به نام خدا

فاز اول پروژه درس پردازش زبان و گفتار

قسمت دیتا

دکتر بهروز مینایی

بهار ۱۴۰۱

تیم حل تمرین:

محمد یارمقدم

امیرحسین امینی مهر

ثمین حیدریان

مهسا انوریان

توحید عابدینی

رضا قهرمانی

تحویل مرحله اول پروژه در قالب یک ریپازیتوری گیتهاب میباشد. ساختار ریپازیتوری بصورت زیر میباشد. چنانچه ریپازیتوری خصوصی است به اساتید حل تمرین دسترسی بدهید.

- توضیح کلی پروژه در ReadMe.md
- یک فایل گزارش در ریشه ریپازیتوری تحت عنوان P1_Report.pdf
- پوشه src شامل کد های پروژه
- پوشه data شامل داده های پروژه

مجموعه داده

لازم است مجموعه داده به شکل خام و پیش پردازش شده موجود باشد.

- داده خام باید بشکلی ذخیره شده باشد که با اجرای مجدد دستور / کد / اسکریپت جمع آوری داده، داده خام بصورت خودکار بروز رسانی شود (مثلا در پوشه‌ای به نام raw)
- داده پیش پردازش شده لازم است بصورت مرحله-مرحله در پوشه/فایل های جداگانه ذخیره شود (مثلا در پوشه هایی با نام های cleaned، sentence_broken، word_broken)
- داده ها باید بشکلی ذخیره شده باشند که برچسب های براحتی در هر مرحله قابل تفکیک باشند.
- چنانچه منابع مختلفی برای جمع‌آوری داده استفاده شده، ساختار/ نام فایل / پوشه به گونه‌ای باشد که منبع جمع‌آوری داده در آن مشخص باشد.

کد جمع‌آوری و پردازش داده

کدهای لازم برای پروژه به سه منظور نوشته شده‌اند.

- جمع‌آوری / استخراج داده(کرال)
- پیش پردازش داده (شامل تمیز کردن داده، شکستن جملات، شکستن کلمات)
- استخراج آمار

لازم است کد جمع‌آوری داده و پردازش آن بصورت ماجولار نوشته شده باشد. بشکلی که از از خط فرمان بتوان مراحل مختلف پیش پردازش و دریافت داده را اجرا کرد. همچنین لازم است یک اسکریپت/نرم افزاری برای اجرای تمام مراحل جمع‌آوری داده، پیشپردازش و استخراج آمار داده داشته باشید بطوریکه محقق/ دانشجوی دیگر بتواند با اجرای این اسکریپت داد های مشابه پوشه data بدست آورد.

گزارش

در گزارش موارد زیر را قید کنید.

- منبع دقیق داده بطوریکه بازیابی آن با روش مشابه برای یک محقق دیگر قابل انجام و راس تآزمایی باشد.
- روش جمع‌آوری، مراحل و ابزارهای استفاده شده برای جمع‌آوری داده.
- فرمت داده ها (فایل و ساختار پوشه). ساختار هر فایل به چه صورت است و برچسب‌های مختلف چگونه از هم متمایز هستند.
- پیش پردازش‌های انجام شده
    - روش / ابزار تفکیک جملات
    - روش / ابزار تفکیک توکن‌ها / کلمات
    - روش/ معیارهای تمیز کردن داده
    - اندازه داده قبل / بعد تمیز کردن داده
- واحد برچسب‌گذاری (جمله، توییت، صفحه وب،)... و روش برچسب‌گذاری
- آمار داده به تفکیک برچسب در قالب جدول "و" نمودار
    - تعداد "واحد" داده
    - تعداد جملات
    - تعداد کلمات
    - تعداد کلمات منحصر به فرد
    - هیستوگرام تعداد تکرار هر کلمه منحصر به فرد به ترتیب از فرکانس بالا به پایین

نکته مهم:

در انتها **حتما** حداقل داکیومنت و کد جمع‌آوری و مرتب‌سازی داده‌های خود را در یک فایل زیپ قرار داده و در کوئرا آپلود نمایید.

موفق باشید ☺