IPBeja

INSTITUTO POLITÉCNICO DE BEJA

Sistema de apoio à promoção do turismo rural
Fase de Análise de Texto via Machine Learning

Gonçalo Amaro – 17440,
Pedro Tomás – 18962,
Vítor Abreu – 18966

XX de Janeiro, 2022

Conteúdo

1 Introdução

Após descrito os fundamentos de ETL e uma vez que a fase de estamos atualmente é a de transformação será necessário ser usado um tipo de processos bastante conhecido chamado de "extração de palavras chave" e "análise de sentimentos". Estes processos consistem:

- retirar do texto as palavras mais importantes que sejam capazes de descrever o estabelecimento.
- caracterizar o sentimento do texto, ou seja, saber se o texto é positivo ou negativo.

2 Análise de Texto

Uma Análise de Texto é um processo de extração de palavras chave e análise de sentimento. Basicamente os processos acima descritos. Mais especificamente a extração de palavras chave é feita através de um algoritmo de Machine Learning que é chamado de "bag of words" ou "bag of features"; e a análise de sentimento é feita através de um algoritmo de Machine Learning que é chamado de "Naive Bayes" ou "Naive Bayes Classifier", ou usando "transformers" ou "transformer" que é um algoritmo de Machine Learning que é chamado de "Tf-Idf" ou "Term Frequency - Inverse Document Frequency".

O ultimo caso foi descrito com duas opções porque como poderão verificar mais à frente, a nossa tentativa de "Naive Bayes" não foi bem sucedida, e por isso foi usado um "transformer" com um dataset chamado de "Bert" ou "Bert For Classification", este foi criado pela "Google".

Para a realização dos processos acima descritos usaremos a linguagem de programação "Python" e variadas bibliotecas. Das quais:

- **nltk** – Natural Language Toolkit.
- **sklearn** – Scikit-learn.
- **numpy** – Numpy.
- **pandas** – Pandas.
- **transformers** – Transformers.
- **torch** – PyTorch.
- **yake** – YAKE!.

Será feita a realização de um script por cada processo, que será transformado num notebook, via *p2j*, para que possa ser documentado.

2.1 Extração de Palavras Chave

Aqui foi executado o processo de extração de palavras chave, que consistiu em usar o YAKE!, que é uma biblioteca de autores portugueses e um japonês que foi criada para extrair palavras chave de um texto. Os seus autores são residentes em: Instituto Politécnico de Tomar, Universidade da Beira Interior, Universidade do Porto, INESC TEC e Universidade de Kyoto.

A razão especifica do uso desta biblioteca ao invés de fazer de raiz, foi o facto acima descrito, em nome de apoio ao uso de trabalhos de autoria portuguesa.

2.1.1 Processo de Execução

Este "script" funciona da seguinte forma:

- Indicamos o caminho da pasta onde se encontra o texto a ser analisado.
- Importa todos os .csv de reviews que se encontram na pasta.
- Por cada .csv de reviews, extrai as palavras chave.
- Exporta as palavras chave para um .csv.

2.1.2 Processo de Desenvolvimento

2.2 Análise de Sentimento

2.2.1 Processo de Execução

Este "script" funciona da seguinte forma:

- Indicamos o caminho da pasta onde se encontra o texto a ser analisado.
- Importa todos os .csv de reviews que se encontram na pasta.
- Por cada .csv de reviews, faz a análise de sentimento.
- Exporta os sentimentos para um .csv.

2.2.2 Processo de Desenvolvimento

3 Resultados

3.1 Extração de Palavras Chave

3.2 Análise de Sentimento

4 Conclusão

5 Webgrafia