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
- پوشه stats شامل اطلاعات محاسبه شده برای داده‌های پروژه.

۱ مجموعه داده

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

- لازم است داده خام در پوشه data/raw به تفکیک برچسب مطابق پروپوزال ذخیره شده باشد (۱۰ نمره)
  - داده خام باید بشکلی ذخیره شده باشد که با اجرای مجدد دستور/کد/اسکریپت جمع‌آوری داده، داده خام بصورت خودکار بروز رسانی شود. چنانچه به هر دلیل بخشی از جمع‌آوری داده بصورت خودکار قابل انجام نباشد، بخشی از نمره کسر خواهد شد. برای مواردی که بصورت خودکار انجام نمی‌شود، چنانچه توضیح کامل، دقیق و قابل تکرار اضافه نشود، نمره‌ای به این بخش تعلق نخواهد گرفت. (۱۰ نمره)
- لازم است داده تمیز شده در پوشه data/clean ذخیره شده باشد (۵ نمره)
  - داده تمیز شده باید با اجرای دستور/کد/اسکریپت از پوشه data/raw استخراج شده و مطابق گزارش تمیز شده و در پوشه data/clean ذخیره شود. در این مرحله نباید هیچ بخشی بصورت دستی انجام شود. برای کثیفی داده در این قسمت نمره کسر خواهد شد. (۱۰ نمره)
- پوشه‌های data/wordbroken و data/sentencebroken باید شامل داده‌ها به تفکیک واحد داده، برچسب، جملات و کلمات در این دو پوشه موجود باشند. (۱۰ نمره)
  - داده‌های این دو پوشه باید بصورت خودکار از پوشه data/clean با اجرای دستور/اسکریپت/کد ایجاد یا بروزرسانی شوند. (۵ نمره)
- در انتها پس از اتمام مراحل پیش‌پردازش، داده خود را در هاب huggingface کرده و در گیت‌هاب به آن ارجاع دهید. داده باید به گونه‌ای باشد که بتوان آن را بارگیری کرد و از آن به راحتی استفاده کرد. برای آگاهی از نوع داده‌ها و روش ذخیره آن‌ها می‌توانید به نمونه داده‌های آپلود شده در Huggingface مراجعه کنید (۱۰ نمره)

✓ 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 (۲۰ نمره)

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 (۱۰ نمره)

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.
- 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.

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

  - 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

لازم است موارد بالا بصورت جداگانه و در فایل‌های csv یا png در پوشه stats ذخیره شوند (۱۰ نمره)

لازم است گزارش شما به صورت خودکار در صورت تغییر نتایج بدست آمده، آن‌ها را به‌روز کند (۱۰ نمره). برای اینکار از LaTEX استفاده کنید. برای نتایجی که در جدول گزارش می‌شوند می‌توانید پس از محاسبه در فایل‌های CSV قرار داده و در فایل گزارش آنها را به شکل زیر در جدول جاسازی کنید.

```
\begin{align*}
        \csvautotabular{../reports/word2vec/similarity.csv}
\end{align*}
```

برای نمودارها نیز در فایل لاتک بصورت مستقیم می‌توانید فایل png را با دستور `includegraphics` جاسازی کنید.