Enunciado do TP2 - information retrival, Q&A, LLM, word-embeddings

J.João Filipe Cunha

Março 2024

Contents

Neste trabalho prático pretende-se explorar vários pontos, através da construção de pequenos programas / funções exploratórios que funcionem como prova de conceito.

Estamos cientes que não será possível explorar todas as variantes.

Pontos principais

base de conhecimento: Diário da República(DR)

Ficheiros em: https://natura.di.uminho.pt/~jj/spln2324/DR-spln

A base de conhecimento do diário da república é composta por

- um ficheiro de textos completos (8.5 GB, SQL) que contém duas tabelas:
  - dreapp_document – análoga aos metadados JSON
  - dreapp_documenttext – contem os textos do DR
    - o segundo campo (document_id) é uma chave estrangeira que corresponde ao `id` do Json e da tabela `dreapp_document`
- um ficheiro JSON de metadados correspondente (400 MB)

Information retrieval (IR)–

- dado uma expressão de pesquisa
- dada uma coleção de documentos – resumos presentes nos metadados do diários da República (Json de 400 MB) ou um subconjunto destes
- encontrar os documentos mais relevantes.
- Como? Algumas hipótese:
  - usando tf-idf (ver exemplo da aula `tfidf_similarity.py`)
  - usando word-embeddings (ver exemplo da aula `wmd_similarity.py`)
  - . . . ou outros algoritmos de doc. similarity para procurar documentos semelhantes

(No caso de utilizar Word-Embeddings criar um modelo)

Preparação do dataset de textos

Depois de encontrar os documentos mais relevantes, usando os metadados, pretende-se obter os textos correspondentes no ficheiro grande (8.5Gb sql)

Como lidar com a dimensão dos dados? Algumas hipótese:

- guardar cada texto num ficheiro separado no sistema de ficheiros.
- usar uma base de dados (Ex. SQLite, retirar o "public." da query SQL)
- ter o cuidado de garantir a ligação entre os metadados e os textos.

  O ficheiro de Textos têm uma chave-estrangeira que corresponde ao atributo `id` dos metadados Json.

Question and Answering (Q&A)

Pretende-se agora extrair informação dos documentos anteriormente selecionados:

- elaborar um conjunto de questões pertinentes.
- usar LLM (ver exemplo da aula `qa.py`), huggingface;
- procurar outros modelos.

Q&A + IR

- dada uma pergunta
- produrar os documentos mais relevantes
- ... e com base neles tentar construir uma resposta

ou seja, junção dos anteriores.