ソフトウェア演習Ⅲ〔課題 7:文書検索・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（kadai7.py）を作成し、実行結果（kadai7.txt）をつけて Moodle に ZIP をアップロードせよ。〆切りは 11 月 24 日（水）（祝日の翌日）までとする。

20Newsgroup は 20 種類（20 クラス）のニュースグループに関するメール記事データセットである。このデータセットから、クラスごとランダムに検索対象データ用に 100 記事、クエリデータ用に 20 記事を事前に抽出し、それぞれのテキストデータに BERT を適用し、特徴量ベクトル(768 次元)を作成した。検索対象用データは 20news2000.npy とし、クエリデータは 20news400query.npy として、NumPy ファイルで表現している。これらをローカルフォルダにダウンロードせよ。（前者は約 6MB, 後者は約 1.2MB である。）

抽出したデータには、それぞれメール記事の Subject がついており、これを文書タイトルとして、以下の検索結果の表示に用いる。検索対象用の Subject は Subject.txt で、クエリデータの Subject は QuerySubject.txt に格納している。これらもダウンロードせよ。

Subject に加えて、20 クラスのどのクラスに属するかを、検索対象用データは、ClassId.txt に、クエリデータは QueryClassId.txt に格納している。これらもダウンロードせよ。上述のデータはいずれも https://www.kde.cs.tut.ac.jp/~aono/data/20news/においてある。

以上の準備ができたら、以下の課題を行え。なお、ダウンロードしたデータをプログラムで読み込む際、ファイル名はプログラムに埋め込んでよいとする。

(1) [0,399]の間の整数の乱数を、現在時刻を初期値とする乱数で 5 種類発生させよ。以降、これを「クエリ文書 ID セット」（あるいは単にクエリ ID セット）と呼ぶこととする。

(2) (1)で得られたクエリ ID セットの個々のクエリ ID に対して、検索対象データセット（合計 2000 データ）と、文書ベクトル間のコサイン類似度を計算し、クエリ文書ベクトルの情報（クエリ文書 ID、クエリの Subject、クエリのクラス）とそれに類似する上位 10 個の文書のタイトル（Subject.txt から得られた文字列）、コサイン類似度、ならびに、個々の文書が属するクラス ID をプリントせよ。（この操作を異なる 5 つのクエリ ID で実行せよ。その際、少なくとも 2 つ以上の異なるクラスのクエリとすること。）

(3) (2)で行う 1 回ごとの検索で、上位 10 文書にクエリ文書のクラスと同じクラスが含まれている割合（パーセント）を上位 10 個の文書をプリントした末尾に表示せよ。（後述の例を参照）

【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 6 つのファイルを読み込みます。BERT で処理され 768 次元のベクトルで表現されたベクトルデータは NumPy 形式の

バイナリデータとなっています。実際は、検索対象用が 2000(文書数＝N) x 768（BERT の特徴量ベクトルの次元＝T）の重みが、row-major の順に並んでいます。つまり NumPy ファイルの中身は N x T の行列データです。クエリ用のベクトルデータも同様で、サイズが 400(文書数＝Q) x 768 となります。

2 つの多次元ベクトルのコサイン類似度は、ライブラリを使うこともできますが、資料 7 でも紹介していますので、プログラム内に関数として定義してください。すなわち、2 つの N 次元ベクトル $\mathbf{x} = (x_1, x_2, ..., x_N)$ と $\mathbf{y} = (y_1, y_2, ..., y_N)$ のコサイン類似度は以下の式で計算されます。

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

検索結果のソーティングは、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。

実行例(ここでは3回のみ実行)は以下のようです。（注：実行するたびに現在時刻を初期値とする整数乱数でクエリIDを決めるので、結果はプログラム内で5回発生させても、別々に実行し、結果を連結してkadai7.txtを作成しても結構です。その際、作者や日付の部分が5回登場しても構いません。

```
$ .python kadai7.py
**************************************
青野雅樹, 01162069
日付: 2021-11-16 07:32:46.978122
内容：文書ベクトルの類似度を用いた文書検索
**************************************
Query ID =    145    class = [ 19 ] Subject: Re: Christian Owned Organization list
----------------------------------------------------------------
Rank 1 ( 1947 ) 0.941 Subject: Re: Pgp, PEM, and RFC's (Was: Cryptography Patents)    class =
[ 11 ]
Rank 2 ( 1673 ) 0.94 Subject: Re: ADB woes    class = [ 4 ]
Rank 3 ( 204 ) 0.94 Subject: Re: Christian Owned Organization list    class = [ 19 ]
Rank 4 ( 1216 ) 0.939 Subject: Re: some thoughts.    class = [ 0 ]
Rank 5 ( 309 ) 0.937 Subject: Re: ATF BURNS DIVIDIAN RANCH! NO SURVIVORS!!!    class =
[ 16 ]
Rank 6 ( 1629 ) 0.937 Subject: Re: 68040 Specs.    class = [ 4 ]
```

Rank 7 (1377) 0.937 Subject: Re: Where can I buy a BIOS? class = [3]
Rank 8 (983) 0.936 Subject: Re: Please Gentlemen class = [8]
Rank 9 (252) 0.934 Subject: Re: Is it good that Jesus died? class = [19]
Rank 10 (393) 0.931 Subject: Re: Who's next? Mormons and Jews? class = [16]
同じクラスがランク10位までに見つかった割合 = 20.0 %
****

Query ID = 252 class = [18] Subject: Re: Why not concentrate on child molesters?

Rank 1 (1749) 0.944 Subject: Re: Wanted: Trombone for a beginner class = [6]
Rank 2 (1874) 0.939 Subject: Re: Divine providence vs. Murphy's Law class = [15]
Rank 3 (378) 0.936 Subject: Re: The Cold War: Who REALLY Won? class = [16]
Rank 4 (908) 0.935 Subject: Re: Shaft-drives and Wheelies class = [8]
Rank 5 (418) 0.932 Subject: Re: electronic parts in NYC? class = [12]
Rank 6 (1176) 0.93 Subject: Re: The Evidence class = [18]
Rank 7 (1193) 0.93 Subject: Re: Why not concentrate on child molesters? class = [18]
Rank 8 (425) 0.928 Subject: Re: Lead Acid batteries & Concrete? class = [12]
Rank 9 (1249) 0.928 Subject: Re: university violating separation of church/state? class = [0]
Rank 10 (284) 0.928 Subject: Evolution as Fact and Theory class = [19]
同じクラスがランク10位までに見つかった割合 = 20.0 %
****

Query ID = 371 class = [9] Subject: Re: Best Homeruns

Rank 1 (626) 0.945 Subject: Re: Where's Roger? class = [10]
Rank 2 (31) 0.934 Subject: Re: Early BBDDD Returns? class = [9]
Rank 3 (79) 0.933 Subject: Re: A true story - Way to go Omar class = [9]
Rank 4 (88) 0.932 Subject: Re: Players Rushed to Majors class = [9]
Rank 5 (93) 0.928 Subject: Re: Torre: The worst manager? class = [9]
Rank 6 (43) 0.928 Subject: Re: The Babe v. The Pride of the Yankees class = [9]
Rank 7 (928) 0.927 Subject: Re: Drinking and Riding (eww, gross) class = [8]
Rank 8 (54) 0.926 Subject: Re: Tickets etc.. class = [9]
Rank 9 (84) 0.925 Subject: Re: NL vs. AL? class = [9]
Rank 10 (91) 0.925 Subject: Re: quick way to tell if your local beat writer is dumb. class = [9]
同じクラスがランク10位までに見つかった割合 = 80.0 %
****