مستندی که در مقابل شما قرار دارد، گزارش نحوه انجام و ایده‌پردازی برای تمرین اول درس پردازش زبان‌های طبیعی می‌باشد که توسط محمدمهدی قیدی به شماره دانشجویی ۹۸۱۰۵۹۷۶ تهیه و ایجاد شده است.

من در حال حاضر در شرکت دیوار به عنوان مهندس نرم‌افزار مشغول هستم و برای این تمرین فکر کردم که چه چیز بهتر از بررسی داده‌های خام و موجود روی سایت! به همین خاطر با کمک گرفتن از دیدی که نسبت به کار کردن سایت و API های دیوار داشتم این تمرین را انجام دادم.

برای این تمرین، ما نیاز به جمع‌آوری و انجام داده‌کاوی روی یک مجموعه داده داشتیم. برای این منظور سایت دیوار را انتخاب کرده‌ایم. سایت دیوار مرجعی برای فروش کالاهای دسته دوم و نو در ایران می‌باشد که برای غالب مردم کشور، اولین مرجع مورد انتخاب برای مراجعه و آگهی‌کردن کالای خودشان است.

در سایت دیوار، مانند بسیاری از پلتفورم‌های معروف فروش در داخل و خارج از کشور، دسته‌بندی‌های مختلفی وجود دارد. به طور مثال املاک، خودرو، کالای دیجیتال و لوازم شخصی ۴ دسته‌بندی با بیشترین استفاده بر اساس مشاهده میدانی و داده‌های تحلیل شده می‌باشند.

در تمرین اول درس NLP، پس از اندکی گشت و گذار در سایت دیوار و یافتن Endpointها و APIهای مورد استفاده، ابتدا تلاش کردیم با استفاده از ابزار Scrapy در پایتون، داده‌های دسته‌بندی خودرو را Crawl کنیم. اما متاسفانه این امر به دلیل CSR Client Side Rendering بودن سایت دیوار به کمک ابزار های رایج(Scrapy) ممکن نبود و ابزار Scrapy توانایی استخراج داده از این سایت را نداشت. به همین دلیل با استفاده از کتابخانه requests در پایتون، یک اسکریپت به نام scraper.py تنها برای این سایت نوشته شد که داده‌های مورد نیاز ما را جمع آوری کرد. این سری داده‌ها در پوشه data در فایل زیپ ارسال شده به عنوان پاسخ تمرین در CW موجود می‌باشند. سعی شده مقدار کافی و مورد نیاز دیتا جمع‌آوری شده تا هم از نظر حد میزان دیتای مشخص شده در تمرین(۱ تا ۵ مگابایت) و هم از نظر کیفیت و فراوانی داده در انجام تسک به مشکل نخوریم.

در مورد داده‌های دسته‌بندی خودرو در دیوار میتوان به این مورد اشاره کرد که در این دسته‌بندی انواع فروشنده‌ها اعم از نمایشگاهی، شخصی و حتی نمایندگان کارخانه‌ها آگهی می‌گذارند. همچنین در توضیحات آگهی‌ها بسته به نوع فروشنده و آگهی، انواع ادبیات(عامیانه، کتابی و ...) دیده می‌شود و می‌توان روی این داده‌ها بررسی خوبی انجام داد که در این تمرین هدف را همین مهم قرار دادیم.

در واقع فرد آگهی‌دهنده تمامی مواردی که برای فروش لازم می‌داند که اشاره کند را در این بخش می‌آورد. اما همان طور که انتظار داریم این داده به هیچ وجه تمیز نیست و مشکلات فراوانی دارد. مانند گذاشته نشدن درست علائم نگارشی، رعایت نشدن موارد املایی مثل نیم‌فاصله و ... و مواردی مانند تبدیل عدد به حروف(مثلا پراید مدل ۹۳ -> پراید مدل نود و سه) که می‌توان در پایپ‌لاین بررسی کرد وجود دارند.
در این مرحله، با استفاده از قسمت‌هایی از پایپ‌لاین ابزار Hazm با استفاده از Normalizer و Lemmatizer و همچنین Word Tokenizer و Sentence Tokenizer به تمیزسازی داده و آنالیز موارد داده‌ای روی آن خواهیم پرداخت. همچنین StopWord ها را از داده‌ها حذف خواهیم کرد تا بتوانیم داده‌ها را قبل و بعد از Preprocess با یکدیگر مقایسه کنیم.
همچنین با توجه به کراول شده بودن دیتا از سایت دیوار، مواردی از پایپ‌لاین را خودمان نیاز به پیاده‌سازی داشتیم که آن‌ها را در نوت‌بوک و بسته به نیاز پیاده‌سازی و استفاده کرده‌ایم.

پس از آماده شدن داده، در شهرهای تهران، ارومیه، تبریز، ساری، اصفهان، سنندج، کرج و رشت بررسی می‌کنیم هر آگهی از چند جمله به طور میانگین تشکیل شده است. همچنین برای هر آگهی بررسی خواهیم کرد که تعداد کلمات مورد استفاده افراد در توضیحات آگهی چه عددی بوده است. در نهایت، فرکانس کلمات پر کاربرد آگهی‌های هر شهر را یافته و بررسی می‌کنیم که آیا غالبا مردم یک شهر هنگام آگهی کردن یک خودرو آیا از دسته کلمات متفاوتی استفاده میکنند یا اینکه این مورد در تمام قومیت‌ها و در سطح کشور یکسان است.

همچنین سایر توضیحات را سعی کردیم به صورت کامنت در کد و در cell های تکست داخل نوت‌بوک بیاوریم.

لازم به ذکر است حجم کل دیتای جمع آوری شده ۵.۴ مگابایت بود که اگر هم بخواهیم این حجم را کاهش دهیم یا کم کنیم صرفا می‌توان فایل مربوط به یکی از شهرها را حذف کرد تا به حجم زیر ۵ مگابایت برسیم.