دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکدهٔ ریاضی و علوم کامپیوتر

پروژهٔ سوم درس داده‌کاوی

خوشه‌بندی متنی و استخراج موضوع از مقالات کووید–۱۹

استاد درس: دکتر فاطمه شاکری

طراحان پروژه: مریم صادقی، مهدی رجالی

زمستان ۱۴۰۲

۱ معرفی مجموعهٔ داده

طی چندسال گذشته که با همه گیری کووید ۱۹ روبه رو شدیم مقالات زیادی در این حوزه منتشر شد. متخصصان جوامع علمی و محاسباتی با این چالش مواجه شدند که اطلاعات مورد نیازشان را در اسرع زمان از بین حجم انبوهی مقاله پیدا کنند.
مجموعه دادهٔ پیش‌رو، مجموعه‌ای از مقالات و تحقیقات مرتبط با بیماری کووید−۱۹ است. اطلاعات 10000 مقاله در قالب یک فایل `csv`. به شما داده‌است، که شامل موارد زیر است:

- `paper_id`: این ویژگی صرفاً هش [1] ساخته‌شده از فایل‌های PDF مقالات است. از آنجا که برخی از مقالات دارای چند فایل PDF بودند، ممکن است بیش از یک هش برای یک مقاله موجود باشد؛ اما اغلب مقالات دارای صفر هش (یعنی ذخیره‌نشده) یا یک هش هستند.
- `doi`: شناساگر اشیاء دیجیتال [2] که برای ارجاع دادن به مقالات استفاده می‌شود.
- `abstract`: چکیدهٔ مقاله (در صورت وجود)
- `body_text`: متن اصلی مقاله
- `authors`: نویسندگان مقاله، یک قالب یک لیست از رشته‌ها
- `title`: عنوان مقاله
- `journal`: ژورنالی که مقاله در آن ثبت شده است
- `abstract_summary`: خلاصهٔ چکیده

۲ هدف

با توجه به تعداد زیاد مقالات مرتبط و گسترش سریع کووید−۱۹، برای متخصصان سلامت دشوار است که خود را با اطلاعات جدید به‌روز نگه دارند. حال سوال این است که آیا خوشه بندی مقالات تحقیقاتی مشابه می‌تواند جستجوی انتشارات مرتبط را ساده‌تر کند؟ ضمناً آیا امکان استخراج موضوعات کلیدی هر خوشه موجود است؟ پاسخ‌دهی به این سوالات، به تسهیل بررسی بسیاری از نشریات مرتبط با این ویروس و تصمیم‌گیری مناسب متخصصان کمک می‌کند.

۳ مراحل مورد نیاز

برای این پروژه، یک فایل ژوپیتر نوتبوک [3] در اختیار شما قرار گرفته‌است که در آن برای هر جزئی که می‌بایست پیاده‌سازی کنید، توضیحی آورده شده است. در مورد الگوریتم‌ها و اجزاء خواسته‌شده تحقیق کنید و خلاصهٔ مختصری از کاربرد و چگونگی کارکرد آنها ارائه دهید. استفاده از تمامی کتاب‌خانه‌ها و توابع آماده برای پیاده‌سازی اجزاء مجاز است.

[1] Hash
[2] Digital object identifier
[3] Jupyter notebook

نحوهٔ ارزیابی

این پروژه دارای ۱۰۰ نمرهٔ اصلی و ۱۰ نمرهٔ امتیازی است، که تقسیم‌بندی ۱۰۰ نمرهٔ اصلی به شکل زیر می‌باشد:

بخش	نمره
پیش‌پردازش متنی	۲۰
استخراج ویژگی و خوشه‌بندی	۳۰
کاهش بعد، مصورسازی و مدل‌سازی تاپیک‌ها	۴۰
گزارش	۱۰

نمرهٔ امتیازی مختص خوشه‌بندی به کمک الگوریتم‌های دیگر (غیر از K-Means، مثلاً DBSCAN یا Hierarchical) و مدل‌کردن تاپیک‌ها به کمک روش‌های غیر از LDA است.

نحوه و مهلت ارسال

مهلت ارسال

شما برای ارسال این پروژه، تا پایان ۱۵ خرداد (یعنی ساعت ۲۳:۵۹ روز مذکور) فرصت دارید.

قالب ارسالی

فایل ارسالی شما باید به فرمت `zip`. با نام‌گذاری `project_3_<GroupID>.zip` باشد که در آن می‌بایست به جای `<GroupID>`، شمارهٔ گروه خود را قرار دهید. فایل فشردهٔ ارسالی می‌بایست شامل **هر سهٔ** مورد زیر باشد:

۱. یک فایل `pdf`. حاوی گزارش پروژه

۲. یک فایل `ipynb`. حاوی کد **کامنت‌گذاری‌شدهٔ** مرتبط با پروژه

۳. یک فایل `txt`. حاوی لینک گوگل کولب [۴] منطبق با ژوپیتر نوتبوک [۵] بخش قبل

در صورتی که ترجیح می‌دهید گزارش را در داخل خود نوت‌بوک بنویسید، حتماً از قالب سلول‌های متنی از پیش قرار داده‌شده استفاده کنید و برای فرمت‌دهی مناسب به متن، از تگ‌های HTML و دستورات LaTeX استفاده کنید.

[4] Google Colaboratory
[5] Jupyter Notebook