مشروع عملي نظم استرجاع المعلومات 2026

عنوان المشروع: بناء Information Retrieval System

وصف المشروع: مطلوب بناء محرك بحث مخصص يمكنه البحث واسترداد الوثائق من مجموعتي بيانات **مختلفتين**.
يجب استخدام مبادئ استرجاع المعلومات لتصميم وتنفيذ محرك بحث يمكنه التعامل مع استعلامات المستخدمين وإرجاع النتائج ذات الصلة باللغة الطبيعية نصياً.

على كل مجموعة اختيار مجموعتي بيانات **مختلفتين** من الرابط التالي https://ir-datasets.com بشرط أن تحتوي المجموعة الواحدة على أكثر من 200K documents. و على أن تحتوي على data testing و qrel (أي أمثلة للاستعلامات عليها والنتائج الأكثر ملائمة لهذه الاستعلامات).

يجب التحقق من ما يلي:

بناء و تحقيق نظام استرجاع معلومات يحتوي على:

1. **معالجة البيانات Data Pre-Processing :**
   بعد تحميل البيانات يجب معالجتها وفق ما تراه المجموعة مناسب (Stemming, Lemmatization, Normalization , …. etc).

2. **يجب تمثيل الوثائق في كل مجموعة بيانات باستخدام كل من الطرق التالية :**
   - تمثيل VSM_TF-IDF
   - تمثيل Embedding مثل Word2Vec, BERT…. etc
   - تمثيل BM25
   - تمثيل هجين Hybrid Representation
     - تسلسلي (Serial)
     - متوازي (Parallel)

ملاحظات :

1. عند استخدام التمثيل الهجين بطريقة متوازية (Parallel)، يجب استخدام طرق دمج النتائج (Fusion Methods) لاحتساب الدرجات النهائية (Scoring) للوثائق.
2. في BM25 يجب توفير طريقة لرؤية تغيير المعاملات حسب ال query عند التنفيذ (في واجهة المستخدم) أو التوضيح بالتقرير سبب اعتماد المعاملات عند قيم معينة.

3. يجب تطبيق التمثيل الهجين **مرتين** مرة تفرعي ومرة تسلسلي **مع توفير خيار بواجهة المستخدم** للبحث باستخدام التمثيل الهجين التسلسلي أو التمثيل الهجين التفرعي.

4. يمكن استخدام أكثر من نوع من نماذج ال Embedding مع بعضها عند التمثيل الهجين التفرعي.

3. الفهرسة Indexing :

بناء فهرس أو أكثر بحيث يكون مناسب لكل مجموعة بيانات (مثل Inverted Index.. الخ) لإستيراد المستندات بسرعة وفعالية. واختيار مصطلحات الفهرسة بكفاءة.

4. معالجة الاستعلامات Query Processing :

يجب معالجة الاستعلامات باستخدام نفس تقنيات المعالجة المسبقة وتمثيلها بنفس طرق تمثيل الوثائق لضمان التوافق بين الاستعلامات و الوثائق المسترجعة.

5. (Query refinement (query formulation assistance, query suggestion ..etc)) :

تطبيق تحسينات على الاستعلامات لزيادة دقة النتائج مثل تثقيل استعلام المستخدم بمعلومات من سجل بحثه السابق أو اضافة مرادفات على استعلامه، أو تصحيح الاستعلام لغويا ..الخ

6. مطابقة الاستعلام وترتيب النتائج Query Matching & Ranking :

بناء تابع لمطابقة تمثيل الاستعلام مع الوثائق وترتيب النتائج حسب أعلى درجات التشابه ويتم اعتماد طريقة المطابقة المناسبة لكل نموذج تمثيل (.. cos similarity -> embidding | vsm : ex) .

7. تطبيق مبدأ SOA (Service Oriented Architecture) :

يجب تصميم النظام وفق مفهوم **(Service Oriented Architecture (SOA** بحيث يتم تقسيم النظام إلى مجموعة من الخدمات المستقلة، ويكون كل جزء مسؤولاً عن مهمة محددة وقابلاً للتشغيل والتطوير بشكل منفصل.

على سبيل المثال يمكن فصل النظام إلى خدمات مثل:

- خدمة معالجة البيانات (Preprocessing Service)
- خدمة الفهرسة (Indexing Service)
- خدمة البحث والاسترجاع (Retrieval Service)
- خدمة الترتيب والتقييم (Ranking & Evaluation Service)
- خدمة Query Refinement
- خدمة الواجهة الأمامية أو الـ API Gateway

ويجب مراعاة ما يلي:

- تحقيق فصل واضح للمسؤوليات بين الخدمات.
- اعتماد أسلوب تواصل مناسب بين الخدمات (REST API, Message Queue, RPC …).
- كتابة كود منظم وقابل للصيانة والتوسعة.
- إمكانية تشغيل أو اختبار كل Service بشكل مستقل.
- تطبيق أفضل الممارسات والـ Design Patterns المناسبة لتحسين الأداء وقابلية التطوير.
- توضيح بنية النظام وآلية التواصل بين الخدمات ضمن التقرير باستخدام مخطط معماري (Architecture Diagram).

وتزداد علامة هذا القسم كلما كان تصميم الخدمات أكثر تنظيماً ومرونة واحترافية من ناحية:

- Clean Architecture
- Scalability
- Maintainability
- Loose Coupling
- Reusability

كما يفضل توضيح سبب اختيار البنية والتقنيات المستخدمة وكيف ساهمت في تحسين أداء النظام أو تسهيل تطويره مستقبلاً.

8. تقييم النظام

يجب تقييم أداء نظام استرجاع المعلومات باستخدام المقاييس القياسية المعتمدة في مجال الـ IR، وذلك للتحقق من جودة النتائج المسترجعة ومدى ملاءمتها للاستعلامات.

المطلوب حساب المقاييس التالية على الأقل لكل نموذج تمثيل مستخدم ولكل Dataset:

- (Mean Average Precision (MAP
- Recall
- Precision@10
- (Normalized Discounted Cumulative Gain (nDCG

ويجب إجراء التقييم في الحالات التالية:

- قبل تطبيق الميزات الإضافية.
- بعد تطبيق الميزات الإضافية.

مع تقديم تحليل واضح للنتائج بالتقرير يوضح:

- تأثير كل طريقة تمثيل على جودة الاسترجاع.
- مقارنة أداء النماذج المختلفة (TF-IDF، BM25، Embeddings، Hybrid).
- مدى مساهمة الميزات الإضافية في تحسين النتائج أو سرعة الاسترجاع.
- تبرير اختيار النماذج والمعاملات المستخدمة اعتماداً على النتائج العملية.

كما يُرفض أي نظام يحقق نتائج تقييم منخفضة جداً أو غير منطقية مقارنة بطبيعة مجموعة البيانات المستخدمة

9. واجهة مستخدم (UI Web or Mobile Application) :

بناء واجهة استخدام سهلة تتضمن:

- اختيار مجموعة البيانات من الواجهة قبل تنفيذ الاستعلام وقبول استعلامات المستخدمين.
- عرض النتائج ذات الصلة من مجموعة البيانات.
- إمكانية اختيار التنفيذ على الطلبات الأساسية فقط وإمكانية اختيار خيار آخر التنفيذ الأساسية مع الإضافية.
- إمكانية التحكم من خلال الواجهة بمعاملات النموذج الاحتمالي وإمكانية اختيار نموذج التمثيل الهجين من خلال الواجهة

<u>ميزات إضافية للنظام</u>

10. RAG (Retrieval-Augmented Generation)
11. Use vector stores
12. Multilingual retrieval system
13. Crawling
14. Distributed information retrieval
15. Documents clustering
16. Personalization
17. Topic detection
18. Agents
19. LTR

ملاحظة : يمكن اقتراح ميزة مختلفة على أن يتم الموافقة عليها من أستاذ العملي أو الدكتور.

<u>التسليم:</u>

- تقرير مفصل باللغة العربية يصف تصميم وتنفيذ نظام استرجاع المعلومات مع إضافة المصادر.

- توصيف ال Dataset المستخدمة (يجب بناء النظام على 2 Datasets من المقترحات وليس إحداها).
- توصيف خطوات المشروع ولكل service ضمن النظام.
- توصيف بنية النظام (Architecture System) يوضح فيه البنية لل Services وكيفية التواصل بينها وفق مفهوم ال SOA (Service Oriented Architecture) حسب ماذُكر سابقاً.
- تقارير التقييم حسب ماذُكر سابقاً.
- تقسيم العمل بين أعضاء المجموعة.
- 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 أن تحتوي المجموعة على 5 – 7 أفراد مع مراعاة التالي:
  - في حال كان العدد 5 طلاب المطلوب تحقيق ميزة إضافية واحدة.
  - في حال كان العدد 6 طلاب المطلوب تحقيق ميزتين إضافيات.
  - في حال كان العدد 7 طلاب المطلوب تحقيق ثلاث ميزات إضافية.
- يمكن استبدال واحدة من ال datasets المقترحة بأخرى خارجية على أن يتم الموافقة عليها من أستاذ العملي.
- **يمنع استخدام Antique dataset** .

موعد التسليم النهائي في 7/3 .

مدرس النظري: د. أُبي صندوق .

مدرسو العملي: م. مروة الداية, م. سليمى المحايري.