به نام خدا

بازیابی پیشرفته اطلاعات

نیم‌سال اول ۰۳-۱۴۰۲

مدرس: دکتر حمید بیگی

دانشکده‌ی مهندسی کامپیوتر

سری اول (۱۰۰ نمره) | کوییز | زمان: ۲۴ مهر

سوال ۱: پاسخ کوتاه (۳۵ نمره)

برای سوالات زیر پاسخ کوتاه بنویسید.

- نقش idf در ساخته شدن بردار tf-idf چیست؟

کلمات پرتکرار در کل مجموعه اسنادی که روی آن کار می‌کنیم، وزن کمتری بگیرند. چرا که اطلاعات کمتری دارند.

- تفاوت Lemmatizaion و Stemming در چیست؟

هر دو فرایندهای مورد استفاده در پردازش زبان طبیعی هستند که به تجزیه و تحلیل کلمات می‌پردازند. تفاوت اصلی بین Lemmatization و Stemming در پردازش زبان طبیعی این است که Lemmatization سعی می‌کند کلمات را به شکل پایه یا لم ساختاری خود تقلیل دهد، در حالی که Stemming فقط به دنبال کاهش کلمات به شکل پایه‌های ساده‌تر است.

(Stemming فرایندی است که با حذف پسوندها و پیشوندهای کلمات، کلمات را به ریشهٔ آنها یا stem تقلیل می‌دهد. این روش به صورت قواعدی و الگویی عمل می‌کند و بدون در نظر گرفتن معنی واقعی کلمه، تنها تغییرات شکلی را در نظر می‌گیرد. بنابراین، خروجی Stemming ممکن است یک کلمهٔ کوچکتر، غیر قابل فهم یا وجود داشتن کلمات غیرموجود باشد.

مثال: فرض کنید دو کلمهٔ "running" و "runner" داریم. با استفاده از روش ،Stemming هر دو کلمه به ریشهٔ "run" تقلیل می‌یابند. نتیجهٔ Stemming برای این دو کلمه، یکسان است.

Lemmatization نیز یک فرایند کاهش کلمات است، اما با توجه به دانش لغت و معنی واقعی کلمه. در این روش، کلمات به شکل اصلی یا لمّا می‌رسند. برای این کار، از دیکشنری‌ها و منابع زبانی استفاده می‌شود تا ریشهٔ واقعی کلمه واکشی شود. نتیجهٔ Lemmatization همچنین یک کلمهٔ مفهومی و موجود در واژه‌نامه خواهد بود.

مثال: برای مثال، اگر کلمهٔ "better" را داشته باشیم، با استفاده از ،Lemmatization به ریشهٔ "good" تقلیل می‌یابد، زیرا "better" به معنای "بهتر" است و ریشهٔ آن "good" است.

به طور خلاصه، تفاوت اصلی بین Lemmatization و Stemming در روش کاهش کلمات و در نظر گرفتن معنی واقعی کلمه است. Stemming به صورت قواعدی و الگویی عمل می‌کند و تنها تغییرات شکلی را در نظر می‌گیرد، در حالی که Lemmatization با در نظر گرفتن معنی کلمه، کلمات را به شکل اصلی یا لمّا می‌رساند.)

- ،مطابق الگوی مبتنی بر برنامه‌ریزی پویا اگر هزینه‌ی افزایش و کاهش برابر ۱ و هزینه‌ی تغییر برابر ۳ باشد کمترین هزینه کلمات زیر را بدست آورید.
  - فرسایش - آسایش

    فر -> آ : ۱+۱+۱ = ۳

    ۲ حرف حذف شده و یکی اضافه شده است.
  - تقطیر - تغییر

    قط -> غ : ۱+۱+۱ = ۳

    ۲ حرف حذف شده و یکی اضافه شده است.

سوال ۲: Sorting Algorithm (۲۰ نمره)

External Sorting Algorithm را توضیح داده و ۲ مشکل آن را بیان کنید.

این الگوریتم برای مرتب‌سازی خارجی استفاده می‌شود که به طور کلی نمی‌توانند به یکباره در حافظه اصلی سورت شوند. این الگوریتم برای مرتب‌سازی حجم بزرگی از داده‌ها که نیاز به استفاده از حافظه ثانویه (مثل دیسک سخت) دارند، استفاده می‌شود.
عملکرد الگوریتم مرتب‌سازی خارجی بر اساس مبدأ تقسیم و مرتب‌سازی است. فرایند اصلی شامل مرحله‌های زیر است:

- تقسیم داده‌ها: ابتدا داده‌ها به بخش‌های کوچکتر تقسیم می‌شوند که به حافظه اصلی مناسب باشند. این بخش‌ها به صورت موازی مرتب می‌شوند و در حافظه ثانویه ذخیره می‌شوند.

- مرتب‌سازی بخش‌ها: هر بخش در حافظه اصلی مرتب می‌شود. می‌توان از یک الگوریتم مرتب‌سازی داخلی (مانند الگوریتم مرتب‌سازی ادغامی) برای این کار استفاده کرد.
- ادغام بخش‌ها: در این مرحله، بخش‌های مرتب شده در حافظه اصلی ادغام می‌شوند. این ادغام به صورت مرحله‌ای انجام می‌شود و می‌توان از الگوریتم‌های ادغام مانند ادغام دوتایی استفاده کرد.

مشکلاتی که در الگوریتم مرتب‌سازی خارجی ممکن است به وجود آیند عبارتند از:

- نیاز به دسترسی مکرر به حافظه ثانویه: اجرای الگوریتم مرتب‌سازی خارجی نیازمند مکرر خواندن و نوشتن از حافظه ثانویه است که زمان بر و هزینه‌بر است. این می‌تواند باعث کاهش کارایی الگوریتم شود.
- مشکل تعادل بار: در حین ادغام بخش‌ها، برای تعادل بار و جلوگیری از ترشح حافظه اصلی، باید به دقت برنامه‌ریزی شود. اگر بار بین بخش‌ها ناهموار توزیع شود، ممکن است برخی بخش‌ها بیش از حد بزرگ شوند و باعث افزایش زمان اجرا شوند.

سوال ۳: اسناد (۴۵ نمره)

۳ سند زیر را در نظر بگیرید که پس از عملیات پیش پردازش به صورت زیر خواهند بود:

داک ۱: learning mechanics build machine
داک ۲: artificial intelligence machine learning
داک ۳: text mining deep learning

- شاخص معکوس (inverted index) را برای کلمه های learning، mechanics، machine، deep رسم کنید.

کلمه :"deep"
داک ۳
کلمه :"machine"
داک ۱ داک ۲
کلمه :"mechanics"
داک ۱
کلمه :"learning"
داک ۱ داک ۲ داک ۳

- برای کلمه های learning، mechanics، machine، deep ماتریس term-document را رسم کنید.

word	doc۱	doc۲	doc۳
deep	۰	۰	۱
machine	۱	۱	۰
mechanics	۱	۰	۰
learning	۱	۱	۱

- با توجه به ماتریس رسم شده اسنادی که با پرسش "machine learning" مرتبط هستند را بدست آورید.

110 bitwise AND 111 = 110 => doc 1, doc 2