

شرح پروژه اول:

در این پروژه، باید یک مدل بازیابی اطلاعات را به نام مدل بازیابی اطلاعات بولین (Boolean) را با برخی از مفروضات ساده، پیاده سازی کنید. بدین منظور، باید مجموعه‌ای از اسناد ساده (Corpus) مثلاً بیست سند در مورد موضوعات مختلف (هر یک در حد ۱ الی ۲ پاراگراف) را تهیه کنید و درخواست‌های ساده‌ی بولین کاربر را که می‌تواند فقط شامل سه کلمه‌ی T_1 ، T_2 و T_3 با عملگرهای بولین: AND، OR و NOT باشند را پیاده‌سازی کنید. به عنوان مثال، یک پرس و جوی کاربر ممکن است به شکل $(T_1 \text{ And } T_2 \text{ Or } T_3)$ باشد.

برنامه شما باید بتواند پرس و جوی بولین موقعیتی (Positional) را نیز با عملگرهای بولین: AND، OR و NOT انجام دهد. متن پرس و جو در این حالت ممکن است شامل کلمه‌ی "with" و یا شامل کلمه‌ی "Near 2" باشد.

- به عنوان مثال، " $T_1 \text{ with } T_2 \text{ with } T_3$ " یعنی اسنادی برگردانده شوند که سه کلمه‌ی نام برده شده به صورت پشت سر هم " $t_1 \ t_2 \ t_3$ " در متن آمده باشند.
- به عنوان مثالی دیگر، " $T_1 \text{ Near } 2 \ T_2 \text{ Near } 1 \ T_3$ " برای بازگرداندن تمام اسناد که شامل سه کلمه‌ی T_1 و T_2 باشند به طوری که فاصله‌ی کلمه‌ی T_1 و T_2 از هم در متن حداکثر ۲ باشد (یعنی صفر، یک یا دو موقعیت از هم فاصله داشته باشند) و همزمان، فاصله‌ی بین کلمات T_2 و T_3 از هم، حداکثر ۱ باشند.

برای پیاده‌سازی این پروژه نکات زیر را در نظر بگیرید:

- ارزیابی (فایل قضاوت، ارزیابی معیار دقت و بازخوانی و ...) نیاز نمی‌باشد.
- پیاده‌سازی تمام پروژه‌ها تحت پایتون، یک نمره‌ی اضافی خواهد داشت.
- پیاده‌سازی تمام پروژه‌ها با پایتون در محیط اسپارک، به عنوان مثال : (Eclipse for developing with Python and Spark در ویندوز) دو نمره‌ی اضافی خواهد داشت.