هوش مصنوعی تمرین کامپیوتری شماره‌ی 5

مدرسین: **دکتر فدایی و دکتر یعقوب‌زاده**

طراحان: مبینا مهرآذر، امیرحسین عارف زاده، آریا عازم

مهلت تحویل: ۱۸ دی ۱۴۰۴، ساعت ۲۳:۵۹

مقدمه

خوشه‌بندی (Clustering) یکی از تکنیک‌های مهم در Unsupervised Learning است که شامل گروه‌بندی اشیای مشابه بر اساس شباهت‌های ذاتی آن‌ها می‌شود. هدف اصلی در خوشه‌بندی، تقسیم نقاط داده به خوشه‌های مجزا به‌گونه‌ای است که داده‌های درون هر خوشه بیشترین شباهت را به یکدیگر و کمترین شباهت را به داده‌های خوشه‌های دیگر داشته باشند.

با کشف این ساختارهای طبیعی در داده‌ها، الگوریتم‌های خوشه‌بندی می‌توانند اطلاعات ارزشمندی درباره الگوها و روابط پنهان موجود در داده‌ها ارائه دهند. خوشه‌بندی در حوزه‌های مختلفی از جمله تحلیل نظرات کاربران، سیستم‌های توصیه‌گر، دسته‌بندی اسناد متنی، تحلیل احساسات، تشخیص ناهنجاری و تحلیل رفتار مشتری کاربرد گسترده‌ای دارد. در ادامه، با استفاده از برخی الگوریتم‌های خوشه‌بندی، این کاربردها را مشاهده می‌کنیم.

توضیح مسئله

در این پروژه قصد داریم با استفاده از الگوریتم‌های Clustering، به تجزیه و تحلیل نظرات متنی کاربران درباره محصولات غذایی آمازون (Amazon Food Reviews) بپردازیم. هدف آن است که با استفاده از داده‌های متنی موجود (نظرات کاربران)، این نظرات را به خوشه‌هایی معنادار تقسیم کنیم. به‌گونه‌ای که هر خوشه بیانگر یک موضوع، الگوی معنایی یا نوع خاصی محصولات باشد. پس از اعمال الگوریتم‌های خوشه‌بندی، انتظار می‌رود نظراتی که از نظر محتوای معنایی به یکدیگر نزدیک هستند، در یک خوشه قرار گیرند.

آشنایی با مجموعه داده

مجموعه‌داده مورد استفاده در این پروژه، دیتاست Amazon Fine Food Reviews است. در این پروژه، داده‌ها به‌صورت مجموعه‌ای از reviewها در نظر گرفته شده‌اند که هر review شامل اطلاعات توصیفی مربوط به کاربر و محصول، و همچنین داده‌های متنی شامل خلاصه و متن کامل نظر می‌باشد. در ادامه، جدول آشنایی با فیلدهای مختلف مجموعه‌داده و یک نمونه از داده‌های موجود ارائه شده است.

Field Name	توضیحات
Id	شناسه یکتای هر نظر
UserId	شناسه یکتای کاربر ثبت‌کننده نظر
ProfileName	نام نمایشی کاربر در آمازون
HelpfulnessNumerator	تعداد کاربرانی که این نظر را مفید ارزیابی کرده‌اند
HelpfulnessDenominator	تعداد کل رأی‌دهندگان به مفید بودن یا نبودن نظر
Score	امتیاز عددی (بین 1 تا 5) داده‌شده به محصول توسط کاربر
Time	زمان ثبت نظر به‌صورت Timestamp
Summary	خلاصه کوتاه نظر کاربر که ایده اصلی نظر را بیان می‌کند
Text	متن کامل نظر کاربر شامل توضیحات و تجربه مصرف محصول

```
Id: 329849
UserId: A1YZYHCKDK2K4I
ProfileName: Santorini
HelpfulnessNumerator: 2
HelpfulnessDenominator: 2
Score: 5
Time: 1328832000
Summary: My Pooch loves them
Text: Since my pet does not like having his teeth cleaned, I have...
```

پیش‌پردازش و استخراج ویژگی‌ها

در این بخش باید اطلاعات متنی موجود در مجموعه داده را برای تحلیل‌های بعدی پیش‌پردازش کنیم. برای این کار می‌توانید از کتابخانه‌های آماده استفاده کنید یا مراحل لازم را خودتان پیاده‌سازی نمایید. همچنین بخش‌های غیرمرتبط داده که کمکی به دسته‌بندی نظرات نمی‌کنند را حذف کنید.

مراحل پیش‌پردازش شامل (اما نه محدود به) موارد زیر است:

- حذف کلمات توقف (Stop Words)
- حذف علائم نگارشی و کاراکترهای غیرضروری
- تبدیل حروف به حالت یکسان (Lowercase)
- ریشه‌یابی کلمات با استفاده از Stemming یا Lemmatization

روش‌های مختلف پیش‌پردازش را امتحان کرده و ترکیب‌هایی از آن‌ها را بررسی کنید.

سوالات:

۱. کدام روش یا ترکیب روش‌های پیش‌پردازش عملکرد بهتری داشته است؟

۲. دلیل انجام پیش‌پردازش روی داده‌های متنی چیست؟

۳. تفاوت Stemming و Lemmatization را توضیح دهید و مزایا و معایب هرکدام را بیان کنید.

۴. چرا استخراج ویژگی ضروری است و نمی‌توان مستقیما از متن خام برای خوشه‌بندی استفاده کرد؟

فرایند حل مسئله

ابتدا با استفاده از کتابخانه SentenceTransformers و مدل all-MiniLM-L6-v2 بردارهای ویژگی متنی را از متن نظرات (Text یا ترکیب Summary و Text) استخراج کنید. سپس روی بردارهای استخراج‌شده، الگوریتم‌های خوشه‌بندی K-means, DBSCAN, Hierarchical Clustering را پیاده‌سازی کنید.

تمام پارامترهای مدل‌ها (مانند تعداد خوشه‌ها، min_samples، eps و ...) به اختیار شماست و باید با آزمون و خطا مقادیر مناسب را انتخاب کنید. همچنین در گزارش خود مراحل آزمون و خطا و معیار انتخاب پارامترها را بنویسید. همچنین 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.

سوالات:

۱. روش‌های Supervised Learning و Unsupervised Learning را توضیح دهید و آن‌ها را با یکدیگر مقایسه کنید.

۲. دلیل استفاده از بردارهای ویژگی (Embeddings) چیست؟ ویژگی‌های این بردارها شامل چه چیزهایی هستند؟

۳. روش‌های بردارسازی متن را توضیح دهید (Bag of Words، TF-IDF، Word Embeddings، Sentence Embeddings) و آن‌ها را از نظر کیفیت معنایی مقایسه کنید.

۴. درباره مدل‌های Sentence Transformer و مدل all-MiniLM-L6-v2 توضیح دهید.

۵. روش Elbow Method چیست و چگونه برای انتخاب تعداد خوشه‌ها در K-Means استفاده می‌شود؟

۶. روش‌های خوشه‌بندی استفاده شده را توضیح دهید.

کاهش بعد (Dimensionality Reduction)

بردارهای استخراج‌شده از مدل‌های زبانی دارای ابعاد بالا هستند. برای نمایش داده‌ها به‌صورت دو یا سه‌بعدی و تجسم بصری خوشه‌ها، لازم است از روش‌های کاهش بُعد استفاده شود. در این پروژه از روش **PCA** و یک روش دیگر به دلخواه استفاده کنید. خوشه‌های حاصل از هر الگوریتم را با استفاده از داده‌های کاهش‌یافته نمایش دهید و نتایج بصری خوشه‌بندی الگوریتم‌های مختلف را با یکدیگر مقایسه کنید.

به این صورت عمل کنید که بعد از کاهش بعد توسط دو روش، دوباره الگوریتم‌های خوشه‌بندی را اجرا کرده و نتایج را تحلیل کنید.

سوالات:

۱. درباره PCA تحقیق کرده و نحوه عملکرد آن را به‌طور خلاصه توضیح دهید.

۲. عملکرد دو روش کاهش بعد استفاده شده را مقایسه کنید.

ارزیابی و تحلیل نتایج

در این بخش باید نتایج خوشه‌بندی را به‌صورت کمی و کیفی ارزیابی کنید. ابتدا معیارهای ارزیابی خوشه‌بندی را معرفی کرده و نحوه محاسبه آن‌ها را توضیح دهید. حالا از معیارهای Silhouette و Homogeneity استفاده کنید. اگر از معیاری نمی‌توانید استفاده کنید، دلیل مناسب نبودن آن را بنویسید.
پس از اجرای هر الگوریتم خوشه‌بندی، مقادیر معیارها را محاسبه و گزارش کنید، نمودار خوشه‌بندی را رسم کرده و تحلیل نمایید.
نتایج خوشه‌بندی بعد از انجام کاهش بعد را با نتایج قبل از آن مقایسه کنید.
علاوه بر این، از هر خوشه حداقل 4 نمونه چاپ کنید و آن‌ها را از نظر شباهت معنایی و موضوعی مقایسه کنید.
در آخر بگویید که با توجه به نتایج بدست آمده، کدام روش خوشه‌بندی در این مسئله عملکردی بهتری داشته و دلایل برتری آن 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. حجم توضیحات گزارش شما هیچ گونه تاثیری در نمره نخواهد داشت و تحلیل 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.
- فایل‌های خود را در قالب یک فایل فشرده با فرمت AI_CA5_[stdNum].zip در سامانه ایلرن بارگذاری کنید. به طور مثال AI_CA5_810101234.zip.
- محتویات پوشه باید شامل گزارش و کدهای شما باشد.

موفق باشید