بسم الله الرحمن الرحیم

تمرین ۱: تحلیل احساسات پست‌های فضای مجازی در بازه ۵ ساله با استفاده از پردازش زبان طبیعی

درس: پردازش زبان طبیعی پیشرفته

شماره تمرین: ۱

تاریخ انتشار: ۱۴۰۴/۸/۱۳

مهلت تحویل: ۱۴۰۴/۹/۱۴

مقدمه

این تمرین برای آشنایی عملی با مراحل کلیدی پردازش زبان طبیعی(NLP) طراحی شده است. تمرکز بر تحلیل احساسات پست‌های فضای مجازی در بازه ۵ ساله است، که شامل جمع‌آوری داده از منابع آنلاین historical، پیش‌پردازش متن و استفاده از مدل‌های زبانی بزرگ می‌شود. این پروژه به شما کمک می‌کند تا مهارت‌های وب‌اسکریپینگ historical، مدیریت داده و تجسم را تقویت کنید و درک کنید چگونه احساسات عمومی نسبت به پست‌های گذشته می‌تواند بازتاب‌دهنده روندهای اجتماعی بلندمدت باشد، با تمرکز بر اینکه آیا مردم خوشحالند که این پست‌ها را می‌گذارند یا ناراحت.

هدف کلی: استخراج و تحلیل احساسات پست‌های فضای مجازی (مانند توییتر/X، اینستاگرام یا تلگرام) در بازه ۵ سال گذشته (از ۱۳۹۹ تا ۱۴۰۴ شمسی، معادل ۲۰۲۰ تا ۲۰۲۵ میلادی) و بررسی روندهای سالانه یا ماهانه آن‌ها.

نکته مهم: پروژه را به گونه‌ای طراحی کنید که تعمیم‌پذیر باشد (چه برای پلتفرم‌های اجتماعی دیگر و چه برای بازه زمانی historical دیگر). از هر زبان و کتابخانه ای که دوست دارید میتوانید استفاده کنید.

بخش ۱: منابع داده

- انتخاب منابع: ۵ پلتفرم یا حساب اجتماعی پرطرفدار در ایران را انتخاب کنید (مانند حساب‌های توییتر/X، کانال‌های تلگرام یا صفحات اینستاگرام). دلایل انتخاب را در گزارش توضیح دهید.

- نوع داده‌ها: برای هر پست گذشته، موارد زیر را استخراج کنید:

- تیتر یا متن پست

- تاریخ انتشار (historical، با timestamp دقیق)

- متن کامل پست

- دسته‌بندی پست (اختیاری: مانند سیاسی، اقتصادی، اجتماعی یا ...)

- واکنش‌های عمومی (اختیاری: کامنت‌ها، ریتوییت‌ها یا لایک‌ها برای تحلیل احساسات پست‌کننده و مخاطبان)

بخش ۲: جمع‌آوری داده‌ها

- از ابزارهایی مانند APIهای historical X/Twitter، BeautifulSoup/Scrapy برای scraping archives، یا APIهای رسمی (مانند Twitter API با since/until) استفاده کنید. برای historical، batch processing برای بازه ۵ ساله را پیاده‌سازی کنید.

- داده‌ها را در فایل CSV یا JSON ذخیره کنید (مثال ستون‌ها: post_text, timestamp, category, source, reactions – reactions شامل متن کامنت‌ها یا پست‌های مرتبط).

- مدیریت مسائل:

- داده‌های ناقص (مانند timestamp نامشخص) را علامت‌گذاری کنید (NaN در pandas).

- رعایت قوانین: robots.txt را چک کنید و از تأخیر (delay) در درخواست‌ها برای جلوگیری از بلاک استفاده کنید. برای historical، از pagination و rate limiting بهره ببرید.

- خروجی این بخش: فایل داده خام (raw_historical_posts.csv).

بخش ۳: پیش‌پردازش متن

۱. حذف تگ‌های HTML و نویسه‌های غیرضروری (لینک‌ها، ایموجی‌ها) با regex.

۲. حذف اعداد، نمادها و فاصله‌های اضافی.

۳. نرمال‌سازی فارسی

۴. متن‌ها را توکن‌سازی کنید تا آماده ورودی مدل شوند (هم متن پست و هم واکنش‌ها).

- خروجی: فایل پیش‌پردازش‌شده با ستون clean_post_text و clean_reactions.

بخش ۴: تحلیل احساسات

- لیبل‌زنی: با مدل زبانی بزرگ، هر پست و واکنش‌های مرتبط را در دسته‌های زیر طبقه‌بندی کنید (تمرکز روی احساسات پست‌کننده: آیا خوشحالند که پست می‌گذارند یا ناراحت):

لیبل ها به این صورت است

خوشحال، ناراحت، عصبانی، مظطرب، خنثی، نگران

بخش ۵: خروجی و تجسم

- نمودارها

- نمودار خطی یا میله‌ای: روند احساسات در بازه ۵ ساله بر اساس timestamp (سالانه یا ماهانه) و پست‌های گذشته.

- جدول آماری:

احساسات مردم در طول زمان(سالانه)

هر پلتفرم چقدر مردم را امیدوار و ناامید میکند (بر اساس پست‌ها و واکنش‌ها).

بخش ۶: نکات تکمیلی و ارزیابی

- تعمیم‌پذیری: کد را modular بنویسید (توابع جدا برای استخراج historical، پیش‌پردازش، تحلیل).

ارزیابی

٪۱۰ کیفیت گزارش کار.

٪۲۰ کیفیت و تمیزی کد.

٪۷۰ نتیجه

تحویل

- گزارش کار: PDF وWord با توضیح مراحل، نتایج و تحلیل.

- کد تمیز (با کامنت‌های انگلیسی).

- خروجی‌ها: به صورت json یا csv باشد.