گزارش کار پروژه 5

هوش مصنوعی

کسری کاشانی

810101490

سوال 1) هر دو روش Stemming و Lemmatizing برای کاهش کلمات به ریشه یا فرم پایه ی خود استفاده می‌شوند. Stemming فرآیندی است که کلمه را به ریشه ی خود تقلیل می‌دهد، اما این کار را صرفاً با حذف یا تغییر حروف انتهایی کلمه انجام می‌دهد، بدون در نظر گرفتن معنی دقیق کلمه. اما lemmatizing فرآیندی است که کلمات را به فرم لغوی پایه ای خود تقلیل می‌دهد، با در نظر گرفتن معنی کلمه و نقش دستوری آن با کمک گرفتن از dictionary.

در نتیجه، روش Lemmatizing که ما از آن استفاده کردیم، دارای دقت بیشتر اما سرعت کمتری نسبت به روش Stemming می باشد.

سوال 2) پیش پردازش متون یکی از مراحل مهم در پردازش زبان طبیعی (NLP) است که باعث بهبود کیفیت داده‌ها برای تحلیل و استفاده در مدل های یادگیری ماشین می‌شود. برخی از دلایل اصلی انجام پیش پردازش بر روی داده‌های متنی عبارتند از:

1. کاهش نویز و افزایش کیفیت داده‌ها: داده‌های خام متنی اغلب شامل نویز هستند (مانند کاراکترهای غیرضروری، علائم نگارشی، لینک‌ها، اعداد و موارد مشابه).
2. کاهش ابعاد ویژگی ها: پیش پردازش با روش‌هایی مثل حذف کلمات غیرضروری (Stop Words)،

Stemming، و Lemmatization، تعداد ویژگی‌ها (کلمات) را کاهش می دهد و پردازش را ساده تر می‌کند.

3. حذف اطلاعات غیرضروری: متون ممکن است شامل کلمات و عباراتی باشد که تأثیر چندانی در تحلیل ندارند. (مانند "is", "and", "the")
4. آماده‌سازی برای الگوریتم‌های یادگیری ماشین: مدل های یادگیری ماشین نمی توانند مستقیماً متن خام را پردازش کنند. متن باید به شکل عددی یا برداری تبدیل شود.

سوال 3) استخراج ویژگی‌ها از داده‌های متنی یکی از مراحل اساسی در پردازش زبان طبیعی (NLP) و یادگیری ماشین است. همچنین به علت بی ساختار بودن داده های خام، حجم زیاد و نویز، نیاز به تبدیل متن به اعداد دلایل نمی توانیم به خواندن متن خام بسنده کنیم. اهمیت این فرآیند به شرح زیر است:

1. غیر قابل درک بودن داده های متنی برای کامپیوترها: داده‌های متنی در حالت خام، برای ماشین ها و الگوریتم های یادگیری ماشین قابل استفاده نیستند، زیرا متن ها به صورت string هستند و الگوریتم ها نیاز به داده هایی دارند که به صورت عددی یا برداری باشند. استخراج ویژگی‌ها متن را به نمایه های عددی تبدیل می‌کند که الگوریتم‌ها بتوانند پردازش کنند.

2. کاهش پیچیدگی و برجسته سازی اطلاعات مفید: داده های متنی ممکن است شامل نویز و اطلاعات غیر ضروری باشند. استخراج ویژگی‌ها به تمرکز بر اطلاعات کلیدی و حذف نویز کمک می‌کند.
3. استخراج روابط معنایی و ساختاری: ویژگی ها می توانند اطلاعات معنایی و نحوی متن را استخراج کنند.

سوال 4) در یادگیری نظارت شده، مدل با استفاده از داده‌های برچسب دار آموزش می بیند. این یعنی هر نمونه در داده های آموزشی، شامل ورودی و خروجی مشخص است، و هدف مدل پیشبینی خروجی برای ورودی‌های جدید است. از کاربرد های این نوع این یادگیری می توان به پیشبینی، طبقه بندی و رگرسیون اشاره کرد. همچنین مزایای آن پیشبینی دقیق خروجی‌ها و معایب آن نیاز به داده های برچسب دار می باشد.

در یادگیری نظارت نشده، مدل با استفاده از داده‌های بدون برچسب آموزش می بیند. هدف این روش کشف الگو ها، ساختارها یا گروه‌ها در داده ها بدون داشتن برچسب های مشخص است. از کاربرد های این نوع این یادگیری می توان به خوشه بندی، بخش بندی داده ها و کشف روابط پنهان اشاره کرد. همچنین مزایای آن کارآمدی در داده های بدون برچسب و معایب آن دشواری در ارزیابی و تفسیر نتایج می باشد.

سوال 5) بردار ویژگی، نمایشی عددی از داده های اصلی است که اطلاعات مهم و مرتبط با داده را در قالب اعداد فشرده می‌کند. این بردار یکی از اجزای اصلی در پردازش داده ها و یادگیری ماشین است. همچنین از ویژگی های بردار ویژگی می توان به عددی و قابل پردازش بودن توسط الگوریتم ها، دارای ابعاد مشخص و تعریف شده، فشرده سازی داده ها، امکان استفاده در الگوریتم های مختلف اشاره کرد. برخی از دلایل استفاده از بردار ویژگی:

1. تبدیل داده های خام به فرمت قابل پردازش برای مدل ها: داده های خام (مانند متن، تصویر یا صدا) معمولاً غیرقابل پردازش توسط الگوریتم های یادگیری ماشین هستند. بردار ویژگی، داده های خام را به یک نمایش عددی تبدیل می کند که قابل پردازش توسط مدل ها است.
2. کاهش ابعاد و حذف نویز: داده های خام معمولاً شامل حجم زیادی از اطلاعات غیر ضروری هستند. با انتخاب ویژگی‌های کلیدی، بردار ویژگی به کاهش ابعاد داده و حذف نویز کمک می کند.
3. افزایش دقت و کارایی مدل ها: بردار ویژگی اطلاعات مهم داده ها را حفظ می کند و از ویژگی های غیرضروری صرف نظر می کند. این کار باعث می شود مدل بتواند بهتر الگوها را یاد بگیرد و عملکرد بهتری داشته باشد.

سوال 6) مدل Sentence Transformer یک معماری یادگیری عمیق است که برای تولید بردارهای معنایی از جملات یا متون طراحی شده است که هدف آن، نگاشت جملات یا متون به فضای برداری است به طوری که متن های مشابه، به بردارهایی نزدیک به هم نگاشت شوند. مدل all-MiniLM-L6-v2 یکی از نسخه های پیش پرداخته شده ی این معماری با دقت و سرعت بالا است که به طور خاص برای کاربرد های متداول مانند جستجوی معنایی، تطابق جملات، و خوشه بندی متون بهینه سازی شده است. تمام جملات یا متون به بردارهایی با طول 384 تبدیل می‌شوند.

سوال 7) در روش K-Means، تعداد خوشه ها یعنی k از قبل دریافت می شود. سپس k نقطه تصادفی به عنوان مراکز اولیه خوشه‌ها انتخاب می شوند و هر داده به نزدیک ترین مرکز خوشه تخصیص داده می شود. سپس مراکز خوشه ها، بر اساس میانگین نقاط در هر خوشه، به روز رسانی می شوند. این مراحل تا زمانی که تغییرات جزئی شوند یا به حداکثر تکرار برسند، تکرار می شوند. مزایای این روش سادگی و سرعت بالا و معایب آن نیاز به مشخص کردن تعداد خوشه ها و حساسیت به نقاط دورافتاده و تنها مناسب بودن برای خوشه های با شکل دایره‌ای یا کروی می باشد.

در روش DBSCAN، داده‌ها بر اساس تراکم نقاط خوشه بندی می شوند. برای هر نقطه، تعداد نقاط موجود در یک شعاع مشخص یا همان epsilon محاسبه می‌شود. همچنین اگر تعداد نقاط بیشتر از

حداقل تعداد مورد نیاز باشد، نقطه به عنوان نقطه ی هسته‌ای در نظر گرفته می شود و نقاط نزدیک به نقاط هسته ای به همان خوشه تخصیص داده می شوند. همچنین نقاطی که به هیچ خوشه ای تعلق ندارند به عنوان نویز برچسب گذاری می‌شوند. مزایای این روش عدم نیاز به مشخص کردن تعداد خوشه ها و قابلیت تشخیص خوشه های با اشکال پیچیده و مقاوم بودن در برابر نقاط دورافتاده و معایب آن عملکرد ضعیف در داده‌هایی با تراکم متغیر و کارایی کمتر نسبت به ابعاد بالاتر می باشد.

در روش Hierarchical، خوشه ها به صورت سلسله مراتبی و درخت مانند ایجاد می شوند. هر داده به عنوان یک خوشه ی جداگانه شروع می‌شود. خوشه های نزدیک به هم ادغام می شوند تا زمانی که همه داده ها در یک خوشه قرار گیرند. در نهایت نتیجه به صورت یک دندروگرام نمایش داده می شود. مزایای این روش عدم نیاز به مشخص کردن تعداد خوشه ها و مناسب برای داده‌های کوچک و تحلیل سلسله و معایب آن حساسیت به نویز و نقاط دور افتاده و نیاز به انتخاب مناسب معیار فاصله می باشد.

سوال 8) elbow method یکی از روش‌های رایج برای تعیین تعداد بهینه خوشه ها یا k در الگوریتم K-Means است. این روش با استفاده از تحلیل inertia، تعداد مناسب خوشه ها را شناسایی می‌کند. inertia معیاری برای ارزیابی کیفیت خوشه بندی است و برابر با مجموع فاصله نقاط هر خوشه از مرکز خوشه خود است.

در این روش، نمودار k بر حسب inertia رسم می شود و نقطه ای که در آن کاهش inertia به طور چشمگیری کاهش می‌یابد و نمودار برای اولین بار دچار شکست می شود، به عنوان تعداد بهینه خوشه ها انتخاب می شود.

سوال 9) امتیاز سیلوئت معیاری برای ارزیابی کیفیت خوشه بندی است. این امتیاز میزان جدایی بین خوشه ها و فشردگی درون هر خوشه را اندازه گیری می‌کند و عددی بین -1 و 1 است. هرچه امتیاز سیلوئت بالاتر و به 1 نزدیکتر باشد، خوشه بندی بهتر است.

لذا با توجه به امتیاز سیلوئت، خوشه بندی K-Means عملکرد بهتری داشته است.

سوال 10) PCA یک تکنیک کاهش ابعاد است که برای خلاصه کردن داده های با ابعاد بالا به تعداد کمتری از ابعاد استفاده می‌شود، در حالی که حداکثر اطلاعات ممکن را حفظ می‌کند. این تکنیک از جبر خطی برای تبدیل داده ها استفاده می کند و اطلاعات موجود در چندین ویژگی را به یک فضای مختصر و جدید نگاشت می‌کند. فرآیند کلی PCA شامل مراحل استاندار سازی داده ها، محاسبه ماتریس کوواریانس، محاسبه مقادیر ویژه و بردار های ویژه، انتخاب مولفه های اصلی و در نهایت تبدیل داده ها به فضای جدید می باشد.

سوال 11) معیار silhouette میزان خوب بودن خوشه بندی را با استفاده از فاصله بین نقاط آن خوشه و خوشه های دیگر اندازه گیری می‌کند. این معیار برای هر داده محاسبه می شود و سپس میانگین آن برای کل داده ها به عنوان معیار کلی در نظر گرفته می شود. دامنه آن نیز بین -1 تا 1 می باشد.

معیار homogeneity بررسی می کند که آیا تمام نمونه های یک خوشه به یک برچسب کلاس مشابه تعلق دارند یا نه. این معیار به برچسب های واقعی داده ها نیاز دارد و به عنوان یک معیار بیرونی استفاده می شود. دامنه آن نیز بین 0 تا 1 می باشد.

سوال 12)

K-Means: دارای امتیاز سیلوئت 0.47522180752131005 که قابل قبول است اما همچنان می‌توان آن را بهبود داد.

DBSCAN: دارای امتیاز سیلوئت -0.5927740405414702 که غیرقابل قبول است و نشان می‌دهد که این روش برای این داده ها یا تنظیمات فعلی مناسب نیست.

Hierarchical: دارای امتیاز سیلوئت 0.39453516996540244 که قابل قبول است اما خیلی می‌توان آن را بهبود داد.