0.51111111111111110.5111111111111111

Rec: 0.460.46

F1:0.48421052631578950.4842105263157895

4. Class Identifier - Naive Bayes

O segundo dos nossos dois algoritmos de classificação, segue o princípio de utilização semelhante ao primeiro

Before and after

Class Identifier (Naive Bayes)

Phrase Classification

Result

Matriz de confusão

Class Identifier (Naive Bayes)

Phrase Classification

'Cough Medicine If you are loc

Submit

Result

Negative

Product of probabilities:
2.9452711296885496e-137,
2.6910416946697644e-131

Matriz de confusão

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP) 72	False Negative (FN) 128
Actual Negative	False Positive (FP) 48	True Negative (TN) 152

Métricas

Prec: 0.6

Rec: 0.36

F1:0.45

5. Swagger

Anexos

Anexo I

Matriz de confusão

Métricas

Prec -

Rec -

F1 -