تقرير مشروع مادة

Information Retrieval System

عمل الطلاب

اسماعيل الرفاعي

لبنى الهندي

محمد سعادة

رهف حسون نصر

بإشراف المهندسة

لين قويدر

ال Datasets المستخدمة:

1- Lotte/lifestyle باللغة الانكليزية
2- Wikir/es13k باللغة الاسبانية

البنية البرمجية للتطبيق

يحتوي التطبيق على العديد من الصفوف Classes, و هي كالتالي:

1- Text Tokenizer

و هو عبارة عن صف مسؤول عن اجراء عمليات تقطيع النصوص و تحويلها الى مصفوفة من ال terms و ذلك عن طريق القيام بالعمليات التالية:

- Normalization
- Tokenization
- Stemming
- Lematization
- Remove stopwords

ثم يقوم باعادة مصفوفة من ال terms النهائية التي تعتبر هامة في النص المدخل، و يستخدم في عملية ال query و ايضا اثناء بناء ال Inverted Index.

2- Document

عبارة عن صف يستخدم لتمثيل ال documents الموجودة داخل ال dataset، حيث يحتوي على معلومات اضافية يتم ارفاقها بكل document و هي:

- Document id
- Original text
- List of document terms
- Vector to represent the document in VSM
- Counter for each term in the document

- Unique set of terms

3- Datasets

عبارة عن صف يستخدم في اجراء عمليات تحميل ال dataset من مكتبة ال ir_datasets وتمريرها الى صف ال InvertedIndex الذي بدوره يقوم بتحليلها.

اضافة الى انه يقوم بتخزين غرضين مختلفين من صف ال InvertedIndex كل واحد منهما يحتوي على dataset مختلفة.

يحتوي هذا الصف على تابع يقوم باخذ الاستعلام و اسم ال dataset و يقوم بتمرير الاستعلام الى ال dataset المناسبة للحصول على نتيجة الاستعلام.

يقوم ال api باستدعاء التابع السابق و يقوم التابع السابق باستدعاء تابع اخر في صف ال InvertedIndex

4- Inverted Index

هو الصف الرئيسي في التطبيق و يحتوي على ما يلي:

- data structure تقوم بتمثيل ال dataset الممررة له بشكل مناسب، و تخزين ال invertedIndex الخاص بكل term من ال terms الموجودة في ال documents الموجودة في ال dataset.
- بنى تقوم بتخزين القيم الرقمية للتوابع التالية (queries , VSM, TF_IDF, IDF, TF information).
- تابع يقوم بعملية بناء الinverted index و تعبئة البنى الاخرى بالمعلومات الرقمية الصحيحة، حيث يقوم بالمرور على الdocumenst كاملة و يقوم بمعالجتها واحدا تلو الاخر.
- تابع خاص باجراء استعلام على ال dataset حيث ياخذ الاستعلام ك parameter، يقوم بالبداية باجراء عمليات ال TextTokenizer على الكويري و تمثيلها بشكل Document (اي حساب قيم ال vector الخاص بال query) ثم نقوم بايجاد ال cosine similarity

بين الqueryو ال documents الموجودة، و نعيد في ال response كل ال documet التي تجاوز ال cos_sim الخاص بها عتبة معينة (0.3).

- بالاضافة الى تابع يقوم بعمليات التقييم الخاصة بالاستعلام حيث يقوم بمقارنة نتيجة النظام الخاص بنا بالنتيجة الصحيحة الماخوذة من مكتبة ال ir_datasets، اضافة الى انه يقوم بحساب المقاييس ال 5 الخاصة بالنظام و هي:
(precision, recall ,avp ,map, mrr).

اللغات البرمجية المستخدمة:

1- Front-end Application
   a. Javascript
   b. ReactJs
   c. TailwindCSS

2- Back-end Application
   a. Python
   b. FastAPI
   c. NLTK
   d. Ir_datasets
   e. Contractions

تقسيم العمل بين أعضاء الفريق:

اسماعيل الرفاعي backend application, Text processing

لبنى وسيم الهندي inverted index, query answering

رهف حسون نصر evaluation

محمد سعادة frontend application

المصادر:

janujaishree94/searchit-an-information-retrieval-system-@/https://medium.com
33d2af956da4