تکلیف شماره 3: نمایه‌گذاری به کمک لوسین

1. ابتدا نسخه‌ی فعلی لوسین را دانلود کنید. برای اینکار می‌توانید از ادرس https://lucene.apache.org/ کمک بگیرید.
2. به کمک توضیحات موجود در ادرس https://lucene.apache.org/core/8_6_3/demo/overview-summary.html ، فرایند نصب را دنبال کرده و با نحوه‌ی استفاده از کتابخانه آشنا شوید.
3. نمایه‌سازی را برای داده‌های مرتبط با خودتان، انجام دهید. برای اینکار فرضیات زیر را درنظر بگیرید:

- هر گروه به ترتیب 103 داکیومنت از مجموعه داده انتخاب کند (گروه اول 103 داکیومنت اول، گروه دوم 103 داکیومنت دوم و...).
در مجموعه داده هر سند به صورت زیر نمایش داده میشود:

. I 1

. W

correlation between maternal and fetal plasma levels of glucose and free fatty acids. correlation coefficients have been determined between the levels of glucose and ffa in maternal and fetal plasma collected at delivery. significant correlations were obtained between the maternal and fetal glucose levels and the maternal and fetal ffa levels. from the size of the correlation coefficients and the slopes of regression lines it appears that the fetal plasma glucose level at delivery is very strongly dependent upon the maternal level whereas the fetal ffa level at delivery is only slightly dependent upon the maternal level.

در این نمایش، خط اول شماره سند و خطوط بعد از W. ، محتوی سند می‌باشد.

- قبل از نمایه‌گذاری پیش‌پردازش‌های مناسب را انجام دهید (مثلا توکن‌سازی، حذف stop word، lemmatization، stemming)
- سیستم نمایه‌گذاری شما باید لیست داکیومنت‌ها را به عنوان ورودی گرفته و لیستی از term ها با تعداد تکرار آن در هر سند را نمایش دهد. به عنوان مثال:

Term1: doc1(3), doc10(1) , ..., doc100(6)

4. یک داکیومنت ایجاد کرده، در آن بخش‌های مختلف کد را توضیح داده و در انتهای آن، خروجی نهایی را قرار دهید.