پروژه دوم :

قسمت اول (اجباری): با استفاده از مفهوم مدل فضای برداری (Vector Space Model)، از جمله ریشه‌یابی، وزن دهی TF-IDF وCosine Similarity، برنامه‌ای بنویسید که قادر به گرفتن پرس و جو (Query) به عنوان ورودی، بازیابی و رتبه‌بندی اسناد مربوطه باشد. حداقل تعداد سندها، ۲۰ سند می‌باشد و تعداد کلمات Query می‌تواند حداکثر ۴ کلمه باشد (یعنی ورودی query می‌تواند یک کلمه ای، دو کلمه‌ای، سه کلمه‌ای و چهار کلمه‌ای باشد).

قسمت دوم (اختیاری: یک نمره اضافی): بعد از انجام مرحله‌ی اول، مدل احتمالاتی (Probabilistic Model) را روی نتایج قسمت دوم با استفاده از فرمول وزن‌دهی مجدد (F4) اعمال کنید و مجدداً اسناد بازیابی شده را رتبه‌بندی کنید .

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. بدین منظور توصیه می‌شود از یکی از پنج محیط توسعه یکپارچه (IDE) پایتون برای یادگیری ماشین یعنی (Jupyter، PyCharm، SPYDER، Rodeo یا Geany) استفاده کنید.
- پیاده‌سازی تمام پروژه‌ها با پایتون در محیط BigDataی اسپارک، به عنوان مثال (Eclipse for developing with Python and Spark : در ویندوز یا لینوکس) دو نمرهی اضافی خواهد داشت.