Homework #1 - N-gram：前後文的關係

Prof. Sheng Luen Chung

October, 2021

Due date: 10/26 上課前上傳至 Moodle 作業繳交區

1. 前後文的關係

句子中組成的字詞會有一定的結構性，最明顯的一個特徵是，同樣的一顆字通 常很有可能發生在其他幾個特定字的前面或後面，這種關係即所謂的情境或是 上下文 (context)。

掌握各個字辭在前後文中發生的頻率，有助於語音辨識的效果，其中 n-gram 指的意思是，在觀察到 $x_1x_2 \dots x_{N-1}$ 之後，接下來在「**第 N 顆字**」，會 看到 x_n 的機率為何？按造數學的定義即為：

$$p(x_n|x_1x_2 \dots x_{N-1})$$

然而，*N-gram* 的機率很難決定，它必須從海量的資料中，由各種組合字串 的出現頻率進行計算而得。但當我們手上的資料不夠海量而有限時，折衷的辦 法，就是假想資料量夠大，而從這有限資料集一是資料文檔中所觀察到出現頻 率，即**實證 (empirical) 機率**當作近似。

本作業，即以一個包含 500 句的中文文字稿：**病歷 500 句子.txt** 當作用來計 算的資料文檔。每個句子均以標點符號「。」結尾。請在程式中，逐項計算並 且印出以下各小題的答案：

1. 統計一共出現多少不同的字 (character)、每個字出現幾次。請選出出現 最多的 10 個字。有哪些字是只出現過一次。
2. 請計算當 N＝2, 3, 4 時的 N-gram，請印出機率最大的前十名：

$$p(x_n|x_1x_2 \dots x_{N-1})$$

3. 請列出當看到哪些字 x_1 時，可以 100% 斷定其下一個字 x_2 出現的字。
4. 所有的字一共有幾種不同的音 (phoneme)？如：{要、藥} 屬於同一個音 的同音字 (homophone)。哪一個「音」發生的次數最多？ 請參考：「Python 中拼音庫 PyPinyin 的用法！這個庫有點意思哈。」 參請考網址：https://kknews.cc/code/klmpzxr.html
5. 請列出連續兩個音都相同，但其實是對應不同的兩個字 x_1x_2，如：{中 止、終止、中指}。

2. 如何交結果？

請按：「HW01_學號_姓名」的方式命名，將你有含執行結果的 *.ipynb 檔，以及同樣命名方式的 *.pdf 檔案 (此為 .ipynb 檔案的文件副本)，兩個檔案一起在 10/26 中午 12:00 之前上傳至課程 Moodle。逾時遲交計分。
同時，請各自寫自己的作業，不可抄襲