دانشگاه فردوسی مشهد
دانشکده مهندسی
گروه مهندسی کامپیوتر

تمرین شماره 1
مبانی بازیابی اطلاعات و جست و
جوی وب

پیش‌پردازش متن و نمایه‌گذاری

زمستان 1400، بهار 1401

در این تمرین از شما خواسته شده است تا مراحل پیش‌پردازش و نمایه‌گذاری معکوس[1] را بر روی دیتاستی که به شما معرفی می‌شود، اعمال کنید.

دیتاست

دیتاست در نظرگرفته شده از این لینک قابل دسترسی است. این دیتاست، یک فایل با پسوند xlsx می‌باشد که حاوی فراداده‌ها[2]ی کتب یکی از سایت‌های دانلود کتاب آنلاین فارسی است. ویژگی‌های موجود در این فایل به شرح زیر هستند:

[1] Inverted indexing
[2] Metadata

- عنوان(title): نام کتاب
- تاریخ(date): تاریخ درج کتاب در سایت
- محتوا(content): خلاصه محتوای کتاب
- دسته‌بندی(category)
- نویسنده(author)
- دیدگاه‌ها(comments)

پیش‌پردازش

در فاز پیش‌پردازش باید روش‌هایی مانند normalization، حذف کلمات توقف[3]، ریشه‌یابی[4]، بن‌یابی[5] و ... (که از دید شما منطقی و لازم است) را متناسب با نیازهای مسئله بر روی ویژگی(ستون) «محتوا» در دیتاست اعمال کنید. در پایان فاز پیش‌پردازش باید توکن‌های «محتوا» مربوط به هر یک کتب را استخراج کرده باشید.

نمایه‌گذاری معکوس

در این فاز، باید توکن‌های استخراج شده از فاز پیش‌پردازش را به صورت معکوس نمایه‌گذاری کنید. برای نمایه‌گذاری کردن می‌توانید از شناسه‌های افزایشی[6] استفاده کرده و یا با تولید اعداد تصادفی و منحصر به فرد، شناسه تولید کنید.

در پایان فاز نمایه‌گذاری معکوس، پیاده‌سازی شما باید بتواند یک واژه[7] را به عنوان ورودی دریافت کرده و لیست شناسه کتاب‌هایی که آن واژه را در محتوای خود دارند، خروجی دهد.

نکته: برای پیاده‌سازی، محدودیتی در استفاده از زبان یا کتابخانه وجود ندارد.

---

[3] Stop words
[4] Stemming
[5] Lemmatization
[6] Incremental
[7] Term

موارد تحویلی

فایل ارسالی شما، یک فایل فشرده‌شده(zip یا rar) خواهد بود که باید شامل موارد زیر باشد:

1- گزارش و مستندات کد پیاده‌سازی(pdf یا word)

توضیح مراحل پیاده‌سازی و قطعه‌کدهای مربوط به هر قسمت(ترجیحا عکس کد خود را به گزارش اضافه کنید)، ذکر مشخصات محیط پیاده‌سازی و ماژول‌های اصلی، بیان دلایل استفاده از روش‌ها و ابزارها، بیان چالش‌ها، تعیین فعالیت‌های انجام‌شده توسط هر یک از اعضای گروه و ...

2- کد پیاده‌سازی

3- فایل خروجی فاز نمایه‌گذاری معکوس(csv یا xlsx)

در نهایت فایل فشرده‌شده را به صورت HW01-GroupX نام‌گذاری کنید. به جای X، شماره گروه خود را جایگزین نمایید.

موفق باشید.