به نام خدا

فاز اول پروژه درس پردازش زبان و گفتار

دکتر بهروز مینایی

بهار ۱۴۰۱

تیم حل تمرین:

محمد یارمقدم

امیرحسین امینی مهر

ثمین حیدریان

مهسا انوریان

توحید عابدینی

رضا قهرمانی

هدف پروژه این درس تحقیق و نوآوری در روش‌ها و ابزارهای پردازش زبان‌های طبیعی نیست. هدف تسلط شما روی تسک‌ها و استفاده مناسب از آن‌ها روی داده جدید می‌باشد. با توجه به داده جدید جمع‌آوری شده توسط شما، انتظار می‌رود که در نهایت هر دانشجو یک مقاله فنی در رابطه با نتایج بدست آمده در پروژه خود ارائه کند. چنانچه به تسک خاصی (مثلا چت بات یا ترجمه ماشینی) علاقه‌مند هستید، سعی کنید با راهنمایی استاد حل تمرین، داده خود را به گونه‌ای انتخاب کنید که علاوه بر تسک‌هایی که به تدریج اعلام می‌کنیم، مناسب تسک مورد علاقه شما نیز باشد. در بخش پایانی پروژه امکان انجام یک تسک انتخابی به جای تسک معرفی شده را خواهید داشت. برای آشنایی با تسک‌های مختلف می‌توانید nlpprogress.com را ببینید.

مرحله اول پروژه انتخاب موضوع و جمع‌آوری مجموعه‌داده مورد نظر می‌باشد. در این مرحله انتظار می‌رود شما یک مجموعه‌داده جدید در رابطه با موضوع مورده علاقه خود جمع‌آوری کنید. لازم است که جمع‌آوری داده به گونه‌ای باشد که مطابق با موضوع انتخابی باید حداقل ۷۵ هزار کلمه باشد. اگر موضوع شما در قالب موضوع های تحلیل مقایسه‌ای و طبق بندی است داده شما می‌بایست حداقل شامل دو دسته باشه و هر دسته شامل خداقل ۵۰ هزار کلمی باشد.

مجموعه‌داده می‌تواند به هر زبانی باشد اما لازم است شما به آن موضوع آشنا و علاقمند بوده به نحوی که قبل از انجام تسک‌ها در مورد نتیجه تسک حدس و نظر داشته باشید و چنانچه نتیجه تحلیل با حدس شما متفاوت بود، بتوانید دیباگ کنید و علت این تفاوت را تحلیل کنید. برای شروع پروژه، سه موضوع دلخواد خود را به ترتیب الویت‌تان در قالب یک فایل در کوئرا ارسال نمایید. برای انتخاب موضوع خود سعی کنید به سوال‌های زیر پاسخ دهید.

- موضوع داده چیست؟
- زبان داده چیست؟
- آیا داده به زبان محاور های است یا رسمی؟
- دسته بندی داده (حداقل ۲ دسته) به چه شکل است؟
- داده از چه جهت برای شما جذاب است؟
- آیا فرضیه‌ای در رابطه با داده دارید؟
- نحوه جمع‌آوری داده؟
- آیا برای جمع آوری داده از کتابخانه یا api خاصی استفاده خواهید کرد؟ آیا دسترسی به این کتابخانه/ api دارید؟
- چه محدودیتی برای جمع‌آوری داده دارید؟ پیش‌بینی می‌کنید چه حجم از داده بتوانید جمع‌آوری کنید؟

برای انتخاب موضوع پروژه می‌توانید از میان موضوعات ذیل نیز انتخاب نمایید:

۱. تشخیص آهنگ پاپ از سنتی

۲. تشخیص اخبار جمهوری‌خواه از دموکرات

۳. تشخیص ژانر فیلم از روی خلاصه فیلم

۴. تولید فیلمنامه

۵. تجزیه و تحلیل احساسات متون مربوط به بازارهای مالی (بیتکوین)

۶. تشخیص موضوع متون مربوط به بازارهای مالی (بیتکوین)

۷. تشخیص غلط‌های املایی موجود در متون توییتر

۸. استخراج رابطه در زبان فارسی با بررسی وابستگی جهانی seraji

۹. استخراج رابطه در زبان فارسی با بررسی وابستگی جهانی perDT

۱۰. تشخیص دیالوگ کاراکتر در یک فیلم یا سریال

۱۱. تشخیص نوع خبر از روی تیترهای اخبار

۱۲. تشخیص احساس موجود در متن یک توییت

۱۳. ایجاد یک سیستم پرسش و پاسخ در زمینه هوش مصنوعی

۱۴. خلاصه سازی متون علمی و اخبار

۱۵. تشخیص زبان متون

۱۶. مصحح غلط های املایی فارسی

۱۷. تشخیص سبک شعر

۱۸. تولید زیرنویس برای تصویر

۱۹. استخراج کلمات کلیدی

۲۰. تجزیه و تحلیل نظرات محصولات فروشگاه اینترنتی

۲۱. تشخیص شباهت بین متون

۲۲. استخراج نظرات نامناسب (toxic)