هوش مصنوعی **تمرین کامپیوتری شماره‌ی ۵**

مدرسین: **دکتر فدایی و دکتر یعقوب‌زاده**

طراحان: امیرحسین عارف‌زاده، مبینا مهرآذر، امیرعلی رحیمی

مهلت تحویل: جمعه 23 خرداد ۱۴۰۴، 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 ترانه‌های انگلیسی 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، ترانه‌ها تا حد ممکن در خوشه‌ی مناسب و با معنا قرار گرفته باشند.

آشنایی با مجموعه داده

مجموعه داده‌ای که در این پروژه استفاده می‌شود، یک مجموعه شامل متن ترانه‌های انگلیسی از چند ژانر مختلف موسیقی 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 (stop words)، تبدیل کلمات به ریشه آن‌ها (stemming یا lemmatization)، حذف علائم نگارشی و حروف بی‌اهمیت مانند $\n$ و $\r$ استفاده کنید. روش‌های متفاوت را امتحان کرده و ترکیب هر کدام از آنها که به مدل شما بیشتر کمک می‌کرد را اجرا کنید. در گزارش‌کار خود، توضیح دهید که کدام روش موثرتر بوده و چرا.

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

دلیل انجام پیش‌پردازش روی مجموعه داده متنی چیست؟

در مورد جایگزین کردن کلمات با روش stemming یا lemmatization را توضیح دهید.

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 مجموعه داده معرفی‌شده 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 (**K-Means**, **DBSCAN**, **Hierarchical Clustering**) داده‌ها را خوشه‌بندی 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.

در مورد هر یک از روش‌های یادگیری Supervised و Unsupervised توضیح دهید و این دو روش را با یکدیگر مقایسه کنید.

دلیل استفاده از بردار ویژگی و ویژگی‌های آن را در گزارش توضیح دهید.
در مورد روش‌های خوشه‌بندی فوق توضیح دهید.
در مورد روش‌های بردارسازی متنی و نحوه کار آن‌ها و مزایا و معایب این روش‌ها را توضیح دهید.

همینطور در مورد مجموعه مدل‌های Sentence Transformer و مدل all-MiniLM-L6-v2 به طور کلی و به اختصار توضیح دهید.
روش استفاده از elbow method در روش K-means را توضیح دهید.
خروجی حاصل از این نوع خوشه‌بندی‌ها را با هم مقایسه کنید. کدام روش روی این مجموعه داده بهتر جواب داده است؟ دلیل آن چیست؟

کاهش بُعد

بردارهای استخراج‌شده توسط مدل زبانی دارای تعداد زیادی ویژگی هستند. برای نمایش این بردارها به‌صورت دو یا سه‌بعدی (جهت تجسم بصری)، باید از روش‌های کاهش بعد استفاده کنیم. برای حل این مشکل، از روش‌های کاهش بعد مثل PCA استفاده می‌شود.

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 الگوریتم‌های فوق را با یکدیگر مقایسه کنید. برای کاهش بعد می‌توانید از کتابخانه sklearn استفاده کنید.

ارزیابی و تحلیل

ابتدا در گزارش راجب معیارهای زیر و نحوه محاسبه آنها توضیح داده و سپس از تمامی معیارهای مناسب برای تحلیل و ارزیابی نتایج حاصل از پیاده‌سازی روش‌های خوشه‌بندی این مسئله استفاده کنید. همچنین، دلایل عدم قابلیت استفاده از معیارهایی که کنار گذاشته شده‌اند را به‌طور دقیق و مستند بیان کنید.

- **Silhouette**
- **Homogeneity**

شما باید پس از اجرای هر روش خوشه‌بندی، معیارهای مربوطه را محاسبه کرده و نتایج آن‌ها را نمایش دهید. همچنین نمودار مربوط به خوشه‌بندی را رسم کرده و تحلیل نمایید.

علاوه بر این، از هر خوشه دو نمونه چاپ کنید و محتوای آن‌ها را از نظر شباهت معنایی و موضوعی مقایسه نمایید.

در نهایت، با تحلیل نتایج به‌دست‌آمده، روش برتر را از نظر عملکرد و کیفیت خوشه‌بندی انتخاب کرده و دلایل انتخاب خود را شرح دهید.

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_810102999.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