به نام خدا

پروژه ترم اینجانب شامل ابهام‌زدایی معنایی از 3 کلمه شیر، کرم و شانه است.

نحوه کار به این صورت بوده که ابتدا فایلهای متنی پیکره متنی متعلق به هر یک از کلمات بر اساس تعداد جملات/پاراگراف‌ها که برچسب‌دهی شده بودند به 4 قسمت تقریبا مساوی تقسیم شده و طی 4 مرحله هربار یک قسمت از آنها مورد آزمون قرار می‌گیرد. فایلهای متنی پیکره که نسبت به فایلهای ارسال‌شده کمی تغییر کرده‌اند در فولدرهای shir_4 folds ، krm_4 folds و shane_4 folds قرارداده‌شده‌اند.

فیچرهای استخراج شده از پیکره آموزش شامل فیچرهای callocational که خود شامل باهم‌آیی‌های 2تایی و 3تایی کلمات (بایگرم و ترایگرم)، کلمات قبل و بعد (بایگرم‌های) تارگت-ورد می‌شود و فیچرهای bag of words که در 3 مرحله در پنجره های ±5تایی، ±5 تا ±10 اطراف تارگت-ورد و از ±10 تا ابتدا و انتهای جمله مورد سنجش قرار گرفته‌اند، می‌باشد. لیست این فیچرها در فولدر total extracted features list قرارداده‌شده و مقادیر از فایل features_list_2 به داخل کدبرنامه فراخوانی می‌شوند.

برای ابهام‌زدایی از هر کدام از کلمات 4 نوع کد برای آموزش و 4 کد برای آزمون نوشته‌شده است.

برای مثال :

برای ابهام‌زدایی کلمه کرم، کد wsd_krm_1_train.py که در فولدر krm_4 folds_codes قرارگرفته است با استفاده از اطلاعات استخراج‌شده از فایل متنی krm_1_train.txt که در فولدر krm_4 folds قراردارد نوشته‌شده و با اجرای آن یک فایل متنی به نام krm_1_train_sorted_decision_list.txt ایجادمی‌شود که لیست تصمیم مورد استفاده در برنامه تست به نام wsd_krm_1_test.py را در خود جای‌داده‌است.

لیست‌های تصمیم بر اساس مقادیر بیشینه Log Likelihood ratio محاسبه‌شده برای هریک از فیچرها مرتب‌سازی شده‌اند.

پس از اجرای هر 4 قسمت برنامه آموزش و ایجاد 4 لیست تصمیم، 4 برنامه تست متناظر با آن‌ها را اجرا نموده و مقدار Average_precision از میانگین 4 precision اجرا، به‌دست‌می‌آید.

نتـــایج حاصـــل از یـــک نمونـــه از اجـــرای 4 برنامـــه تســـت بـــرای هـــر کلمـــه (مـــثلا بـــا نـــام krm_4 tests_average_precision_exe بـــرای کلمـــه کـــرم) در فولـــدر krm_4 folds_codes قرارداده‌شده‌است.