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

- توضیح کلی پروژه در ReadMe.md
- یک فایل گزارش در ریشه ریپازیتوری تحت عنوان Phase1-Report.pdf
- پوشه src شامل کدهای پروژه
- پوشه data شامل داده‌های پروژه

۱ مجموعه داده

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

- داده خام باید بشکلی ذخیره شده باشد که با اجرای مجدد دستور/کد/اسکریپت جمع‌آوری داده، داده خام بصورت خودکار بروز رسانی شود (مثلا در پوشه‌ای به نام raw).
- داده پیش‌پردازش شده لازم است بصورت مرحله-مرحله در پوشه/فایل‌های جداگانه ذخیره شود (مثلا در پوشه‌هایی با نام‌های cleaned, sentence_broken, word_broken, …).
- داده‌ها باید بشکلی ذخیره شده باشند که برچسب‌های 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.

۲ کد جمع‌آوری و پردازش داده

کدهای لازم برای پروژه به سه منظور نوشته شده‌اند.

- جمع‌آوری/استخراج داده (کرال)
- پیش‌پردازش 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 پیش‌پردازش و دریافت داده را اجرا کرد. همچنین لازم است یک اسکریپت/نرم‌افزاری برای اجرای تمام مراحل جمع‌آوری داده، پیش‌پردازش و استخراج آمار داده داشته باشید بطوریکه محقق/دانشجوی دیگر بتواند با اجرای این اسکریپت داده‌ای مشابه پوشه data بدست آورد.

۳ گزارش

در گزارش موارد زیر را قید کنید.

- منبع دقیق داده بطوریکه بازیابی آن با روش مشابه برای یک محقق دیگر قابل انجام و راست‌آزمایی باشد.
- روش جمع‌آوری، مراحل و ابزارهای استفاده شده برای جمع‌آوری داده.
- فرمت داده‌ها (فایل و ساختار پوشه). ساختار هر فایل به چه صورت است و برچسب‌های مختلف چگونه از هم متمایز هستند.

تعریف پروژه

مرحله اول – جمع آوری داده

مهلت تحویل – ۱۲ اردیبهشت قبل از کلاس

- پیش‌پردازش‌های 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

أ. تعداد «واحد» داده

ب. تعداد جملات

ج. تعداد کلمات

د. تعداد کلمات منحصر به فرد

ه. تعداد کلمات منحصر به فرد مشترک و غیر مشترک بین برچسب‌ها

و. ۱۰ کلمه پرتکرار غیر مشترک هر برچسب

ز. ۱۰ کلمه مشترک برتر هر برچسب نسبت به برچسب‌های دیگر بر اساس معیار زیر.

$$\text{RelativeNormalizedFrequency}(w_i) = \frac{\frac{\text{count}(w_i)}{\sum_{w \in \text{Label1}} \text{count}(w)}}{\frac{\text{count}(w_i)}{\sum_{w \in \text{Label2}} \text{count}(w)}}$$

ح. ۱۰ کلمه برتر هر برچسب بر اساس TF-IDF(w_i)[1] (در اینجا یک داکیومنت برابر است با تمام داده‌های متناظر با یک برچسب)

ط. هیستوگرام تعداد تکرار هر کلمه منحصر به فرد به ترتیب از فرکانس بالا به پایین

[1] https://en.wikipedia.org/wiki/Tf-idf