هدف پروژه

هدف اصلی این پروژه، طبقه‌بندی محصولات یک فروشگاه آنلاین با استفاده از توضیحات مربوط به هرکدام به دسته‌های از پیش تعریف‌شده (شامل لوازم الکترونیکی، کتاب، پوشاک و لوازم خانگی) است. در این راستا، دو روش بسیار متداول و مهم برای نمایش و برداری‌سازی متون، یعنی TF-IDF و Word2Vec، پیاده‌سازی و عملکرد آن‌ها با یکدیگر مقایسه می‌شود.

دیتاست

برای این پروژه، از مجموعه داده E-commerce Classification Text استفاده می‌شود. این دیتاست شامل هزاران نمونه از توضیحات محصولات به همراه دسته‌بندی آن‌ها است که شامل موارد زیر است:

- Electronics
- Household
- Books
- Clothing

فاز اول: پیش‌پردازش داده‌ها

در این مرحله، هدف اصلی آماده سازی متون برای تحلیل و مدل سازی است. شما باید مطمئن شوید که داده‌ها عاری از نویز و به شکلی مناسب برای استفاده در مدل‌های پردازش زبان طبیعی (NLP) باشند. از تکنیک‌های مختلفی برای پیش‌پردازش داده‌ها استفاده کنید. انتخاب این روش‌ها بر عهده خودتان است.

فاز دوم: TF-IDF و Word2Vec

هدف این فاز این است که متون پیش‌پردازش شده در مرحله قبل را تبدیل به بردارهای عددی کنید. این کار به دو روش مختلف انجام می‌شود:

- **TF-IDF**
- **Word2Vec**

توجه داشته باشید که TF-IDF باید توسط خودتان پیاده سازی شود و اجازه استفاده از کتابخانه را ندارید.

فاز سوم: طبقه بندی و visualization

در این فاز، هدف اصلی، طبقه‌بندی بردارهای عددی است که در مرحله قبل ساخته‌اید، است. این کار به شما کمک می‌کند تا بفهمید کدام یک از روش‌های بردارسازی (TF-IDF یا Word2Vec) عملکرد بهتری داشته است. برای این کار دیتاست بدست آمده از مرحله پیش را به دو بخش train و test تقسیم کنید و طبقه بندی را انجام دهید. انتخاب الگوریتم طبقه بندی بر عهده خودتان است.

فاز چهارم: ارزیابی عملکرد

پس از آموزش، عملکرد هردو مدل را بر روی مجموعه داده تست بسنجید و باهم مقایسه کنید. نتایج با استفاده از معیارهای استاندارد ارزیابی (مانند accuracy، precision، recall و confusion matrix) ثبت و با یکدیگر مقایسه شوند. نتایج را در داکیومنت خود بررسی، مقایسه و تحلیل کنید. در نهایت، بعد از طبقه بندی به بررسی داده های قرار گرفته در کلاس های مشابه بپردازید. تعدادی از بردارها را ویژوالایز کنید و نتیجه را تحلیل کنید.

نکات تکمیلی

- علاوه بر سورس کد پروژه، فایل مستندات نیز باید آپلود شود.
- نام اعضای گروه در فایل مستندات ذکر شود و فقط یکی از اعضا پروژه را آپلود کند.
- هر گونه شباهت نامتعارف بین کد شما و کد سایر گروه ها تقلب محسوب مي شود و نمره ای برای این پروژه دریافت نخواهید کرد.
- در صورت نوشتن داکیومنت تمیز (برای مثال با LATEX) نمره اضافه برای شما در نظر گرفته خواهد شد.
- فایل شامل سورس کد پروژه و مستندات را در قالب فایل zip و با نام شماره دانشجویی خود ذخیره و ارسال نمایید.
- در صورت داشتن هرگونه سوال می توانید با fatemeh_dehbashii در ارتباط باشید یا در گروه درسی مطرح کنید.

موفق باشید؛
تیم حل تمرین