هوش مصنوعی **تمرین کامپیوتری شماره‌ی ۶**

مدرس: **دکتر فدایی** طراحان: نیما مدیر، نوید دهبان

مهلت تحویل: جمعه ۲۹ دی ۱۴۰۲، ساعت ۲۳:۵۹

مقدمه

خوشه بندی یا Clustering تکنیکی است که شامل گروه‌بندی اشیا مشابه بر اساس شباهت‌های ذاتی آن‌ها می‌شود. به عبارت دیگر، هدف آن است که نقاط داده را به خوشه‌های مجزا تقسیم کند، به صورتی که نقاط درون یک خوشه بیشتر به یکدیگر شباهت داشته باشند تا به خوشه های دیگر. با کشف این گروه‌بندی‌های طبیعی، الگوریتم‌های خوشه‌بندی می‌توانند بینش‌های ارزشمندی را در مورد ساختار زیربنایی داده‌ها ارائه دهند. خوشه‌بندی در حوزه‌های مختلفی از جمله تقسیم‌بندی مشتری، دسته‌بندی تصاویر و اسناد، تشخیص ناهنجاری و سیستم‌های توصیه کاربرد دارد.

تعریف مسئله

در این پروژه قصد داریم با استفاده از الگوریتم‌های Clustering، به تجزیه و تحلیل خبرهای سایت تحلیلی خبری عصر ایران[1] بپردازیم و سعی کنیم با استفاده از داده‌هایی که در اختیار داریم، آن‌ها را در دسته‌بندی‌های مختلف قرار دهیم، به طوری که بعد از اعمال الگوریتم خوشه‌بندی تا حد ممکن در خوشه درست خودشان قرار گرفته باشند.

آشنایی با مجموعه داده

مجموعه داده در فرمت csv در اختیار شما قرار گرفته است. در هر داده، متن خبر و همینطور دسته بندی آن خبر مشخص شده است. شما در انتها از دسته‌بندی‌های این مجموعه داده برای محاسبه دقت خوشه بندی خود استفاده خواهید کرد.

در این مجموعه داده شش دسته وجود دارند که به صورت زیر می باشد:

سلامت، سیاسی، ورزشی، فناوری، حوادث و فرهنگی/هنری

[1] https://www.asriran.com/

	label	content
0	فناوری	... گزارش های منتشر شده حاکی از آن است که کاربران
1	ورزشی	...سوپر استار سینما و از قهرمانان سابق ووشو ـ کو
2	حوادث	...مدیرعامل شرکت عمران آب کیش از فوت یک نفر در آت
3	فناوری	... یک نوجوان انگلیسی به اتهام هک حساب‌های کاربری
4	سلامت	...دانشمندان در جدیدترین مطالعات خود اثرات جدید و

دو فایل در اختیار شما قرار گرفته است که یکی برای آموزش و دیگری برای ارزیابی مدل شما است. فایل مربوط به آموزش مدل به عنوان train.csv و همینطور فایلی که مربوط به ارزیابی مدل شما است با نام test.csv در اختیار شما قرار گرفته است. دقت داشته باشید که تعداد سطرها به ازای هر موضوع دسته‌بندی در هر فایل به صورت متوازن قرار داده شده است و نیازی به یکسان کردن تعداد خبرها از دسته بندی های متفاوت یا resampling نیست. این کار برای از بین بردن bias موجود در داده هایی که تعداد کلاس های خروجی آن ها با هم برابر نیست استفاده می شود.

۱. در صورتی که داده ها نامتوازن بودند، چه مشکلاتی در فرآیند خوشه بندی پیش می آمد؟ چه راهکاری را برای برطرف کردن این مشکل ارائه می‌دهید؟ توضیح دهید.

فاز اول: پیش پردازش داده

در فاز اول باید اطلاعات متنی داخل مجموعه داده را برای تحلیل‌های بعدی پیش‌پردازش کنیم. برای این کار می‌توانید از کتابخانه‌ی Parsivar[2] یا هضم[3] استفاده کنید یا خودتان موارد مورد نیازتان را پیاده‌سازی کنید. شما باید عنوان و توضیحات‌هایی که موجود است را تا حد ممکن Normalize کنید(روش‌های ممکن، شامل حذف کلمات پرتکرار یا همان stop words، تبدیل کلمات به ریشه آنها و ... است).
دقت کنید که این کار هم روی داده‌های train و هم روی داده‌های test باید انجام شود و لزوماً اجرای هر نوع پیش‌پردازشی باعث بالا رفتن دقت مدل شما نخواهد شد. روش‌های متفاوت را با استفاده از کتابخانه یا بدون آن امتحان کنید و ترکیب هر کدام از آن‌ها که به مدل شما بیشتر کمک می‌کرد را اجرا کنید.

[2] https://github.com/ICTRC/Parsivar
[3] https://github.com/sobhe/hazm

البته به جز موارد توضیح داده شده می‌توانید تنها به حذف ایست واژه‌ها و کاراکترهای بی‌اهمیت مانند \n و \r بسنده کنید. اما لازم است تا تاثیر انواع دیگر پیش پردازش‌ها را نیز مشاهده کنید و در گزارش خود توضیحی در مورد آن‌ها ارائه دهید.

۲. در گزارش کار خود، جایگزین کردن کلمات با روش stemming یا lemmatization را توضیح دهید.

فاز دوم: فرایند مسئله

هدف کلی در این بخش استفاده از روش های clustering برای خوشه بندی متون دیتاست است.

ابتدا با استفاده از کتابخانه gensim و مدل doc2vec، یک مدل روی داده‌های آموزش پیش‌پردازش شده آموزش دهید و با استفاده از مدل آموزش داده شده، بردار ویژگی[4] داده‌های آموزش و تست را استخراج کنید.

در قدم بعدی، روی بردارهای ویژگی استخراج شده، با استفاده از روش‌های خوشه‌بندی که یاد گرفته‌اید(K-Means و DBSCAN)، داده‌هایتان را خوشه‌بندی کنید.

تمامی پارامترهای مدل‌های مورد استفاده دست شماست. توجه داشته باشید که در روش K-Means، انتخاب مقدار مناسب برای K اهمیت بسیاری دارد و احتمالاً با تعداد دسته‌های خبر باید تناسب داشته باشد. این موضوع در ارزیابی نتایج به شما کمک خواهد کرد.

۳. دلیل استفاده از بردار ویژگی و ویژگی‌های آن را در گزارش توضیح دهید.

۴. در مورد نحوه کار word2vec و doc2vec و تبدیل متن به بردار ویژگی توضیح دهید.

۵. در مورد روش های K-means و DBSCAN و مزایا و معایب این روش‌ها نسبت به هم توضیح دهید.

۶. خروجی حاصل از دو نوع خوشه‌بندی را باهم مقایسه کنید.

فاز سوم: کاهش بُعد(امتیازی)

در این بخش، می‌خواهیم خوشه‌های استخراج شده در فاز قبلی را نمایش دهیم. نکته مهمی که در نمایش خوشه‌ها وجود دارد، این است که معمولاً ابعاد بردار ویژگی زیاد بوده و همین موضوع باعث می‌شود که نتوان آن را در صفحه دو/سه‌بعدی به صورت مستقیم نمایش داد. برای حل این مشکل، از روش‌های کاهش بعد مثل PCA استفاده می‌شود.

۷. درباره PCA تحقیق کنید و نحوه عملکرد آن را به اختصار توضیح دهید.

حال روی بردارهای ویژگی بدست آمده کاهش بعد را انجام دهید و با استفاده از بردارهای کاهش یافته، خوشه‌ها را نمایش دهید و خوشه‌های بدست آمده توسط دو الگوریتم را با یکدیگر مقایسه کنید. برای کاهش بعد می‌توانید از کتابخانه sklearn استفاده کنید.

[4] Embedding

ارزیابی و تحلیل نتایج

در این بخش به ارزیابی نتایج حاصل از پیاده سازی روش‌ها می‌پردازیم. برای ارزیابی روش‌های خوشه‌بندی، می‌توان دقت خوشه‌بندی را با استفاده از دسته[5]های واقعی داده‌ها و بدون استفاده از آن اندازه‌گیری کرد. برای مطالعه این روش‌ها می‌توانید از این لینک استفاده کنید. برای روش‌های مبتنی بر true label، از معیار homogeneity و برای روش‌های غیر از آن از امتیاز silhouette استفاده می‌کنیم.

۸. در مورد نحوه محاسبه معیار silhouette و homogeneity توضیح دهید.

۹. نتایج حاصل از معیارهای ذکر شده را برای هر یک از روش‌ها گزارش کنید.

۱۰. راهکارهایی پیشنهاد کنید که بتوان عملکرد مدل‌ها را بهبود داد.

نکات پایانی

- توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. از ابزارهای تحلیل داده مانند نمودارها استفاده کنید.
- پس از مطالعه کامل و دقیق صورت پروژه، در صورت وجود هرگونه ابهام یا سوال با طراحان پروژه در ارتباط باشید.
- نتایج، گزارش و کدهای خود را در قالب یک فایل فشرده با فرمت AI_CA6_[stdNumber].zip در سامانه ایلرن بارگذاری کنید.
- محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- دقت کنید که نیازی به آپلود مجموعه داده‌ها در سامانه ایلرن نیست.
- هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید

موفق باشید

[5] Label