ソフトウェア演習Ⅲ〔課題 6:ソーティングと文書ランキング〕青野雅樹

以下の問題に対する解答を、Python 言語でのプログラムを作成し、実行結果 (kadai6.txt) をつけて Moodle にアップロードせよ。〆切りは 11 月 27 日（金）まで（定期テストの週の金曜日の夜まで）とする。

課題 1 で扱った 20Newsgroup から得られたすべての文書（873 ファイル）https://www.kde.cs.tut.ac.jp/~aono/data/20news.zip から抽出した小文字の英単語データをマージして、文書頻度が 3 以上 （3 ファイル以上に出現するもの）の単語 （合計 4921 単語）に TF-IDF (Term Frequency – Inverse Document Frequency)モデルで重み付けした文書単語データを NumPy 形式で事前に作成している。これを https://www.kde.cs.tut.ac.jp/~aono/data/meta20news/tfidf.npy においてある。また、4921 単語とその文書頻度をペアにして 1 行ごとアルファベット順に列挙したものを https://www.kde.cs.tut.ac.jp/~aono/data/meta20news/DFlist.txt においてある。873 個のファイルは https://www.kde.cs.tut.ac.jp/~aono/data/meta20news/filelist.txt にあり、それぞれの文書を表すタイトルとして、メールの Subject を抽出したデータを https://www.kde.cs.tut.ac.jp/~aono/data/meta20news/Subject.txt においてある。

(1) 上記のファイル群を各自の作業ディレクトリにダウンロードし、このうち tfidf.npy と Subject.txt の合計 2 ファイルをプログラムから読み込め。ファイル名はプログラムに埋め込んでよい。

(2) [0,872]の間の整数の乱数を（現在時刻を初期値とする乱数の初期化をした上で発生させよ。これを、文書 ID（(3)ではクエリ ID (Query ID)と呼ぶ）を表すとする。

(3) (2)で得られた文書 ID と、自分自身を含む 873 個の文書ベクトルの間でコサイン類似度を計算し、クエリに類似する上位 10 個の文書のタイトル（Subject.txt から得られた文字列）と類似度をプリントせよ。（この操作を少なくとも異なる 5 つの文書 ID で実行せよ）

【コメントとヒント】

ファイルの入力に関して、今回は、バイナリデータを含む２つのファイルを読み込みます。文書総数は、ファイル総数に等しく 873 です。単語自体は今回の課題では用いませんが、単語は事前にこちらが絞り込んだ 4921 個の単語を用い、各単語の文書頻度は上記の https://www.kde.cs.tut.ac.jp/~aono/data/meta20news/DFlist.txt においてあります（これは参考のためで、今回はプログラムから読込む必要はありません)。

バイナリファイルの tfidf.npy は、873 文書それぞれが、4921 次元のベクトルで表現された **Numpy フォーマット**のバイナリデータとなっています。実際は、`873(文書数＝D) x 4921`（単語数＝T）の重みが、`row-major` の順に並んでいます。

Python からは、import numpy as np のあと、array = np.load('tfidf.npy')とするだけで、データを読み込むことができます。

2つのN次元ベクトルのコサイン類似度は以下の式で計算されます。

$$similarity(x, y) = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} y_i^2\right)}}$$

ソーティングは、自前のものでも、リストのsort関数、組み込み関数のsorted関数等を使っても結構です。インデックスのソートも必要となると思います。

今回の課題では、特にクラスを作成する必要はありません。一方、コサイン類似度は関数にしておくことを勧めます。

実行例(ここでは3回のみ実行)は以下のようです。

$ ./kadai6

****
作者：青野雅樹, 01162069
日付:2020年11月16日17時1分36秒
内容：ソーティングと文書ランキング
****
Query ID = 309
1.000000 [38908] Subject: Re: I donwloaded a .bin file from a unix machine - now what?
0.971074 [38845] Subject: Re: I donwloaded a .bin file from a unix machine - now what?
0.220208 [10176] Subject: Re: ? Required File format of WORD for MS-WINDOW File (.Doc)
0.169034 [10844] Subject: Re: MAC DISKS IN WINDOWS?
0.157513 [10150] Subject: Print to file: how do I print the file later?
0.155026 [9992] Subject: re: BBBBIG problem with W4W print file. Help!!!!
0.147993 [9465] Subject: Wanted - dialog box to select file(s) for DOS apps
0.129405 [60142] Subject: Re: DOS 6.0

0.128829 [38881] Subject: Re: Need gif/iff file format
0.128647 [10815] Subject: Re: MAC DISKS IN WINDOWS?

Query ID = 280
1.000000 [38762] Subject: Re: 48-bit graphics...
0.443049 [38784] Subject: Re: 48-bit graphics...
0.248834 [38945] Subject: Re: XV problems
0.193037 [38948] Subject: Colour Transform for Red/Green Colour Blindness
0.181192 [37917] Subject: Re: 16 million vs 65 thousand colors
0.17846 [37950] Subject: Re: 16 million vs 65 thousand colors
0.16521 [38893] Subject: Re: WANTED: 24 bit viewer
0.160858 [38299] Subject: Need help with Mitsubishi P78U image printer
0.150774 [38710] Subject: Problems grabbing a block of a Starbase screen.
0.143919 [60711] Subject: Do the 2MB ATI Ultra Pro 16 and 24 bit Windows Drivers Work?

Query ID = 440
1.000000 [60248] Subject: Re: Help with 24bit mode for ATI
0.208604 [10136] Subject: Re: Windows NT und X-Windows?
0.183432 [60805] Subject: Re: Do the 2MB ATI Ultra Pro 16 and 24 bit Windows Drivers Work?
0.182058 [10006] Subject: Re: Actix video card drivers for windows
0.153982 [10095] Subject: Re: Actix video card drivers for windows
0.120984 [9698] Subject: Re: ATI ultra pro Drivers? [bad ATI ultra]
0.113719 [60711] Subject: Do the 2MB ATI Ultra Pro 16 and 24 bit Windows Drivers Work?
0.105244 [10141] Subject: Re: (Some info) The DOS/MSW meltdown is progressing nicely
0.100141 [9556] Subject: Re: Cirrus Logic 5426 Graph Card
0.0935943 [60831] Subject: Re: Can't set COM4