گزارش فاز اول پروژه بازیابی پیشرفته اطلاعات- دکتر بیگی

شبنم قاسمی راد ۹۴۱۰۵۸۰۳

پرند علیزاده علمداری ۹۴۱۰۰۰۲۴

تمام قسمت‌ها به جز قسمت امتیازی مطابق مطالب درس پیاده سازی شده است. تقسیم کار در این پروژه به صورت مساوی انجام شده است. زبان برنامه نویسی پایتون است.

در فایل main.py عملیات خواسته شده در صورت پروژ هرکدام جداگانه پیاده سازی شده است. در ابتدای کار زبان مورد نظر برای index کردن پرسیده می‌شود و در یک حلقه دستور مورد نظر کاربر و سپس کوئری و یا شماره داکیومنت (در صورت نیاز) پرسیده می‌شود و عملیات مورد نظر انجام و نتایج نمایش داده می‌شوند.

فایل utils یک فایل کمکی برای پارس کردن فایل‌های csv و xml در شروع برنامه است. برای csv از کتابخانه‌ی csv پایتون و برای xml نیز از element tree استفاده شده است. در فایل pre_process عملیات یکسان سازی متن و پیش پردازش (قسمت ۱) انجام می‌شود و در فایل indexing همه عملیات مربوط به index کردن، تصحیح پرسمان و جستجو انجام می‌شود و در فایل index_compression هم توابع مربوط به فشرده سازی و معکوس آن قرار گرفته اند.

اعداد قرار داده شده برای تشخیص کلمات پرتکرار را با سعی و خطا و به صورت تجربی به دست آورده ایم.

خروجی ایندکس های فشرده شده برای هر یک از اسناد فارسی و انگلیسی و با استفاده از هر دو روش گاما کد و بایت مد به صورت جداگانه در فایل‌های مربوطه قرار داده شده‌اند اما به دلیل حجم بالا (حدود ۴۰ مگابایت برای فایل فارسی) از آپلود کردن آنها خودداری می‌کنیم. در هر صورت کد تولید کننده‌ی این فایل‌ها در فایل index_compression موجود است و در صورت نیاز قابل بازسازی می‌باشد.