آزمایشگاه تحلیل پیشرفته کلان داده (ابدال)

تمرین تحلیل داده‌های متنی

استاد درس: دکتر حسین رحمانی

تاریخ: ۱۴۰۰/۰۳/۱۰

نسخه: ۱

مقدمه

- پاسخ به سوالات این تمرین باید در قالب یک گزارش کامل و با فرمت PDF ارائه شود.
- توجه داشته باشید که بخش زیادی از ارزیابی تمرین با توجه به گزارش شما انجام خواهد شد، پس تمامی مراحل و نتایج به خوبی در آن انعکاس پیدا کند و در نگارش آن نیز دقت فرمایید.
- برای بازنمایی و توضیح نتایج هر مرحله تا حد ممکن از نمودار و شکل استفاده کنید.
- از قراردادن کد و توضیح آن در متن گزارش خودداری فرمایید.
- برای پاسخ به تمرینات باید حتما از زبان پایتون استفاده شود.
- فایل کدهای اجرا شده به صورت خوانا و با توضیحات لازم درباره هر بخش آن به صورت کامنت، به همراه گزارش تمرین بارگذاری شود.
- تمامی فایل‌های مربوط به این تمرین (گزارش و کدها)، در قالب یک فایل فشرده با نام‌گذاری زیر ارسال شود.

StudentNumber_FirstName_LastName_HW#.zip

- فایل تمرین را از طریق سامانه LMS ارسال نمایید.

۱- فایل ورودی

داده‌های ما در دو فایل قرار دارند. فایل movie_synopsis شامل plot_synopsis فیلم‌ها به عنوان داده متنی موجود در دیتاست ما است. فایل دیگر (movie_info) نیز شامل اطلاعات فیلم‌ها از جمله عنوان و ژانر است. این دو دیتاست را می‌توانید به وسیله local_id با یک دیگر ادغام کنید.

۲- پیش پردازش داده

یکی از مهم‌ترین مراحل تحلیل داده‌های متنی، پیش‌پردازش داده‌ها است. به منظور نتیجه‌گیری بهتر از بسیاری از الگوریتم‌های داده‌کاوی، لازم است تغییرات و یا اصلاحاتی بر روی داده‌های خام انجام شوند تا کیفیت الگوها و قواعد کاوش شده از داده‌ها، به بیشترین حد ممکن افزایش یابد. از جمله این موارد می‌توان به حذف علایم نگارشی، حذف stop word ها، ریشه‌یابی کلمات و ... اشاره کرد. کتابخانه‌های مختلفی در زبان‌های برنامه‌نویسی مختلف برای انجام پیش‌پردازش طراحی شده‌اند. در زبان برنامه‌نویسی پایتون کتابخانه nltk برای زبان انگلیسی طراحی شده‌است که با مراجعه به مستندات این کتابخانه‌ها می‌توانید اطلاعات بیشتری از قابلیت‌های آن‌ها به‌دست بیاورید.

سوال ۱: تفاوت stemming و lemmatization را با ذکر مثال توضیح دهید.

تمرین ۱: در این مرحله لازم است پیش‌پردازش‌های مورد نیاز را روی داده‌های بخش ۱ انجام دهید و نتیجه را با داده‌های خام مقایسه کنید.

۳- استخراج ویژگی

استخراج ویژگی از متون، مرحله‌ای بسیار مهم در پردازش زبان‌های طبیعی است. برای اجرای بسیاری از الگوریتم‌های داده‌کاوی و یادگیری ماشین، باید هر سند در قالب یک بردار (مجموعه‌ای از ویژگی‌ها) نمایش داده شود. روش‌های متعددی در این زمینه مورد استفاده قرار می‌گیرند که یکی از این روش‌ها tf است. با استفاده از tf می‌توان هر جمله یا سند را در قالب یک بردار نمایش داد. یکی دیگر از روش‌های رایج برای تبدیل کلمه به بردار، Word2vec است. Word2vec برای هر کلمه یک بردار در نظر می‌گیرد که با استفاده از آن می‌توانیم شباهت معنایی بین کلمات را پیدا کنیم. برای دریافت نتایج مناسب از مدل Word2vec نیاز به آموزش بر روی مجموعه داده‌ی زیادی است، اما مدل‌های از پیش‌آموزش دیده شده زیادی در اینترنت موجود است و می‌توان از آن‌ها استفاده کرد. در وب‌سایت https://projector.tensorflow.org/ می‌توانید به صورت آنلاین شباهت بین کلمات را با استفاده از w2v مشاهده کنید.

سوال ۲: چند نمونه دیگر از روش‌های استخراج ویژگی را نام برده و یکی از آن‌ها را در چند سطر توضیح دهید.

تمرین ۲: با استفاده از یک روش به دلخواه خود، استخراج ویژگی انجام دهید.

۴- پردازش داده

بعد از پیش‌پردازش بر روی داده‌های متنی نوبت به استخراج ویژگی رسید، همانطور که دیدیم روش‌های مختلفی برای این امر وجود دارد که انتخاب هر یک از آن‌ها تاثیر مستقیم بر روی نتیجه الگوریتم‌های داده‌کاوی دارند. بعد از مرحله استخراج ویژگی نوبت به استخراج دانش از داده‌ها می‌رسد. در این مرحله می‌توان با اجرای الگوریتم‌ها و تکنیک‌های رایج داده‌کاوی به نتایج جالبی رسید. یکی از وظایف مرسوم داده‌کاوی خوشه‌بندی است که الگوریتم‌های مختلفی برای انجام آن وجود دارد و هرکدام مزیت‌ها و معایبی دارند.

سوال ۳: با توجه به مزایا و معایب روش‌های خوشه‌بندی، یک روش مناسب برای این دیتاست با ذکر دلیل انتخاب کنید؟

تمرین ۳ : روش خوشه‌بندی انتخابی را بر روی خروجی حاصل از مرحله قبل پیاده‌سازی کنید.

۵- پساپردازش

از ابتدا تا انتهای مراحل پیش‌پردازش، استخراج ویژگی و پردازش داده، تنها بخشی از فرایند داده‌کاوی است. یکی از مهمترین مراحل داده‌کاوی تحلیل نتایج به‌دست آمده است که معمولا کمتر به آن توجه می‌شود. یک داده‌کاو خوب، نتایج را تحلیل و بررسی می‌کند.

تمرین ۴ : به صورت دستی به بررسی نتایج حاصل از خوشه‌بندی بپردازید و نمونه‌ای از نتایج جالب را بیان کنید، برای تحلیل این قسمت می‌توانید از ژانر و موضوعات فیلم‌ها استفاده کنید.