مشروع عملي نظم استرجاع معلومات 2023-2024

عنوان المشروع:

بناء information retrieval system مخصص لمجموعتي بيانات مختلفتين.

وصف المشروع:

مطلوب بناء محرك بحث يمكنه التعامل مع مجموعتين مختلفتين من البيانات وتمثيلهم وفق VSM وتنفيذ عمليات استرجاع البيانات وترتيبها والإجابة على استعلامات المستخدم.

على كل مجموعة اختيار Datasets 2 من الرابط التالي: https://ir-datasets.com

بشرط احتواء كل Dataset على أكثر من 200k Documents بالإضافة إلى Qrel (مجموعة استعلامات والنتائج الملائمة لكل استعلام).

يجب تحقيق ما يلي:

1. بناء وتحقيق نظام استرجاع معلومات يحتوي على الأقل:

- معالجة البيانات Data Preprocessing:
بعد تحميل مجموعات البيانات يجب معالجتها وفق ما تراه المجموعة مناسباً لمجموعات البيانات المختارة (Stemming, Lemmatization, Normalization, ..etc.).

- تمثيل البيانات Data Representation:
تمثيل الوثائق في كل مجموعة بيانات بطرق التمثيل المناسبة.

- الفهرسة Indexing:
بناء فهرس خاص بكل مجموعة بيانات لاسترداد المستندات بسرعة وبفعالية عالية واختيار ال indexing terms بكفاءة.

- معالجة الاستعلامات Query Processing:
معالجة الاستعلامات وتمثيلها بالطرق المناسبة.

- مطابقة الاستعلام وترتيب النتائج Matching & Ranking:
بناء تابع مطابقة الاستعلام مع المحتوى المشابه وترتيب المستندات المستردة بناءً على مدى صلتها باستعلام المستخدم.

- واجهة المستخدم UI (Web Application or Mobile Application):

بناء واجهة مستخدم سهلة الاستخدام تتضمن التالي:

i. اختيار مجموعة البيانات من الواجهة قبل تنفيذ الاستعلام وقبول استعلامات المستخدمين.

ii. عرض النتائج ذات الصلة من مجموعة البيانات.

ميزات إضافية للنظام

- Query refinement (query formulation assistance, query suggestions)
- Use advanced word embedding models
- Multilingual retrieval system
- Crawling
- Distributed information retrieval
- Conversational search (Chatbot)
- Documents clustering
- Personalization
- Use link analysis
- Topic detection
- Use vector stores
- Multimedia
- Object Search
- يمكن اقتراح ميزة مختلفة على أن يتم الموافقة عليها من أستاذ العملي أو الدكتور

التسليم:

- تقرير مفصل باللغة العربية يصف تصميم وتنفيذ نظام استرجاع المعلومات.
- توصيف ال Dataset المستخدمة (يجب بناء النظام على 2 Datasets) المقترحات وليس إحداها.
- توصيف خطوات المشروع ولكل service ضمن النظام.
- توصيف لبنية النظام system architecture يوضح فيه البنية لل service وكيفية التواصل بينها وفق مفهوم SOA (Service Oriented Architecture).
- النتائج والتقييم على ضمن مجموعة البيانات المرفقة، المطلوب حساب:

Mean Average Precision (MAP) – Recall – Precision@10
Mean Reciprocal Rank (MRR)

- إعادة حساب التقييم بدون ومع الميزة الإضافية للنظام مع كتابة تحليل للنتيجة وتبريرها.
- تقسيم العمل بين أعضاء المجموعة.
- المصادر.
- نسخة تنفيذية لمحركي البحث جاهزة للاستعلام ضمن المقابلة.
- رابط GitHub للكود البرمجي الخاص بالمشروع مع readme واضح لبنية الكود.

ملاحظات تنظيمية:

- لغة العمل حصراً Python.
- يمكن ان تحتوي المجموعة على 4 – 6 أفراد (في حال كان العدد 5 المطلوب تحقيق ميزتين إضافيات من الطلب الثاني أما في حال كان العدد 6 المطلوب تحقيق 3 ميزات إضافية).
- يمكن استبدال واحدة من ال datasets المقترحة بأخرى خارجية.

مُدرس النظري: د. أُبي صندوق.

مُدرسو العملي: م. عبد الرحمن شيباني – م. مروة الداية.