دانشگاه کردستان
University of Kurdistan
زانکۆی کوردستان

تمرین سوم جستجو و بازیابی اطلاعات

موضوع: TF-IDF

استاد

دکتر دانشفر

حل تمرین

پارسا زاهدی

مقدمه:

TF-IDF در بازیابی‌اطلاعات، یک آمار عددی است که میزان اهمیت یک کلمه نسبت به یک سند در یک مجموعه‌ای از اسناد را نشان می‌دهد. در واقع هدف این سیستمِ وزن‌دهی، نشان‌دادن اهمیت کلمه در متن است؛ که اغلب در جستجوهای درون بازیابی‌اطلاعات، متن کاوی و مدل‌سازی کاربر (به انگلیسی: User modeling) استفاده می‌شود. مقدار TF-IDF به تناسب تعداد تکرار کلمه در سند افزایش می‌یابد و توسط تعداد اسنادی که در مجموعه هستند و شامل کلمه نیز می‌باشند متعادل می‌شود. به این معنی که اگر کلمه‌ای در بسیاری از متون ظاهر شود احتمالاً کلمه‌ای متداول است و ارزش چندانی در ارزیابی متن ندارد. در حال حاضر TF-IDF یکی از محبوب‌ترین روش‌های وزن‌گذاری اصطلاحات می‌باشد و امروزه بیش از ۸۳ درصد از سامانه‌های توصیه‌گر در کتابخانه‌های دیجیتال از این روش وزن‌دهی اصطلاحات استفاده می‌کنند. (ویکی‌پدیا)

توضیح:

TF مخفف عبارت term frequency است، یعنی تعداد دفعاتی که یک کلمه در هر داکیومنت استفاده شده است. IDF مخفف عبارت inverse document frequency است، یعنی معکوس تعداد داکیومنت‌هایی که یک کلمه در آن‌ها به‌کار رفته است. IDF میزان اهمیت یک کلمه را نسبت به کل corpus اندازه‌گیری می‌کند. اگر TF را در IDF ضرب کنیم، مقدار TF-IDF score به دست می‌آید. به مجموعه‌ای از داکیومنت‌ها (document) کورپس (corpus) می‌گوییم.

$$tf(t,d) = \frac{\#\text{of t in d}}{\#\text{ total words in d}}$$

$$idf(t,D) = \log\frac{\text{N}}{(\#\text{ Docs with t}) + 1}$$

N => # of Documents

t => term

D => corpus

d = document

دستورالعمل:

دیتاست: از دیتاست‌های "wiki_25.txt" و "wiki_90.txt" استفاده کنید.

کدنویسی: باید از زبان پایتون استفاده کنید. فایل تحویلی ترجیحا ipynb باشد. می‌توانید از https://colab.research.google.com/drive/ هم استفاده کنید.

پیش‌پردازش: مانند تمرین‌های قبل انجام شود.

- دیتاست را از فایل بخوانید و آن‌ها را با استفاده از NLTK پیش‌پردازش کنید.
- محتوای فایل را به توکن تبدیل کنید.
- علائم نگارشی و stop words را حذف کنید.

فایل‌های خروجی:

نمره TF-IDF را برای همه کلمات به دست آورید و فایل py. یا ipynb. به همراه فایل گزارش (ویس یا pdf) zip کرده و آپلود کنید.

راهنمایی: from sklearn.feature_extraction.text import TfidfVectorizer