بازیابی پیشرفته اطلاعات

نیم‌سال اوّل ۰۰ـ۹۹

دانشکده‌ی مهندسی کامپیوتر

مدرس: حمید بیگی

فاز سوم پروژه (۱۰۰ نمره)

زمان تحویل: ۱۵ بهمن‌ماه

هدف از فاز سوم پروژه پیاده‌سازی الگوریتم‌های خوشه‌بندی و یک خزنده برای واکشی اطلاعات مقالات از سایت Microsoft Academic است. در بخش اول این فاز الگوریتم‌های خوشه‌بندی را برای یک مجموعه داده‌ی فارسی پیاده‌سازی می‌کنید و خوشه‌های به دست آمده را در خروجی برمی‌گردانید. در بخش دوم نیز یک خزنده برای واکشی اطلاعات مقالات از سایت Microsoft Academic پیاده‌سازی می‌کنید و در آخرین بخش PageRank را برای مقالات واکشی‌شده محاسبه می‌کنید.

بخش ۱. خوشه‌بندی (۵۰ نمره)

در بخش اول باید برای یک مجموعه داده به زبان فارسی، چند الگوریتم خوشه‌بندی را پیاده‌سازی کنید. مجموعه داده انتخابی شامل اطلاعات یک مجموعه اخبار در قالب یک فایل json است. در این دادگان، اطلاعات زیر برای هر خبر موجود است:

- title: عنوان خبر
- summary: خلاصه متن خبر
- link: لینک به خبر در وبسایت
- tags: دسته‌بندی خبر. به فرمت «دسته اصلی > دسته فرعی». مثلا: «دانش > پزشکی»

برای تبدیل متون اخبار به فضای برداری باید یک بار از طریق TF-IDF و یک بار از Word2vec استفاده کنید. توجه کنید که برای هر دو روش می‌توانید از توابع و کتاب‌خانه‌های آماده استفاده کنید. ضمنا استفاده از کتابخانه هضم برای پیش‌پردازش متون توصیه می‌شود. الگوریتم‌های خوشه‌بندی که باید پیاده‌سازی شوند نیز عبارت هستند از:

۱. **K-means**

۲. **Gaussian Mixture Model**

۳. **Hierarchical clustering**

توجه: انتخاب تمامی پارامترهای الگوریتم‌های بالا برعهده‌ی خودتان است. نحوه انتخاب بهترین عدد برای تعداد خوشه‌ها و مقادیر سایر پارامترهایی که انتخاب می‌کنید را در گزارش بیاورید. برای پیاده‌سازی الگوریتم‌های خوشه‌بندی نیز می‌توانید از توابع و کتاب‌خانه‌های آماده استفاده کنید. به ازای هر زوج از روش‌های تبدیل به فضای برداری و الگوریتم خوشه‌بندی یک فایل csv در خروجی داشته باشید (مجموعا ۶ فایل) که نتیجه‌ی خوشه‌بندی الگوریتم شما است. این فایل‌ها باید دو ستون داشته باشند: ستون اول لینک خبر و ستون دوم شماره‌ی خوشه برای آن خبر را قرار دهید. همچنین در گزارش خود تعدادی نمودار از نتایج خوشه‌بندی‌های خود (در قالب نمودار دوبعدی یا دندروگرام و یا هر نمودار جالب دیگری) ارائه کرده و در چند سطر به صورت مختصر مشاهدات خود به همراه توضیحاتی ارائه دهید. همچنین حداقل ۲ معیار ارزیابی خوشه‌بندی را به ازای هر بار خوشه‌بندی، در گزارش بیاورید. در حین ارزیابی، از دسته‌بندی اصلی اخبار به عنوان ground truth استفاده کنید.

بارم‌بندی

بارم‌بندی این بخش قطعی نیست. بر حسب امتیازهایی که خوشه‌بندی‌های شما به دست می‌دهند نمره‌تان تعیین می‌گردد.

بخش ۲. پیاده‌سازی خزنده، واکشی اطلاعات مقالات (۴۰ نمره)

در این بخش قصد داریم تا برای سایت Microsoft Academic یک خزنده پیاده‌سازی کرده و با استفاده از آن اطلاعات تعدادی مقاله را واکشی کنیم.

اطلاعاتی که از هر مقاله باید جمع‌آوری شوند عبارت هستند از:

۱. عنوان مقاله

۲. چکیده‌ی مقاله

۳. سال انتشار مقاله

۴. تمامی نویسندگان مقاله

۵. ارجاعات مقاله. توجه کنید که تنها ۱۰ ارجاع اول که در صفحه‌ی مقاله در سایت Microsoft Academic قرار دارد کافی است و نیازی به واکشی تمامی ارجاعات نیست.

خزنده برای آغاز کار باید از چند مقاله‌ای که در فایل start.txt وجود دارند و در صف خزش قرار می‌گیرند شروع کرده و ۵۰۰۰ مقاله (تعداد کل مقالات به عنوان پارامتر ورودی داده می‌شود) را ذخیره نماید. همچنین آدرس ۱۰ مقاله‌ی ابتدایی در لیست ارجاعات مقاله‌ی کنونی به صف خزش خزنده اضافه می‌شود. توجه نمایید که برخی مقالات دارای چند نسخه منتشر شده در وبسایت‌های مختلف هستند و یکی از این نسخه‌ها باید ذخیره شود. هیچ مقاله‌ای نباید بیش از یک بار ذخیره شود. اگر لیست ارجاعات یک مقاله کمتر از ۱۰ مورد باشد ایرادی ندارد. همچنین برخی از ارجاعات به صورت لینک نیستند که می‌توانید از آن‌ها چشم‌پوشی کنید.

یک نمونه فایل json از اطلاعات ذخیره شده در فایل sample.json قرار دارد. برای ذخیره‌سازی اطلاعات مقاله‌ها مشابه این فایل نمونه اقدام نمایید. توجه کنید که برای خزش سایت Microsoft Academic شاید نیاز باشد تا بین درخواست‌های خود تاخیر (delay) بیاندازید.

بارم‌بندی

۱. پیاده‌سازی خزنده (۳۰ نمره)

۲. ذخیره‌ی اطلاعات مقاله‌های به فرمت json و بررسی اولیه اطلاعات استخراج شده (۲۰ نمره)

بخش ۳. PageRank (۱۰ نمره)

در آخرین بخش الگوریتم PageRank را بر روی مقالات واکشی‌شده اجرا کرده و نتایج آن را به دست می‌آوریم. توجه کنید که اگر مقاله‌ی A به مقاله‌ی B ارجاع (reference) داشته باشد، آنگاه پیوندی (link) از مقاله‌ی A به مقاله‌ی B در نظر می‌گیریم. برای این منظور مقدار α مورد نیاز در ورودی گرفته می‌شود و سپس معیار PageRank برای تمامی مقالات محاسبه شده و در خروجی چاپ می‌شود. برای این بخش می‌توانید از ابزارهای آماده استفاده کنید.

بارم‌بندی

۱. محاسبه‌ی معیار PageRank برای مقالات واکشی‌شده و گزارش مقالات با بالاترین رنک (۱۰ نمره)

بخش ۴. نکات

۱. امکان تغییر بارم‌بندی وجود دارد.

۲. نوشتن گزارش فراموش نشود. به قوانین کلاس و پروژه که در پیاتزا قرار گرفته است رجوع کنید.