



جامعة دمشق
كلية الهندسة المعلوماتية

المشروع الفصلي السنة الخامسة

مادة نظم البحث عن المعلومات
قسم هندسة البرمجيات

بإشراف: م لين قويدر

تقديم الطلاب:

طه جمال عبد الناصر ملص

رغد ماجد حنا

عبادة أحمد قرطومة

رغد عبد المنعم النحلاوي

توصيف CACM Dataset

i وهو المعرف الخالص بالمقالة

t.عنوان المقالة

b. تاريخ النشر

a. اسم المؤلف

n.معلومات إضافية

w. محتور الملف

x.المرتبطة معها

توصيف المشروع

قمنا باستخدام اللغة python من أجل بناء search engine من طرف OFFLINE

واستخدمنا Flask من أجل انشاء API

من أجل الواجهات استخدمنا Flutter SDK

بداية ملف server هو المسؤول عن اقلاع السيرفر وتوجيه API routes

قمنا بقرأة الداتا سبت وفصل عناصرها والحصول على المقالات باستخدام التابع

get_data_structerd_dataset

ثم قمنا بتنظيف المقالات من stop words والفراغات, مقابل الاختصار قمنا بتوحيد صيغة التواريخ إضافة إلى أننا استخدمنا lemma_words من مكتبة NLTK من أجل الحصول على جذور الكلمات

عند طلب search query نقوم بفحص الكويري في حال وجود خطأ إملائي نقوم بعرض التصحيح على المستخدم

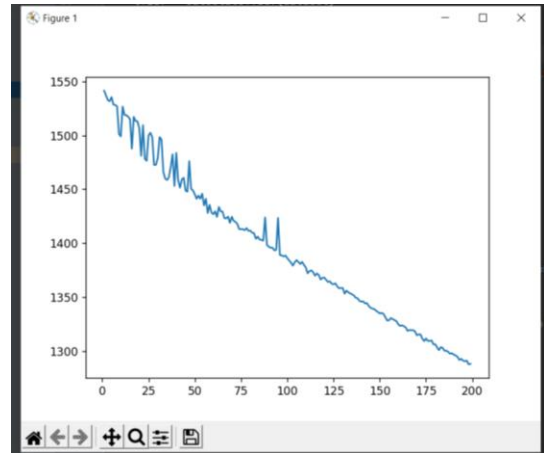
أما في حال عدم وجود خطأ فنقوم بحساب tfidf و cosine_similarity نرتب النتائج حسب الأقرب إلى الواحد ثم نقوم بعرض هذه النتائج

الطلب الإضافي

Document clustering .

قمنا باستخدام cluster

لتحسين عملية matching



تبعاً للمخطط والاحصائية السابقة افترضنا ان عدد cluster يساوي 200

تقسيم العمل

عبادة قرطومة القراءة من clean docs + datasets clean

طه ملص search interface

رغد حنا cos similar + tfidf

رغد النحلاوي clustering

المصادر:

<https://www.pragmalingu.de/docs/guides/how-to-parse>

<https://towardsdatascience.com/create-a-simple-search-engine-using-python-412587619ff5>

<https://towardsdatascience.com/clustering-documents-with-python-97314ad6a78d>