ソフトウェア演習Ⅲ〔課題１：文字列処理〕　　担当：青野雅樹

https://www.kde.cs.tut.ac.jp/~aono/data/20news/ 以下に 20 Newsgroups データの一部（元は http://qwone.com/~jason/20Newsgroups/）を置いている（すべてのニュースグループのファイル（内容はメール）は**数字のファイル**である）。事前準備として、上記のデータのうち、自分の学籍番号の末尾の１ケタとデータファイルの末尾の１ケタが一致するファイルを１つ選択し、そのファイルを自分の作業フォルダにコピーしておくこと。

以下の項目を満たす Python プログラムを作成せよ。期限は 10 月 13 日の夜までとする。

(1) 自分の名前と学籍番号、ならびに（課題）ファイル名・現在時刻を書き出せ。
(2) 上記の URL にあり、ファイルと自分の学籍番号の末尾が一致するファイルをプログラムから読込み、行単位で以下の処理をせよ。
　(ア) ファイル内の「英単語」（ここで「英単語」とは、アルファベット（大文字と小文字は両方ＯＫ）だけからなる文字列とする）を抽出し、Python の辞書形式のデータに加えよ。ここで辞書のキーは英単語とし、値は出現頻度とする。
　(イ) すでに辞書に登録されている「英単語」なら、出現数をカウントアップせよ。
(3) ファイルを最後（EOF）まで読み終わったら、出現した総単語数（出力の末尾）と、すべての「英単語」とその出現頻度をプリントせよ。
(4) プログラムの実行では、必ず、以下のように引数にニュースファイル名を与えること（プログラムにファイル名を埋め込まない）とする。入力ファイルが存在しない場合や、引数にファイル名がない場合は、警告を出して終了すること。

```
$ python  kadai1.py  [ファイル名] > kadai1-output.txt
```

＊kadai1.py, 処理したデータファイル、ならびに出力結果を zip(kadai1.zip)として Moodle にアップロードしてください。

コメントとヒント：

この課題での「英単語」では、特殊文字（例　クオート（’）やダッシュ（－）文字）は含みません。上記の定義から、英単語以外の文字は、単語間のセパレータとして扱うといいと思います。たとえば、What’s called “Mother’s day” is well-known.のような英語の文があった場合、これから抽出される「英単語」（文字列）は、What, s, called, Mother, s, day, is, well, known の 8 種類,9 単語（"s"という単語が 2 回現れていますので 8 種類です）となります。

単語とカウントの保持には、Python 特有の辞書(dictionary)を使うことが条件となっています。

辞書（変数）の初期値は dict()(または{})とし、新しい単語に出会うたびに辞書にエントリを追加するといいと思います。

第 1 回の資料の Python の辞書（https://www.kde.cs.tut.ac.jp/~aono/2020/P-1.html#dictionary）が参考になるかと思います。すなわち、辞書には新しい単語ごと x[word] = 1 のようにして追加（ここで、辞書の変数を x と仮定）し、最後に、x.keys()ですべてのキーが取り出せますので、for ループ等で、取り出したキーと x[key]で値をプリントすればいいかと思います。

「英単語」の抽出にあたり、正規表現を利用することを勧めます。正規表現は re という名前のパッケージでサポートされています。具体的には以下のように使用します。

```
import re
```

としておき、入力された英文のテキストに関して、ある行のデータが line という変数に入っていると仮定したとき

```
new_string = line.strip() # line 前後の余分な文字を除去
# [^a-zA-Z¥n]は a-z でも A-Z でもない任意の文字
new_string = re.sub('[^a-zA-Z¥n]', ' ',new_string)
new_string = re.sub('¥.', ' ', new_string) # ピリオドを半角スペースに
new_string = re.sub('[0-9]', ' ', new_string) # 数字を半角スペースに
words = new_string.split() # スペースで split し、単語リストを得る
```

のような処理で、英単語を取り出せるかと思います。他にも、もっと簡単な取り出し方があるかもしれません。

日付は、datetime パッケージを使うことでプリントできます。以下に日付のプリント例を紹介します。

```
import datetime
#現在の日付と日時
date = datetime.datetime.now()
print(date)
```

```
2020-10-01 12:26:53.617788
```

課題の条件(1) の実行例（出力の先頭あたりの例）は以下のようです。

```
*****************************************
青野雅樹, 011620
現在時刻: 2020-10-06 12:19:26.069516
課題 1:ニュースグループファイルから英単語の抽出
入力ファイル：9140
*****************************************
```

なお、出力結果の英単語がソーティングされている必要はありません。

出力の末尾に

総単語数 = xxx

は忘れないようにしてください。

また、kadai1.py の先頭にも、たとえば、

```
"""
        ファイル：kadai1.py
        作者：青野雅樹
        ID: 011620
        作成日付：20/10/06(火) 12:50:11
        内容：ニュースグループファイルから英単語の抽出
"""
```

のようなコメントを書くと同時に、プログラム中にも、適宜、何をしているかよくわかるようなコメントを必ず書いてください。