گزارش کار پروژه پنجم AI

نرگس بابالار 810101557

سوال1. در گزارش کار خود، جایگزین کردن کلمات با روش **lemmatization** یا **stemming** را توضیح دهید.

**Lemmatization** و **Stemming** هر دو روش‌هایی برای کاهش فرم کلمات به شکل پایه هستند، اما تفاوت‌هایی میان این دو وجود دارد:

1. **Stemming**:
   * کلمات را به ریشه یا پایه خود باز می‌گرداند، اما ممکن است این ریشه الزاماً یک کلمه معنادار نباشد.
   * برای انجام آن از الگوریتم‌های ساده‌ای استفاده می‌شود که اغلب پسوندها و پیشوندها را حذف می‌کنند.
   * مثال: کلمه "running" به "run" تبدیل می‌شود.
2. **Lemmatization**:
   * کلمات را به فرم پایه‌ای که در زبان قابل قبول و معنادار است، تبدیل می‌کند.
   * معمولاً نیاز به دانش زبانی و فرهنگ لغت دارد.
   * مثال: "running" به "run" و "better" به "good" تبدیل می‌شود.

**کاربرد در پیش‌پردازش داده‌ها:**

* **Stemming** سریع‌تر است و بیشتر در کاربردهایی که دقت خیلی بالا نیاز نیست، استفاده می‌شود.
* **Lemmatization** دقیق‌تر است و برای تحلیل‌های زبانی پیچیده یا در شرایطی که نیاز به معنی دقیق کلمات وجود دارد، مناسب‌تر است.

**سوال2**. دلیل انجام پیش‌پردازش روی داده‌های متنی چیست؟

**کاهش پیچیدگی و حجم داده‌ها**:

* داده‌های متنی خام معمولاً شامل اطلاعات زائد مانند کلمات پرتکرار (Stop Words)، اعداد، کاراکترهای خاص و علائم نگارشی هستند. حذف این اطلاعات به کاهش حجم داده‌ها کمک می‌کند و پردازش سریع‌تر می‌شود.

**یکنواخت‌سازی داده‌ها**:

* داده‌های متنی ممکن است شامل کلمات در فرم‌های مختلف (مانند زمان‌های مختلف فعل یا جمع و مفرد بودن اسامی) باشد. استفاده از روش‌هایی مانند Stemming یا Lemmatization داده‌ها را یکدست می‌کند.

**بهبود دقت مدل‌های یادگیری ماشین**:

* پیش‌پردازش کمک می‌کند نویزهای موجود در داده‌ها حذف شوند و مدل بتواند الگوهای واقعی را به‌درستی یاد بگیرد.

**افزایش کیفیت ویژگی‌های استخراج شده**:

* پیش‌پردازش داده‌های متنی به استخراج ویژگی‌های کلیدی و معنادار کمک می‌کند که باعث افزایش دقت در مدل‌سازی و تحلیل‌های بعدی می‌شود.

**آماده‌سازی داده‌ها برای الگوریتم‌های خاص**:

* بسیاری از الگوریتم‌های پردازش زبان طبیعی (NLP) نیازمند ورودی‌های استاندارد و ساختارمند هستند که از طریق پیش‌پردازش داده‌ها فراهم می‌شود.

**سوال3.** علت استخراج ویژگی‌ها از داده‌های متنی چیست؟ چرا تنها به خواندن داده متنی بسنده نمی‌کنیم؟

**استخراج ویژگی‌ها** از داده‌های متنی یکی از مراحل اصلی در پردازش زبان طبیعی (NLP) و یادگیری ماشین است. دلایل اهمیت این مرحله عبارتند از:

1. **نمایش داده‌های متنی در قالب عددی (Numerical Representation)**:
   * مدل‌های یادگیری ماشین و الگوریتم‌های تحلیلی نمی‌توانند با داده‌های متنی خام کار کنند. داده‌های متنی باید به فرم عددی (مانند بردارهای ویژگی) تبدیل شوند تا مدل بتواند آن‌ها را پردازش کند.
2. **استخراج اطلاعات معنادار**:
   * متن خام معمولاً شامل اطلاعات غیرضروری یا نویز است. استخراج ویژگی‌ها به شناسایی و تمرکز بر روی جنبه‌های کلیدی و معنادار متن کمک می‌کند.
3. **بهبود عملکرد مدل‌ها**:
   * ویژگی‌های استخراج شده، داده‌های متنی را به شکلی تبدیل می‌کنند که مدل بتواند بهتر الگوها و ارتباطات را یاد بگیرد، که این امر منجر به بهبود دقت و کارایی مدل‌ها می‌شود.
4. **کاهش ابعاد داده‌ها**:
   * داده‌های متنی می‌توانند بسیار بزرگ و پیچیده باشند. استخراج ویژگی‌ها به کاهش حجم و پیچیدگی داده‌ها کمک می‌کند و پردازش را سریع‌تر و کارآمدتر می‌سازد.
5. **یکنواخت‌سازی داده‌ها**:
   * ویژگی‌های استخراج شده معمولاً داده‌ها را به یک فضای مشخص و قابل مقایسه تبدیل می‌کنند، که این موضوع در تحلیل‌ها و مقایسه‌ها بسیار مفید است.
6. **پشتیبانی از الگوریتم‌های مختلف**:
   * الگوریتم‌های خوشه‌بندی و دسته‌بندی نیازمند ورودی‌هایی هستند که نشان‌دهنده ویژگی‌های متمایز داده باشند. بدون استخراج ویژگی، اجرای این الگوریتم‌ها امکان‌پذیر نیست.

**سوال4. در مورد هر یک از روش‌های یادگیری Supervised و Unsupervised توضیح دهید و این دو روش را با یکدیگر مقایسه کنید.**

**یادگیری Supervised:**  
این روش شامل یادگیری مدل از داده‌های دارای برچسب (Label) است. مدل با استفاده از ورودی‌ها و خروجی‌های مشخص، الگوها را یاد می‌گیرد و هدف آن پیش‌بینی نتایج برچسب‌دار برای داده‌های جدید است.  
**مثال:** رگرسیون خطی، دسته‌بندی تصویر.

**یادگیری Unsupervised:**  
در این روش، داده‌ها فاقد برچسب هستند و مدل باید بدون کمک برچسب، ساختارهای پنهان در داده‌ها را شناسایی کند. خوشه‌بندی نمونه‌ای از این روش است.  
**مثال : ,** K-mean DBSCAN

**مقایسه:**

* Supervised نیازمند داده‌های برچسب‌دار است، در حالی که Unsupervised بدون برچسب کار می‌کند.
* Supervised دقت بالاتری برای پیش‌بینی دارد، اما Unsupervised در کشف الگوهای ناشناخته قدرتمند است.

**سوال5.** **دلیل استفاده از بردار ویژگی‌ها در گزارش توضیح دهید.**

بردار ویژگی‌ها (Feature Vectors) به مدل کمک می‌کند تا داده‌ها را به شکل عددی نمایش داده و شباهت‌ها یا تفاوت‌های موجود در داده‌ها را محاسبه کند. کامپیوترها نمی‌توانند رشته‌های متنی را به‌خوبی پردازش کنند، اما بردارها (که عددی هستند) را سریع‌تر و دقیق‌تر پردازش می‌کنند.این بردارها:

1. نمایش‌دهنده جنبه‌های کلیدی داده هستند.
2. فرایند خوشه‌بندی یا دسته‌بندی را تسهیل می‌کنند.

**سوال6. در مورد مجموعه مدل‌های Sentence Transformer و مدل all-MiniLM-L6-v2 به طور کلی و به اختصار توضیح دهید.**

این مجموعه مدل‌ها برای تبدیل متن به بردار ویژگی (Embedding) استفاده می‌شوند. مدل‌ها مثل all-MiniLM-L6-v2 قادر به محاسبه شباهت معنایی بین جملات هستند و در کارهایی مانند خوشه‌بندی متن و بازیابی اطلاعات استفاده می‌شوند.  
مدل all-MiniLM-L6-v2:  
یک مدل کم‌حجم و سریع از خانواده Sentence Transformers است که تعادل خوبی بین دقت و کارایی دارد.

**سوال7. در مورد روش‌های K-means، DBSCAN و Hierarchical Clustering، نحوه کار آن‌ها و مزایا و معایب این روش‌ها را توضیح دهید.**

K-means:

* نحوه کار:  
  داده‌ها را به k گروه تقسیم می‌کند با هدف کاهش مجموع فاصله‌ها از مراکز کلاسترها.
* مزایا: ساده، سریع.
* معایب: حساس به مقدار اولیه و outliers.

DBSCAN:

* نحوه کار:  
  براساس تراکم، نقاطی که در فاصله مشخصی (epsilon) از تعداد کافی نقاط (min\_samples) هستند را خوشه‌بندی می‌کند.
* مزایا: مناسب برای داده‌های با نویز.
* معایب: تنظیم دشوار پارامترها.

Hierarchical Clustering:

* نحوه کار:  
  خوشه‌ها را در سطوح مختلف به صورت درختی ایجاد می‌کند.
* روش تجمعی (Agglomerative): از نقاط منفرد شروع کرده و خوشه‌ها را یکی‌یکی ادغام می‌کند.
* روش تقسیم‌کننده (Divisive): از یک خوشه کلی شروع کرده و به تدریج خوشه‌ها را تقسیم می‌کند.
* مزایا: نمایش سلسله‌مراتبی.
* معایب: محاسبات سنگین برای داده‌های بزرگ.

**سوال8. روش استفاده از elbow method در روش K-means را توضیح دهید.**

Elbow Method:  
این روش مقدار بهینه k در K-means را مشخص می‌کند. نمودار مجموع مربعات خطاها (SSE) بر حسب k رسم می‌شود و نقطه "زانو" به‌عنوان مقدار بهینه انتخاب می‌شود.

برای این کار :

* محاسبه مقدار Inertia (واریانس درون خوشه‌ها)  
  Inertia معیاری برای ارزیابی کیفیت خوشه‌بندی است. این مقدار مجموع فاصله نقاط هر خوشه از مرکز خوشه خود است.
* شناسایی نقطه Elbow:  
  نقطه Elbow جایی است که کاهش مقدار Inertia به‌طور محسوسی کاهش می‌یابد. این نقطه بهترین مقدار K را نشان می‌دهد، زیرا تعادل بین تعداد خوشه‌ها و کیفیت خوشه‌بندی را برقرار می‌کند.

که در پروژه نیز از همین روش برای پیدا کردن مقدار دقیق K استفاده شده است.

**سوال9. 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؟**

مقایسه:

* K-means: در داده‌های گرد و متراکم بهترین عملکرد را دارد.
* DBSCAN: برای داده‌های نامتقارن و نویز بهتر عمل می‌کند.
* Hierarchical Clustering: نمایش سلسله‌مراتبی ارائه می‌دهد اما برای داده‌های بزرگ ناکارآمد است.

**بهترین روش: K-means**  
**دلیل:**

1. **سادگی و سرعت:** K-means یک الگوریتم سریع و ساده است که برای داده‌های متراکم و خوشه‌های با توزیع کروی عملکرد بسیار خوبی دارد.
2. **نتایج واضح:** مراکز خوشه (Centroids) به وضوح نماینده داده‌های هر گروه هستند.
3. **قابلیت تکرار:** با تنظیم مقدار مناسب k با استفاده از روش( Elbow)، خروجی این روش پایدار و قابل‌اعتماد خواهد بود.
4. **کاربرد در داده‌های متراکم:** K-means در این داده‌ها بهترین عملکرد را نشان داده است و نسبت به دیگر روش‌ها مانند DBSCAN یا Hierarchical Clustering، مناسب‌تر بوده است.

**مقایسه با سایر روش‌ها:**

* **DBSCAN:** اگرچه برای داده‌های با نویز مناسب است، اما در این مجموعه داده نتوانست خوشه‌بندی دقیقی ارائه دهد.
* **Hierarchical Clustering:** به دلیل حجم داده و پیچیدگی خوشه‌ها، محاسبات بیشتری نیاز دارد و کارایی پایین‌تری دارد.

**نتیجه‌گیری:**  
K-means به دلیل عملکرد بهتر در این مجموعه داده خاص، انتخاب بهتری است.

**البته در کد عملکرد بهتر K-**means کاملا قابل مشاهده میباشد.

**سوال10. درباره PCA تحقیق کنید و نحوه عملکرد آن را به اختصار توضیح دهید.**

PCA (Principal Component Analysis) یکی از روش‌های کاهش ابعاد (Dimensionality Reduction) است که با حفظ بیشترین مقدار اطلاعات موجود در داده‌ها، تعداد ویژگی‌ها (Dimensions) را کاهش می‌دهد. این روش به‌خصوص برای نمایش و تجزیه‌وتحلیل داده‌های با ابعاد بالا مفید است.

نحوه عملکرد PCA:

1. مرکز داده‌ها: ابتدا داده‌ها استاندارد شده و میانگین هر ویژگی صفر می‌شود.
2. محاسبه ماتریس کواریانس: همبستگی میان ویژگی‌ها مشخص می‌شود.
3. محاسبه بردارهای ویژه (Eigenvectors) و مقادیر ویژه (Eigenvalues): برای ماتریس کواریانس محاسبه می‌شود تا جهت‌های اصلی داده‌ها تعیین شود.
4. انتخاب مؤلفه‌ها: مؤلفه‌هایی که بیشترین واریانس را پوشش می‌دهند، انتخاب می‌شوند.
5. ایجاد داده کاهش‌یافته: داده‌ها روی مؤلفه‌های اصلی پروجکت (Project) می‌شوند تا ابعاد کاهش یابد.

مزایا:

* کاهش ابعاد داده‌ها بدون از دست دادن اطلاعات مهم.
* حذف ویژگی‌های غیرضروری و کاهش پیچیدگی مدل.
* بهبود کارایی الگوریتم‌های یادگیری ماشین.

معایب:

* از دست رفتن تفسیرپذیری ویژگی‌های اصلی.
* حساس به مقیاس داده‌ها، نیاز به استانداردسازی قبل از اجرا.

**سوال 11. در مورد نحوه محاسبه معیار silhouette و homogeneity توضیح دهید.**

1. معیار Silhouette:
   * این معیار برای بررسی کیفیت خوشه‌بندی است.
   * هدف این است که ببینیم داده‌های یک خوشه چقدر به هم نزدیک هستند (انسجام) و چقدر از خوشه‌های دیگر جدا هستند (تفکیک).
   * اگر داده‌های یک خوشه به هم نزدیک و از بقیه خوشه‌ها دور باشند، Silhouette بالا است (عدد نزدیک به ۱).
   * اگر داده‌ها در مرز خوشه‌ها قرار گرفته باشند یا به اشتباه در خوشه نامربوط باشند، این عدد کاهش می‌یابد (نزدیک به صفر یا منفی می‌شود).
2. معیار Homogeneity:
   * این معیار برای بررسی همگنی خوشه‌ها است.
   * هدف این است که بررسی کنیم آیا تمام داده‌های یک خوشه خاص، به یک برچسب واقعی مشترک تعلق دارند یا خیر.
   * اگر تمام داده‌های یک خوشه به یک دسته‌بندی واقعی تعلق داشته باشند، مقدار Homogeneity برابر ۱ خواهد بود.
   * اگر خوشه‌ها پراکندگی زیادی داشته باشند و داده‌های با برچسب‌های مختلف را شامل شوند، مقدار Homogeneity به ۰ نزدیک می‌شود.

کاربرد:

* Silhouette زمانی استفاده می‌شود که داده‌ها برچسب واقعی ندارند (یعنی وقتی نمی‌دانیم داده‌ها به چه گروهی تعلق دارند).
* Homogeneity زمانی استفاده می‌شود که برچسب‌های واقعی داده‌ها در دسترس هستند و می‌خواهیم خوشه‌بندی را با برچسب‌ها مقایسه کنیم.

**سوال12. نتایج حاصل از معیارهای ذکر شده را برای هر یک از روش‌ها گزارش کنید.**

