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/jp/ 以下に 10 種類の日本語のニュース記事ファイル（UTF-8 符号, ファイル名=utf8-jp-newsX.txt(X=0～9)）を置いている。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 X の部分が一致するファイルを選択し、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 18 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「カタカナ文字列」（ここで「カタカナ文字列」とは、カタカナまたは、「ニュース」のように全角の「ー」を含む文字列とする。ただし、「ー」はカタカナ文字列の先頭にこないと仮定してよいとする）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「カタカナ文字列」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「カタカナ文字列」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、すべての「カタカナ文字列」とその出現頻度、ならびに、出現した「カタカナ文字列」の種類総数を末尾にプリントせよ。
(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:

ここでの「カタカナ文字列」とは、全角カタカナが 1 文字以上続くものとします。ただし、「ー」だけ単独 1 文字のものはカタカナ文字列としません。**全角カタカナは'ァ'（小さい「ア」： 整数値 12449=16 進数 0x30A1）から'ヶ'（小さい「ケ」:整数値 12534=16 進数 0x30F6）までが UTF-8 符号で連続しており、その間にすべての全角カタカナは含まれます。これに「ー」（整数値 12540=16 進数 0x30FC）を加えてください。**たとえば、以下の URL

http://ash.jp/code/unitbl21.htm が参考になると思います。

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/2022/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。

「カタカナ文字列」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()
  new_string = re.sub('[a-zA-Z¥n]', ' ',new_string) # アルファベットは捨てる
  new_string = re.sub('[0-9]', ' ', new_string) # 数字もすてる
  # カタカナでもーでもない文字は捨てる（ここは自分で埋めてください）
  new_string = re.sub('¥.', ' ', new_string) # ピリオドはすてる
  words = new_string.split() # line.split()
```

のような処理で、「カタカナ文字列」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)
```

```
2022-10-01 12:26:53.617788
```

課題の条件(1) の仮想的な実行例（出力の先頭あたりの例）は以下のようです。

```
**********************************************
青野雅樹, 01162069
日付: 2022-10-04 22:48:31.261481
課題 1:日本語ニュースからカタカナ文字列の抽出
入力ニュースファイル: utf8-jp-newsX.txt (X:0-9)
**********************************************
ノーベル　2
トポロジカル　1
コンピューター　1
ガス　1
(中略)

カタカナ文字列総数 =  18
----------------------------------------------------------------
```

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

カタカナ文字列総数 = 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：22/10/05(水) 09:50:11
        バージョン：1.0
        内容：日本語ニュース記事ファイルからカタカナ文字列の抽出
"""
```

のような、そのプログラムの概要、作者、日付当がわかる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。