هوش مصنوعی **تمرین کامپیوتری شماره‌ی ۵**

مدرسین: **دکتر فدایی و دکتر یعقوب‌زاده**

طراحان: محمدرضا علوی، مبینا مهرآذر، صادق فاضلی

مهلت تحویل: دوشنبه ۲۴ دی ۱۴۰۳، ساعت ۲۳:۵۵

مقدمه

خوشه‌بندی 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 خوشه‌های 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.

توضیح مسئله

در این پروژه قصد داریم با استفاده از الگوریتم‌های Clustering، به تجزیه و تحلیل متن گروه‌های مختلف اخبار 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 خوشه‌ی درست خودشان قرار گرفته باشند.

آشنایی با مجموعه داده

مجموعه داده‌ای که در این پروژه استفاده می‌شود، یک مجموعه شامل مقالات خبری از اخبار روزانه در چند گروه خبری به زبان انگلیسی می‌باشد.

پیش‌پردازش و استخراج ویژگی

در این بخش باید اطلاعات متنی داخل مجموعه داده را برای تحلیل‌های بعدی پیش‌پردازش کنیم. برای این کار می‌توانید از کتابخانه‌های موجود استفاده کنید یا خودتان موارد مورد نیازتان را پیاده‌سازی کنید. در این بخش باید از روش‌های ممکن، شامل حذف کلمات پرتکرار یا همان stop words، تبدیل کلمات به ریشه آنها و ... استفاده کنید.

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.

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.

سپس باید با مدلی در ادامه گفته شده است به استخراج ویژگی‌ها از داده‌های متنی بپردازید.

1. در گزارش‌کار خود، جایگزین کردن کلمات با روش stemming یا lemmatization را توضیح دهید.
2. دلیل انجام پیش‌پردازش روی مجموعه داده متنی چیست؟
3. علت استخراج ویژگی‌ها چیست؟ چرا تنها به خواندن داده متنی بسنده نمی‌کنیم؟ توضیح دهید.

فرآیند مسئله

هدف کلی در این بخش استفاده از روش‌های clustering برای خوشه‌بندی متون دیتاست است.

ابتدا با استفاده از کتابخانه SentenceTransformers و مدل all-MiniLM-L6-v2، بردار ویژگی داده‌ها را استخراج کنید.

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

تمامی پارامترهای مدل‌های مورد استفاده دست شماست و سعی کنید با آزمون و خطا به پارامترهای مناسبی برسید. توجه داشته باشید که در روش K-Means، انتخاب K باید با تعداد دسته‌ها تناسب داشته باشد. در

نتیجه حتما از روش elbow method استفاده کرده و نمودار آن را نمایش دهید. این مقدار مناسب برای K اهمیت بسیاری دارد و احتمالا در ارزیابی نتایج به شما کمک خواهد کرد.

4. در مورد هر یک از روش‌های یادگیری Supervised و Unsupervised توضیح دهید و این دو روش را با یکدیگر مقایسه کنید.
5. دلیل استفاده از بردار ویژگی و ویژگی‌های آن را در گزارش توضیح دهید.
6. در مورد مجموعه مدل‌های Sentence Transformer و مدل all-MiniLM-L6-v2 به طور کلی و به اختصار توضیح دهید.
7. در مورد روش‌های K-means و DBSCAN و Hierarchical Clustering، نحوه کار آن‌ها و مزایا و معایب این روش‌ها را توضیح دهید.
8. روش استفاده از elbow method در روش K-means را توضیح دهید.
9. خروجی حاصل از این سه نوع خوشه‌بندی را با هم مقایسه کنید. کدام روش روی این مجموعه داده بهتر جواب داده است؟ دلیل آن چیست؟

کاهش بُعد

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

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

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

ارزیابی و تحلیل

در این بخش به ارزیابی نتایج حاصل از پیاده‌سازی روش‌ها می‌پردازیم. برای ارزیابی روش‌های خوشه‌بندی، می‌توان دقت خوشه‌بندی را با استفاده از دسته‌های واقعی داده‌ها و بدون استفاده از آن اندازه‌گیری کرد. برای مطالعه این روش‌ها می‌توانید از این لینک استفاده کنید. برای روش‌های مبتنی بر label true، از معیار

homogeneity و برای روش‌های غیر از آن از امتیاز silhouette استفاده می‌کنیم. در این قسمت از کتابخانه‌های موجود استفاده کنید.

شما باید بعد از اجرای هر روش در هر قسمت silhouette score مربوط به آن را با محاسبه کنید و نمایش دهید و نمودار مربوط به خوشه بندی آن را رسم کنید. همچنین پس از اجرای هر روش خوشه‌بندی، از هر خوشه سه نمونه خبر را چاپ کرده و گروه خبری آن ها را مقایسه کنید.

11. در مورد نحوه محاسبه معیار silhouette و homogeneity توضیح دهید.
12. نتایج حاصل از معیارهای ذکر شده را برای هر یک از روش‌ها گزارش کنید.

نکات پایانی

- دقت کنید که کد شما باید به نحوی زده شده باشد که نتایج قابلیت بازتولید داشته باشند.
- 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. حجم توضیحات گزارش شما هیچ گونه تاثیری در نمره نخواهد داشت و تحلیل و نمودارهای شما بیشترین ارزش را دارد.
- سعی کنید از پاسخ‌های روشن در گزارش خود استفاده کنید و اگر پیش‌فرضی در حل سوال در ذهن خود دارید، حتما در گزارش خود آن را ذکر نمایید.
- 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_CA5_[stdNumber].zip در سامانه ایلرن بارگذاری کنید. به طور مثال AI_CA5_810101999.zip
- محتویات پوشه باید شامل فایل پاسخ‌های شما به سوالات کتبی، فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- توجه کنید این تمرین باید به صورت تک‌نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد. در صورت مشاهده تقلب به همه افراد مشارکت‌کننده، نمره تمرین 100- و به استاد نیز گزارش می‌گردد. همچنین نوشته نشدن کدها توسط هوش مصنوعی نیز بررسی می‌شود!

موفق باشید