بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته اطلاعات

نیم‌سال دوم ۰۱-۰۰

استاد: احسان‌الدین عسگری

تمرین سری چهارم — مهلت تحویل: **۲۳ خرداد**

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

توضیحات کلی

در این تمرین نیز هریک از گروه‌ها بر روی پروژه‌ی خود به صوت جداگانه کار خواهند کرد در این تمرین در بسیاری از بخش‌ها می‌توانید از حاصل کار عزیزان ترم گذشته که با زحمات تدریس‌یاران درس در قالب کتابخانه parsi.io ایجاد شده بهره ببرید. به امید خدا در ترم‌های آینده حاصل تجمیع زحمات شما عزیزان در قالب محصولات متن‌باز (البته با ذکر نام خودتان) در اختیار دیگر دانشجویان و بلکه جامعه ایرانی قرار می‌گیرد تا در اثر این تلاش‌ها محصولاتی ارزشمند برای پردازش متن‌های فارسی و بلکه زبان‌های ایرانی و فراتر از آن داشته باشیم. می‌توانید به این کتابخانه از طریق این لینک دسترسی داشته باشید

مدل‌های مورد استفاده برای خوشه‌بندی و دسته‌بندی

در آغاز دقت شود که هر گروه فقط یکی از دو تسک خوشه‌بندی و یا دسته‌بندی را برای تمرین سری چهارم باید انجام دهد. به عبارتی نباید هم دسته‌بندی و هم خوشه‌بندی انجام شود. در صورتی که گروهی خوشه‌بندی را انتخاب کند، تسک دسته‌بندی را در زمان پروژه‌ی پایانی باید پیاده کند و در صورتی که گروهی تسک دسته‌بندی را برای تمرین سری چهارم انتخاب کند، خوشه‌بندی همان هنگام تحویل پروژه‌ی پایانی درس تحویل گرفته می‌شود. بنابراین هر دو تسک تا پایان ترم باید انجام شوند ولی برای این تمرین کافی است یکی از آن‌ها را انجام دهید

برای خوشه‌بندی از الگوریتم Kmeans استفاده کنید. برای دسته‌بندی نیز بایستی از یک مدل بر پایه‌ی Transformer ها و یک مدل نیز از روش‌های سنتی دسته‌بندی انتخاب شود. پیش‌پردازش‌های لازم نیز باید توسط خود اعضای گروه در صورت لزوم انجام شود و همچنین اثر آن‌ها در گزارش آورده شود.

نحوه‌ی ارزیابی در تمرین سری چهارم

در این تمرین برای ارزیابی دسته‌بندهای خود از معیار F1-Macro استفاده کنید. همچنین accuracy و ماتریس درهم‌ریختگی را نیز گزارش کنید. برای ارزیابی خوشه‌بندها نیز از معیار RSS و یک معیار دل‌خواه دیگر استفاده کنید. معیار دل‌خواه دیگر برای نمونه می‌تواند از میان معیارهای این لینک و یا هر مرجع دیگری انتخاب شود. در صورتی که در هر یک از بخش‌های تمرین سری چهارم، افزون‌بر معیارهای بالا، معیار دیگری نیز از شما خواسته شود، بایستی خروجی شما شامل آن موارد نیز باشد.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد:

۱. دسته‌بندی باب‌های بوستان یا گلستان سعدی

کتاب بوستان دارای ده باب است. یک مسئله‌ی دسته‌بندی می‌تواند، به این صورت تعریف شود که با داشتن یک شعر کامل یا چند بیت از یک شعر و یا حتی یک تک‌بیت بتوانید باب مربوط به آن شعر یا بیت را پیش‌بینی کنید. برای نمونه در صورتی که ورودی مدل شمابیت «نصیحت کسی سودمند آیدش که گفتار سعدی پسند آیدش» باشد، مدل دسته‌بند باید بتواند به درستی باب مربوط به این شعر را تعیین کند که پاسخ درست آن «در نیایش خداوند» است.

۲. دسته‌بندی وزن‌های عروضی غزلیات سعدی

می‌دانیم که در ادبیات فارسی، هر بیت شعر، یک وزن دارد. به طور دقیق‌تر باید گفت که هر مصرع دارای یک وزن مشخص است. همچنین می‌دانیم که مصرع‌های یک قطعه‌ی شعر، همگی یک وزن مشخص دارند. برای نمونه اگر شعری دارای ده بیت باشد، همه‌ی بیست مصرع آن، باید بر یک وزن باشند. وزن‌های عروضی معروف تعدادشان حدود ۳۰ تا است. برای نمونه یک وزن شعری «فعولن فعولن فعولن فعل» است. لیست وزن‌های شعری را در صورت علاقه می‌توانید با جستجو در اینترنت بیابید. مسئله‌ی دسته‌بندی در این‌جا، آن است که شما با دریافت یک مصرع شعر، بتوانید وزن آن مصرع را پیش‌بینی کنید. برای این کار نیاز به ساختن یک مدل دسته‌بند دارید. از آن‌جا که در وب‌گاه گنجور، وزن همه‌ی شعرها مشخص شده است، داده‌ی مورد نیاز برای ساخت و ارزیابی مدل نیز فراهم است. ذکر این نکته تاکید دوباره می‌شود که داده‌های ارزیابی شما شامل «مصرع‌ها» هستند و نه یک شعر کامل. زیرا وزن همه‌ی مصرع‌ها در یک شعر، یکسان است. از آن‌جا که کل شعرهای بوستان سعدی همگی بر یک وزن مشخص سروده شده‌اند، بنابراین برای داشتن مجموعه داده‌ی متنوع، بر روی غزلیات سعدی این تمرین انجام شود چرا که غزلیات سعدی دارای وزن‌های مختلفی هستند.

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد.

۱. خوشه‌بندی شعرها

در این حالت شما باید همه‌ی بیت‌های شعرهای بوستان سعدی را استخراج کنید و روی آن‌ها خوشه‌بندی اعمال کنید. همچنین در یک حالت دیگر باید خوشه‌بندی را بر روی شعرها انجام دهید (نه تک‌بیت‌ها). در این حالت پس از انجام خوشه‌بندی گزارش کنید که خوشه‌های ساخته شده چقدر شبیه به باب‌های بوستان سعدی

بخش‌بندی شده‌اند؟ با ذکر چند نمونه به صورت شهودی، مشاهده‌های خود را شرح دهید. همچنین نمایش گرافی خوشه‌بند خود را در این حالت ارائه دهید. برای آن‌که برای نمایش گرافی، به مشکل برنخورید می‌توانید موقع نمایش گرافی، نام هر شعر را به‌جای خود آن شعر قرار دهید.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریسیار این track کار کنید.

۱. دسته‌بندی اخبار بر اساس برچسب خبر

مقالات خبری موضوعات مختلفی از جمله سیاست، مسائل اجتماعی، سینما، ورزش و .. را پوشش می‌دهند و بر این اساس برچسب‌گذاری می‌شوند. یک مسئله‌ی دسته‌بندی می‌تواند این باشد که با دریافت عنوان خبر، برچسب‌های آن خبر را مشخص کنید.

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش خوشه‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریسیار این track کار کنید.

۱. خوشه‌بندی اخبار بر اساس عناوین مقالات خبری

در این مسئله، شما باید روی عناوین مقالات خبری خوشه‌بندی را انجام دهید. پس از انجام این خوشه‌بندی، نتایج به دست آمده را طی گزارشی تحلیل کنید. با ذکر چند نمونه به صورت شهودی، مشاهدات خود را شرح دهید.

۲. خوشه‌بندی اخبار بر اساس متن مقالات خبری

در این مسئله، شما باید روی متن مقالات خبری خوشه‌بندی را انجام دهید. پس از انجام این خوشه‌بندی، نتایج به دست آمده را طی گزارشی تحلیل کنید. با ذکر چند نمونه به صورت شهودی، مشاهدات خود را شرح دهید.

دسته‌بندی

یکی از دو سناریوی زیر را برای دسته‌بندی می‌توانید در نظر بگیرید:

۱. دسته‌بندی کلمات کلیدی

در نظر گرفتن کلمات کلیدی مقاله‌ها به عنوان کلاس‌های هدف، و پیش‌بینی آن‌که یک مقاله‌ی ورودی شامل کدام کلیدواژه‌ها است. دقت شود که هر مقاله می‌تواند multiclass نیز باشد.

۲. دسته‌بندی مقاله‌ها بر اساس مرجع داوری آن‌ها

ساخت یک دسته‌بند باینری که مشخص می‌کند هر مقاله در کدام مرجع به چاپ رسیده است.

• دسته‌ی مرجع شماره‌ی یک

مراجع زیر را به عنوان کلاس‌های تسک طبقه‌بندی در نظر بگیرید و دیتای مناسب برای این سه دیتاست تهیه و سپس طبقه‌بندی را انجام دهید.

European Conference on Computer Vision (ECCV)
Conference on Computer Vision and Pattern Recognition (CVPR)
International Conference on Computer Vision (ICCV)

دسته‌ی مرجع شماره‌ی دو

مراجع زیر را به عنوان کلاس‌های تسک طبقه‌بندی در نظر بگیرید و دیتای مناسب برای این سه دیتاست تهیه و سپس طبقه‌بندی را انجام دهید.

Annual Meetings of the Association for Computational Linguistics (ACL)
Association for Computational Linguistics - European Chapter (EACL)
Empirical Methods in Natural Language Processing (EMNLP)
The North American Chapter of the Association for Computational Linguistics (NAACL)

خوشه‌بندی

سناریوی زیر را در نظر بگیرید

۱. خوشه‌بندی مقاله‌ها

مقاله‌ها را بر اساس چکیده، به تعدادی خوشه‌ی معنادار نقسیم کنید و سعی کنید مقدار مناسب برای k پیدا کنید و معیار purity را به ازای مقادیر مختلف k مقایسه کنید. توجه داشته باشید که پیشنهاد می‌شود که از دیتاست جمع آوری شده در تمرین قبل برای این بخش، استفاده نمایید.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. دسته‌بندی ریاضی‌دانان بر اساس دانشگاه محل تحصیل

طبق داده‌هایی در تمرین گذشته استخراج، هر ریاضی‌دان در دانشگاهی تحصیل کرده است. می‌توانید با دریافت نام ریاضی‌دان و استفاده از سایر داده‌های استخراج شده، دانشگاه محل تحصیل او را مشخص کنید. با توجه به اینکه تعداد این دانشگاه‌ها می‌تواند زیاد باشد، کافی‌ست دسته‌بندی را بر اساس پرتکرارترین‌ها انجام دهید.

۲. دسته‌بندی ریاضی‌دانان بر اساس موقعیت مکانی

هر ریاضی‌دان در یک کشور متولد شده، زندگی کرده، و یا فوت کرده‌است. می‌توانید هر یک از این سه مورد را از صفحه‌ی ریاضی‌دانان در وبسایت شجره‌نامه‌ی ریاضی‌دانان و یا از بیوگرافی ریاضی‌دانان استخراج کرده و بر اساس آن ریاضی‌دانان را دسته‌بندی کنید.

۳. دسته‌بندی ریاضی‌دانان بر اساس حوزه‌ی کاری

مشابه مورد قبل، می‌توانید ریاضی‌دانان را بر اساس حوزه‌ی تحقیقاتی یا فیلدشان، که در بیوگرافی و صفحه اصلیشان موجود است، دسته‌بندی کنید. لازم به ذکر است که چون ممکن است یک ریاضی‌دان در چند فیلد کار کرده باشد، توجه به یکی از آنها کافی‌ست.

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش خوشه‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. خوشه‌بندی ریاضی‌دانان بر اساس پایان‌نامه

در صفحه‌ی هر ریاضی‌دان موضوع پایان‌نامه‌ی او ذکر شده است. می‌توانید ریاضی‌دانان را بر اساس این موضوع خوشه‌بندی کنید. توجه داشته باشید که این بخش به زبان‌های مختلف نوشته شده است؛ پس باید ابتدا به انگلیسی ترجمه گردد (ترجمه تحت‌اللفظی کافیست) و سپس خوشه‌بندی انجام شود. هم‌چنین صحت و کلیات این خوشه‌بندی را در گزارشی مختصر بررسی نمایید.

۲. خوشه‌بندی ریاضی‌دانان بر اساس خلاصه بیوگرافی

در صفحه‌ی بیوگرافی ریاضی‌دانان در بخشی تحت عنوان Summary خلاصه‌ای از زندگی‌نامه‌ی آن ریاضی‌دان آمده است. می‌توانید بر اساس این بخش خوشه‌بندی را انجام دهید و تحلیل این خوشه‌بندی را در گزارشی مختصر بیان کنید.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود نیز کار کنید.

۱. دسته‌بندی غذاها بر اساس استان مربوطه

طبق داده‌های موجود در وبگاه ویکی‌پدیا که در صورت تمرین سوم در اختیار شما قرار گرفت، هر غذا به یک یا چند استان اختصاص یافته‌است. یک مسئله‌ی دسته‌بندی می‌تواند این باشد که با دریافت اسم غذا و استفاده از سایر داده‌ها، استان مربوط به آن غذا را مشخص کنید.

۲. دسته‌بندی غذاها بر اساس سطح کالری

طبق داده‌های موجود در دیتاستی که در صورت پروپوزال پروژه در اختیار شما قرار گرفت، در فایل $PP\_recipes$ سطح کالری غذا با IDهای مشخص، داده شده‌است؛ یک مسئله‌ی دسته‌بندی می‌تواند این باشد که با دریافت مواد اولیه بکاررفته در غذا و در صورت نیاز استفاده از سایر داده‌ها، سطح کالری مربوط به آن غذا را مشخص کنید.

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود نیز کار کنید.

۱. خوشه‌بندی بر اساس مواد اولیه بکار رفته در غذا

برای تهیه‌ی غذاهای مختلف از مواد اولیه گوناگونی استفاده می‌شود. یک مسئله‌ی خوشه‌بندی می‌تواند این باشد که غذاها را بر اساس مواد اولیه بکار رفته در تهیه‌ی آن‌ها در خوشه‌های مختلف قرار دهید؛ دقت کنید که منظور ازینکه خوشه‌بندی را بر اساس مواد اولیه انجام دهید، این است که مواد اولیه بکار رفته در آن غذا قرار است نماینده‌ی متن شما باشد و شما عملا برای خوشه‌بندی نیاز به تمام متن دستور غذایی نخواهید داشت و تنها بر اساس مواد اولیه بکار رفته خوشه‌بندی را انجام می‌دهید. پس از انجام خوشه‌بندی، خوب است که تحلیلی از نتایج بدست آمده در این بخش نیز داشته‌باشید و خوشه‌های حاصل را از نظر ۲ معیار معرفی‌شده برای دسته‌بندی نیز بررسی کنید.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. دسته‌بندی صفحات بر اساس موضوع نوشته یا برچسب

هر صفحه‌ای که بازیابی می‌کنید می‌تواند موضوع متفاوتی داشته باشد که شما بر اساس این موضوع یک برچسب به صفحه نسبت می‌دهید. اگر به نظر شما یک صفحه می‌تواند برچسب‌های مختلفی داشته باشد باید برای آن صفحه چند برچسب مشخص کنید. مسئله حالا می‌تواند این باشد که برای هر صفحه که به عنوان ورودی به شما داده می‌شود، برچسب آن را مشخص کنید.

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش خوشه‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. خوشه‌بندی صفحات بر اساس ساختار و شباهت استایل

هر صفحه‌ی وب یک ساختار مشخص و یا استایل دارد که می‌تواند بر اساس آن، با دیگر صفحات مقایسه شده و شباهت آن با بقیه سنجیده شود. هر صفحه بر اساس این معیارها می‌تواند در یک خوشه قرار بگیرد. این می‌تواند یک مسئله‌ی خوشه‌بندی برای این موضوع باشد که بر اساس ساختار صفحات وب، آن‌ها را در خوشه‌های محتلف قرار دهید. می‌توانید پس از انجام مسئله‌ی خوشه‌بندی، نتایج به دست آمده را تحلیل کنید.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. دسته‌بندی متون بر اساس کاربران

در این بخش می‌بایست داده‌ها را از شبکه اجتماعی انتخابی خود استخراج کنید. به این صورت که اگر برای مثال شبکه اجتماعی توییتر انتخاب شده، توییت‌های مربوط به چند کاربر مشهور در توییتر را استخراج کنید. و هنگام دسته‌بندی مشخص کنید که هر متن مربوط به کدام کاربر است.

۲. اعمال تحلیل تمایل بر روی متون

در این بخش به داده برچسب خورده نیاز دارید. در نتیجه برای جمع‌آوری داده‌ها می‌توانید ازدیتاست‌های آماده ـنظیر sentiment۱۴۰ یا دیتاست‌های مربوط در این لینک ، و یا داده‌های آماده به زبان فارسیـ استفاده کنید. همچنین می‌توانید داده‌ها را جمع‌آوری کرده و برچسب بزنید. در نهایت باید با ورودی گرفتن یک توییت بتوانید مثبت بودن، منفی بودن و یا خنثی بودن آن را مشخص کنید.

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش خوشه‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. خوشه‌بندی متون

در این بخش داده‌های جمع‌آوری شده را به کمک روشی که در ابتدای تمرین گفته شده‌است، خوشه‌بندی کنید. سپس نتایج حاصل از خوشه‌بندی را تحلیل و گزارش کنید.

دسته‌بندی

دسته‌بندی داستان های شاهنامه

شاهنامه شامل ۹ داستان می باشد که از بین این ۹ داستان ۴ داستان دوازده رخ، اکوان دیو، رستم و اسفندیار، و سیاوش دارای تعداد ابیات کافی برای یادگیری یک مدل دسته بندی کننده را دارا هستند. در این بخش قصد داریم تا با گرفتن یک کوئری از کاربر تشخیص دهیم که مربوط به کدام داستان می باشد. لازم به ذکر است که نیاز است تا در تعیین داده های یادگیری و تست ملاحظات لازم را جهت رعایت توزیع متوازن داده ها،لحاظ فرمایید.

خوشه‌بندی

خوشه‌بندی ابیات شاهنامه

در این بخش لازم است تا ابیات شاهنامه را در فضای برداری، خوشه بندی کنید و به طور کلی مشخص کنید که موضوعات ابیات شاهنامه از نظر معنایی به چه خوشه هایی تقسیم می شوند. حتما خوشه ها را در فضای کاهش یافته به روش مناسب نمایش دهید. سپس لازم است تا کد شما برای جمله ی کوئری تشخیص دهد که جزو کدام خوشه می باشد.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. دسته‌بندی آیات بر اساس سوره‌ها

همانطور که می‌دانید قرآن بخش‌بندی دیگری که دارد بر اساس سوره است. در قرآن ۱۱۴ سوره وجود دارد که در طول سال‌های مختلف وحی شده است. اما ما در این قسمت به دلیل تعداد کلاس بالا و همگون نبودن طول سوره‌ها از تمام آنها استفاده نمی‌کنیم و تنها ۳۰ سوره‌ای از بقیه بیشترند (این مورد به دلخواه خودتان می‌تواند برحسب تعداد کلمه یا آیه‌ها باشد) را انتخاب کرده و بقیه را کنار می‌گذاریم. سپس با توجه به اطلاعاتی که از سوره‌های قرآن دارید آیات را لیبل بزنید و با استفاده از این داده آماده شده دسته‌بندی روی آیات مختلف را انجام دهید و در آخر

خوشه‌بندی

پیشنهاد زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش خوشه‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. خوشه‌بندی آیات بر اساس بخش های چهارگانه

در این قسمت شما باید سعی کنید چهار بخش مختلف را از درون آیات استخراج کنید که این کار را با استفاده از الگوریتم‌های خوشه بندی انجام می‌دهید.
بنا به نقلی از رسول خدا (صلی الله علیه وآله) و امام باقر (علیه السلام) ربع قرآن درباره اهل بیت (علیهم السلام)، ربع آن درباره دشمنان آن ها، ربع آن فرایض و احکام و ربع دیگر حلال و حرام است. حال در قسمت آخر قصد داریم با توجه به نتایجی که از خوشه‌بندی به دست آمده این حدیث را تحلیل کنیم و بررسی کنیم که آیا خوشه‌بندی انجام شده توانسته است به این حدیث نزدیک شود؟

۲. خوشه‌بندی سوره‌ها بر اساس مکی یا مدنی بودن

می‌دانیم که سوره‌های قرآن یا مدنی و یا مکی هستند. یک رویکر می‌تواند ان باشد که با خوشه‌بندی سوره‌های قرآن آیا می‌توان به چنین دسته‌هایی رسید؟

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. دسته‌بندی صفحات گیتهاب بر اساس زبان‌های برنامه‌نویسی

هر صفحه گیتهاب به طور خاص مشخص کرده است که به طور عمده از چه زبان برنامه‌نویسی‌ای در کد مورد نظر استفاده شده که می‌تواند label شما برای ادامه کار باشد. شما می‌توانید با استفاده از کدی که در مخزن گیتهاب وجود دارد، مخازن را بر اساس زبان‌های برنامه‌نویسی دسته‌بندی کنید.

۲. دسته‌بندی صفحات گیتهاب بر اساس requirements

برخی از صفحات گیتهاب نیازمندی‌ها را در فایل جداگانه‌ای به نام requirements.txt نام برده‌اند. می‌توانید این را label خود در نظر بگیرید و دسته‌بندی را با استفاده از کدی که در مخزن گیتهاب وجود دارد انجام دهید. برای این منظور می‌توانید داده‌های خود را محدودتر انتخاب کنید به طوری که دیتاست شما فقط شامل مخازنی باشد که حاوی این فایل هستند. همچنین می‌توانید نیازمندی‌ها را محدود در نظر بگیرید (به طور مثال ۱۰ نیازمندی در نظر بگیرید که بیشتر در کدها به کار رفته‌اند و متداول‌ترند).

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش خوشه‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. خوشه‌بندی با استفاده از فایل‌های Readme

می‌توانید با استفاده از فایل‌های Readme یک مسئله خوشه‌بندی تعریف کنید. توجه کنید که برخی از فایل‌های Readme می‌توانند حاوی لینک به مقاله مربوط به آن Repository باشند. می‌توانید از مقاله داده شده نیز برای دستیابی به نتیجه بهتر استفاده کنید. همچنین تمام مخازن حاوی فایل Readme نیستند؛ لذا فقط باید مخازنی را در نظر بگیرید که این فایل را دارا می‌باشند. در انتها تفسیری از این خوشه‌بندی ارائه دهید. نمایش گرافی خوشه‌بند را نیز انجام دهید و برای هر مخزن تنها نام آن مخزن را نمایش دهید.

۲. خوشه‌بندی با استفاده از کدهای موجود در مخازن گیتهاب

می‌توانید با استفاده از کدهای موجود در مخازن یک مسئله خوشه‌بندی تعریف کنید. در انتها تفسیری از این خوشه‌بندی ارائه دهید. نمایش گرافی خوشه‌بند را نیز انجام دهید و برای هر مخزن تنها نام آن مخزن را نمایش دهید.

دسته‌بندی

پیشنهادهای زیر برای دسته‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش دسته‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. در این بخش می‌توانید با استفاده از بردارهای تعبیه و برچسب‌های موجود روی دادگان معرفی شده، یک دسته‌بند روی مطالب آموزش دهید که بتواند با دریافت یک نوشتار، دسته‌ی مرتبط آن را مشخص کند.

خوشه‌بندی

پیشنهادهای زیر برای خوشه‌بندی ارائه می‌شود. هریک از گروه‌ها کافی‌ست تنها یک مورد از این پیشنهادها را برای بخش خوشه‌بندی انجام بدهد. علاوه بر این موضوعات، می‌توانید روی موضوع دلخواه خود بعد از هماهنگی با تدریس‌یار این track کار کنید.

۱. با استفاده از بردارهای تعبیه که در تمرین قبل آن‌ها را بررسی کردید می‌توان خوشه‌بندی‌هایی براساس محتوای اخبار و نوشته‌ها ارائه کرد. دادگان معرفی شده در پروپوزال این track شامل فیلدهای دسته و برچسب هستند (دادگان سامانه‌ی نمناک داری دسته‌های سلسله‌مراتبی نیز هستند) که می‌توانید از آن‌ها جهت بررسی امتیاز خوشه‌بندی با معیارهای معرفی شده یا ارائه‌ی خوشه‌بندی مدنظر خودتان استفاده کنید.