# lab05 - Hidden Markov Model and DP

HMM 的訓練資料: corpus.txt
中文常用字與注音對照表: bpmf.txt

中文 VOC_SIZE = 10000

sample input:
ㄓㄏㄇㄍㄓㄈ
ㄗㄖㄩㄧㄔㄌ

sample output:
ㄓㄏㄇㄍㄓㄈ => 中華民國政府
ㄗㄖㄩㄧㄔㄌ => 自然語言處理

♪ Tips
1. 由於語料庫龐大 ，可能要花數分鐘建立LM，請使用pickle來將建立好的LM存檔與讀入
http://docs.python.org/library/pickle.html

Dumping multiple variables:
```
import pickle
```

寫進pickle檔:
```
export = open("EXPORTFILENAME", 'wb')
pickle.dump((uniDict, biDict, Smooth, Total), export)
export.close()
```

讀取pickle檔:
```
pkl_file = open("LMFILENAME", 'rb')
uniDict, biDict, Smooth, Total = pickle.load(pkl_file)
pkl_file.close()
```

2.中文讀檔
```
infile = open("INFILENAME", "rt")
```

3.中文 unigram與bigram

```
for line in infile:
    #unigram
    for unigram in line:
        ...

    #bigram
    for i in range(len(line)-1):
        bigram = line[i:i+2]
```

另外一種寫法
```
wordList = list(line.encode('utf8'))
```

4. obs, state, emission

obs = The string of phonic symbols
input 注音文 (e.g., ㄓㄉㄊㄐ)

states = Chinese character string

emission = 每個input注音對應到的所有可能中文字
e.g., ㄓ:值制針...
　　ㄉ:得定對黨...
　　ㄊ:統團推...
　　ㄐ:計薦...

start_prob = unigram prob of Chinese character
對第一個音所有可能的中文字算出unigram prob.
ex. P(值), P(制), P(針)

transition_prob = bigram prob of Chinese characters
請寫成函數，直接在viterbi裡call
e.g., P(幫|請) = P(請幫) / P(請)

emission_prob = P(ㄑ｜請)
=> 都是1 ("請"開頭音為ㄑ的機率)

P(請) = # unigram(請) / # of unigrams in corpus
P(請幫) = # bigram(請幫) / # of bigrams in corpus

5. bpmf.txt 建成 dictionary, key為開頭注音, value為該注音所有可能的字list

```
if symbol in phonicDict:
     phonicDict[symbol].append(word)
else:
     phonicDict[symbol] = [word]
```

BONUS:接受部分是中文字
bonus input:
ㄓㄕㄧㄅ書
值ㄉㄊㄐ
bonus output:
ㄓㄕㄧㄅ書 => 這是一本書
值ㄉㄊㄐ => 值得他就