تمرین دوم: (تمرینها به صورت گروهی نوشته می شود ولی نمره دهی انفرادی است. هر فرد به صورت جداگانه تمرین را ارائه می دهد.)

پردازش متن برای بازیابی اطلاعات:

برای این تمرین شما میتوانید از پایه کد نویسی کنید و یا با جستجو در وب، کتابخانه های مناسب برای پردازش متن را پیدا کنید. در هر دو صورت هر تابع باید بتواند با ورودی دلخواه خروجی ذکر شده را تولید کند. چنانچه از کتابخانه ای استفاده می کنید باید حتما در گزارشی کوتاه (1 صفحه ای) در مورد این کتابخانه، آدرس آن و اینکه چگونه با آن کار کرده اید توضیح دهید.

کلاسی بنویسید که شامل توابع زیر باشد:

1. یک فایل متنی به عنوان ورودی بگیرد (آدرس فایل متنی در کامپیوتر) و لیست کلمات و تکرار هر کلمه در فایل را پیدا کند:
   1. خروجی: لیستی از سه گانه <fileID, token, tf>
2. یک لیست کلمات بگیرد و تعداد کل کلمات و تکرار هر کلمه را محاسبه کند
3. فایل متنی بگیرد و تمام 2-gram ها و 3-gram ها را و تکرار آنها را پیدا کند.