中文輸入法
注音整句輸入

403410034 資工四 黃鈺程
403410064 資工四 陳煒杰

Outline

- 預處理
  - 資料清理與合併
- 模型
  - **Model 0**：查表
  - **Model 1**：暴搜
  - **Model 2**：最長路徑
  - **Model 3**：PyPinyin + Viterbi
- 感想

Preprocess

國語字典清理 (1/2)

- xls 格式
- 非常髒
  - 含有 gif
  - 全形半形不分 ()() , , 。 ;
  - (變)、(又音)、(讀音)
  - NaN
- 無法單純使用 re
- 手寫處理函式
- 缺字:沒有「我」、「帥」

```python
symbol_regex = re.compile(r'[，；·ˊ]', re.UNICODE)
change_regex = re.compile(r'[\( (][變又語讀](.*)', re.UNICODE)
bracket_regex = re.compile(r'[\( (](.*)[\)) ]', re.UNICODE)

def process_name(name):
    if '.gif' in name:
        return None
    name = name.replace(' ', '')
    name = re.sub(symbol_regex, '', name)
    name = re.sub(bracket_regex, '', name)
    return name

def process_pron(pron):
    pron = pron.replace('  ', ' ')
    pron = re.sub(change_regex, '', pron)
    pron = re.sub(bracket_regex, '', pron)
    res = []
    for token in pron.split(' '):
        if len(token) == 0 or token == '  ':
            continue
        if len(token) > 1 and token[-1] == 'ㄦ' and token[-2] != ' ':
            res.append(token[:-1])
            res.append('ㄦ')
        else:
            res.append(token)
    return res
```

處理注音的函式, 不斷地試誤

國語字典清理 (2/2)

- 許多破音字
  - 扒手、扒開
  - 只取第一個音
- 收錄的詞許多不在詞頻表中
  - 詞頻設為 0
- 合併後的詞的數量為 40 萬
- 轉成 dict 並存成 pkl
  - 查詢字詞注音只需 O(1)

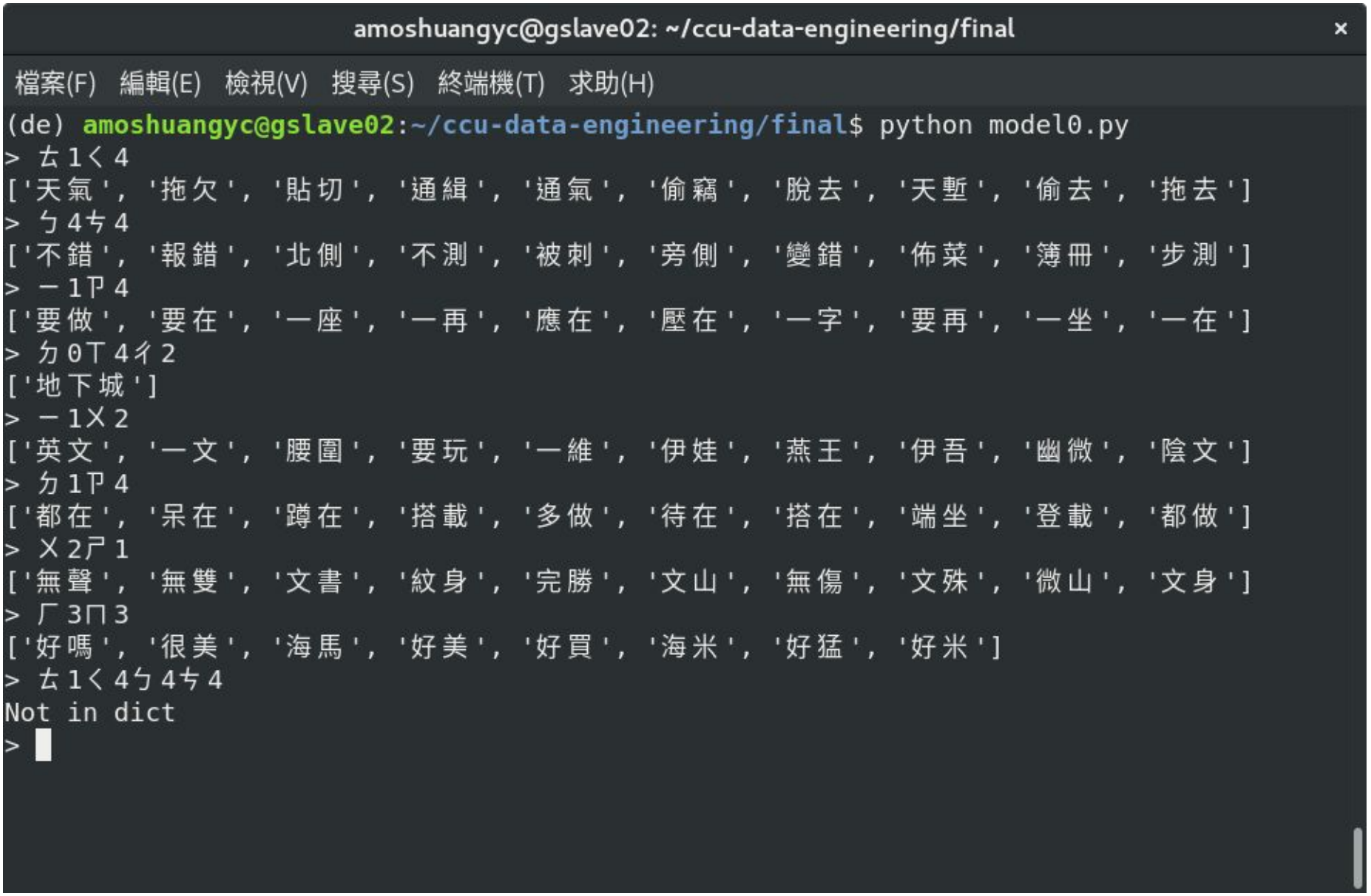
Models

測資

- 天氣不錯: ㄊ1ㄑ4ㄅ4ㄘ4
- 一座地下城: ㄧ1ㄗ4ㄉ0ㄒ4ㄔ2
- 英文單字: ㄧ1ㄨ2ㄉ1ㄗ4
- 吳昇好猛: ㄨ2ㄕ1ㄏ3ㄇ3

Model 0 : 查表

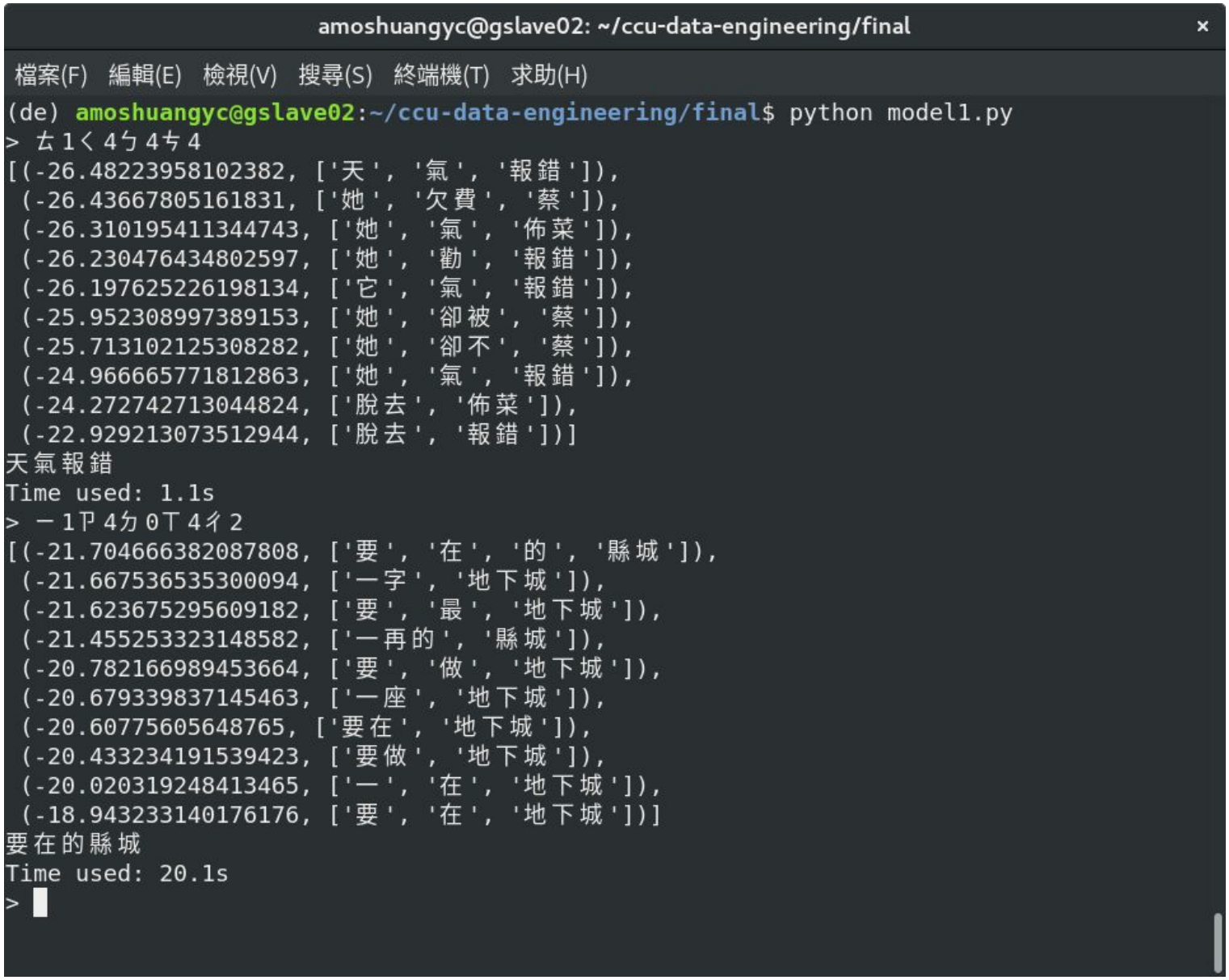
- 使用 dict 建表
  - 注音 -> list of 詞
  - dict[‘ㄊㄧㄢ4’] = ['天氣', '拖欠', '貼切', '通緝', '通氣', '偷竊', '脫去', '天塹', '偷去', '拖去', ...]
- 不處理整句輸入
- 輸出前 10 選項


amoshuangyc@gslave02: ~/ccu-data-engineering/final
檔案(F) 編輯(E) 檢視(V) 搜尋(S) 終端機(T) 求助(H)
(de) amoshuangyc@gslave02:~/ccu-data-engineering/final$ python model0.py
> ㄊ1ㄑ4
['天氣', '拖欠', '貼切', '通緝', '通氣', '偷竊', '脫去', '天塹', '偷去', '拖去']
> ㄅ4ㄎ4
['不錯', '報錯', '北側', '不測', '被刺', '旁側', '變錯', '佈菜', '簿冊', '步測']
> ㄧ1ㄕ4
['要做', '要在', '一座', '一再', '應在', '壓在', '一字', '要再', '一坐', '一在']
> ㄉ0ㄒ4ㄔ2
['地下城']
> ㄧ1ㄨ2
['英文', '一文', '腰圍', '要玩', '一維', '伊娃', '燕王', '伊吾', '幽微', '陰文']
> ㄉ1ㄕ4
['都在', '呆在', '蹲在', '搭載', '多做', '待在', '搭在', '端坐', '登載', '都做']
> ㄨ2ㄕ1
['無聲', '無雙', '文書', '紋身', '完勝', '文山', '無傷', '文殊', '微山', '文身']
> ㄏ3ㄇ3
['好嗎', '很美', '海馬', '好美', '好買', '海米', '好猛', '好米']
> ㄊ1ㄑ4ㄅ4ㄎ4
Not in dict
>

Model 0 結果

Model 1: 暴搜

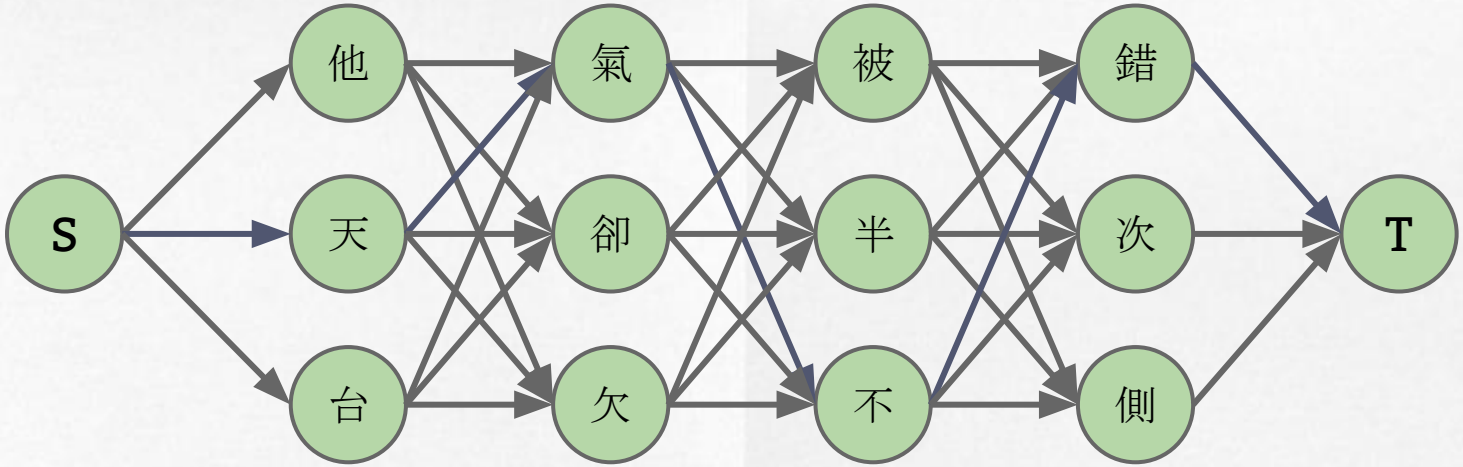
- 整句輸入
  - 不知斷點在哪 => DFS 暴開
- ㄊ1ㄑ4ㄅ4ㄘ4
  - ㄊ1 ㄑ4ㄅ4ㄘ4
  - ㄊ1ㄑ4 ㄅ4ㄘ4
  - ㄊ1ㄑ4ㄅ4 ㄘ4
- 所有組合中找權重和前 10 大
  - 邊權代表機率, 已取 log
  - 上下文無關


amoshuangyc@gslave02: ~/ccu-data-engineering/final
檔案(F) 編輯(E) 檢視(V) 搜尋(S) 終端機(T) 求助(H)
(de) amoshuangyc@gslave02:~/ccu-data-engineering/final$ python model1.py
> 士1く4ㄅ4ㄘ4
[(-26.48223958102382, ['天', '氣', '報錯']),
(-26.43667805161831, ['她', '欠費', '蔡']),
(-26.310195411344743, ['她', '氣', '佈菜']),
(-26.230476434802597, ['她', '勸', '報錯']),
(-26.197625226198134, ['它', '氣', '報錯']),
(-25.952308997389153, ['她', '卻被', '蔡']),
(-25.713102125308282, ['她', '卻不', '蔡']),
(-24.966665771812863, ['她', '氣', '報錯']),
(-24.272742713044824, ['脫去', '佈菜']),
(-22.929213073512944, ['脫去', '報錯'])]
天氣報錯
Time used: 1.1s
> ㄧ1ㄗ4ㄉ0ㄒ4ㄔ2
[(-21.704666382087808, ['要', '在', '的', '縣城']),
(-21.667536535300094, ['一字', '地下城']),
(-21.623675295609182, ['要', '最', '地下城']),
(-21.455253323148582, ['一再的', '縣城']),
(-20.782166989453664, ['要', '做', '地下城']),
(-20.679339837145463, ['一座', '地下城']),
(-20.60775605648765, ['要在', '地下城']),
(-20.433234191539423, ['要做', '地下城']),
(-20.020319248413465, ['一', '在', '地下城']),
(-18.943233140176176, ['要', '在', '地下城'])]
要在的縣城
Time used: 20.1s
>

Model 0 結果

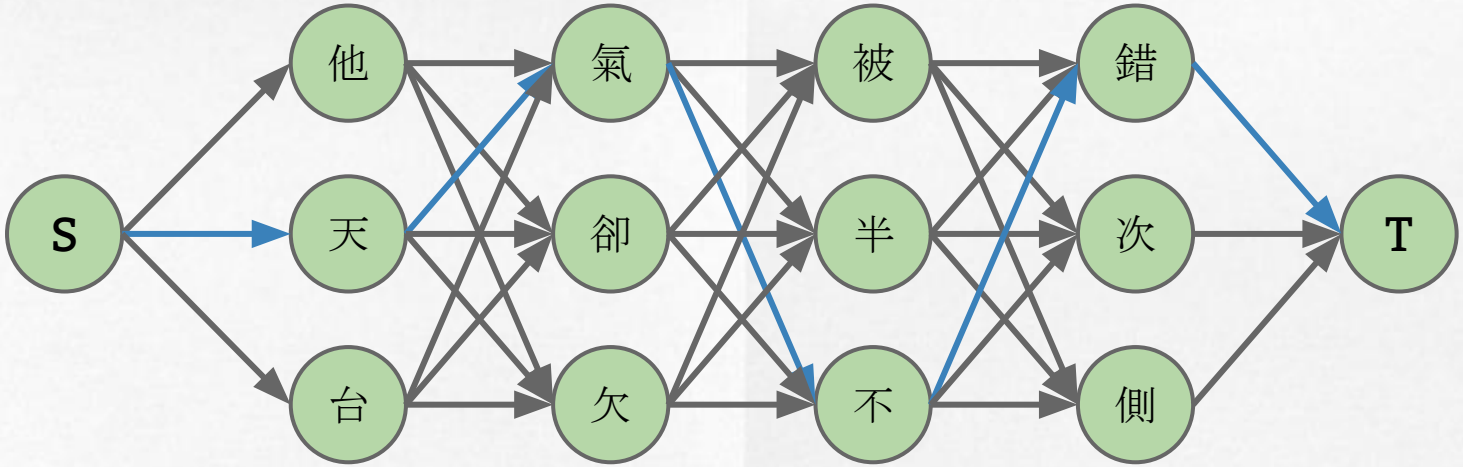
Model 2：最長路徑

- 上下文有關
  - bigram: P(w1|w2) = C(w1, w2) / C(w2)
- ㄊㄧㄢ ㄑㄧˋ ㄅㄧˋ ㄘㄜˋ



Model 2：最長路徑

- DAG
  - 沒有正環
  - Dijkstra




amoshuangyc@gslave02: ~/ccu-data-engineering/final
檔案(F) 編輯(E) 檢視(V) 搜尋(S) 終端機(T) 求助(H)
(de) amoshuangyc@gslave02:~/ccu-data-engineering/final$ python model2.py
> ㄊ1ㄑ4ㄅ4ㄣ4
['踢', '偷', '廳', '貼', '推', '添', '託', '他', '挑', '托', '帖', '她', '脫', '湯', '聽', '天', '通', '拖', '它', '台']
['勸', '確', '倩', '沁', '慶', '棄', '茜', '恰', '妾', '去', '竅', '企', '券', '氣', '趣', '欠', '卻', '汽', '器', '俏']
['不', '便', '抱', '報', '辦', '被', '半', '布', '變', '步', '必', '笨', '併', '部', '病', '北', '爸', '並', '佛', '費']
['挫', '肭', '湊', '脆', '菜', '廁', '刺', '策', '蹭', '側', '錯', '促', '測', '翠', '燦', '措', '簇', '寸', '蔡', '竄']
81
天氣不錯
Time used: 0.0s
> －1ㄗ4ㄉ0ㄒ4ㄔ2
['丫', '么', '腰', '應', '喲', '要', '雅', '一', '衣', '亞', '因', '耶', '煙', '伊', '燕', '依', '英', '音', '押', '壓']
['臟', '罪', '坐', '在', '造', '贊', '贈', '葬', '奏', '自', '藏', '醋', '揍', '醉', '最', '讚', '座', '做', '字', '再']
['底', '的', '地']
['線', '孫', '秀', '下', '縣', '姓', '現', '性', '係', '細', '象', '系', '樣', '信', '像', '項', '謝', '肖', '向', '笑']
['蟲', '晨', '盛', '愁', '茶', '塵', '程', '城', '成', '乘', '潮', '誠', '除', '常', '池', '沈', '船', '場', '呈', '持']
84
一座的下場
Time used: 0.0s
> －1ㄨ2ㄉ1ㄗ4
['丫', '么', '腰', '應', '喲', '要', '雅', '一', '衣', '亞', '因', '耶', '煙', '伊', '燕', '依', '英', '音', '押', '壓']
['吾', '雯', '文', '紋', '王', '玩', '唯', '薇', '聞', '丸', '亡', '娃', '完', '唔', '吳', '維', '魏', '微', '圍', '韋']
['多', '滴', '當', '東', '待', '端', '搭', '追', '燈', '丁', '登', '刀', '呆', '丹', '提', '都', '答', '單', '丟', '低']
['臟', '罪', '坐', '在', '造', '贊', '贈', '葬', '奏', '自', '藏', '醋', '揍', '醉', '最', '讚', '座', '做', '字', '再']
81
要聞都坐
Time used: 0.0s
> ㄨ2ㄕ1ㄏ3ㄇ3
['吾', '雯', '文', '紋', '王', '玩', '唯', '薇', '聞', '丸', '亡', '娃', '完', '唔', '吳', '維', '魏', '微', '圍', '韋']
['傷', '沙', '書', '詩', '勝', '山', '說', '收', '生', '燒', '刷', '聲', '雙', '輸', '刪', '商', '深', '殺', '昇', '師']
['吼', '悔', '浣', '夥', '罕', '很', '喊', '緩', '虫', '火', '好', '謊', '恍', '許', '狠', '伙', '海', '晃', '毀', '黑']
['馬', '瑪', '每', '姆', '閔', '敏', '美', '嗎', '滿', '買', '某', '免', '猛', '畝', '母', '鎂', '抹', '米', '碼', '秒']
81
文山海馬
Time used: 0.0s
>
>
>

Model 0 結果

Demo

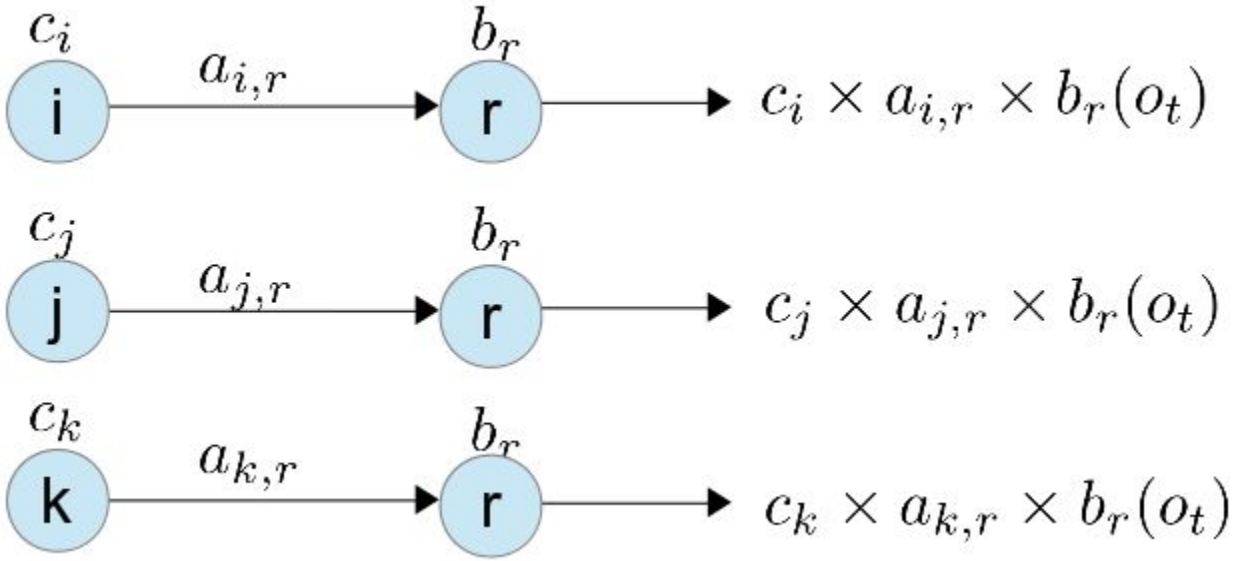
Model 3 : Pypingin + Viterbi

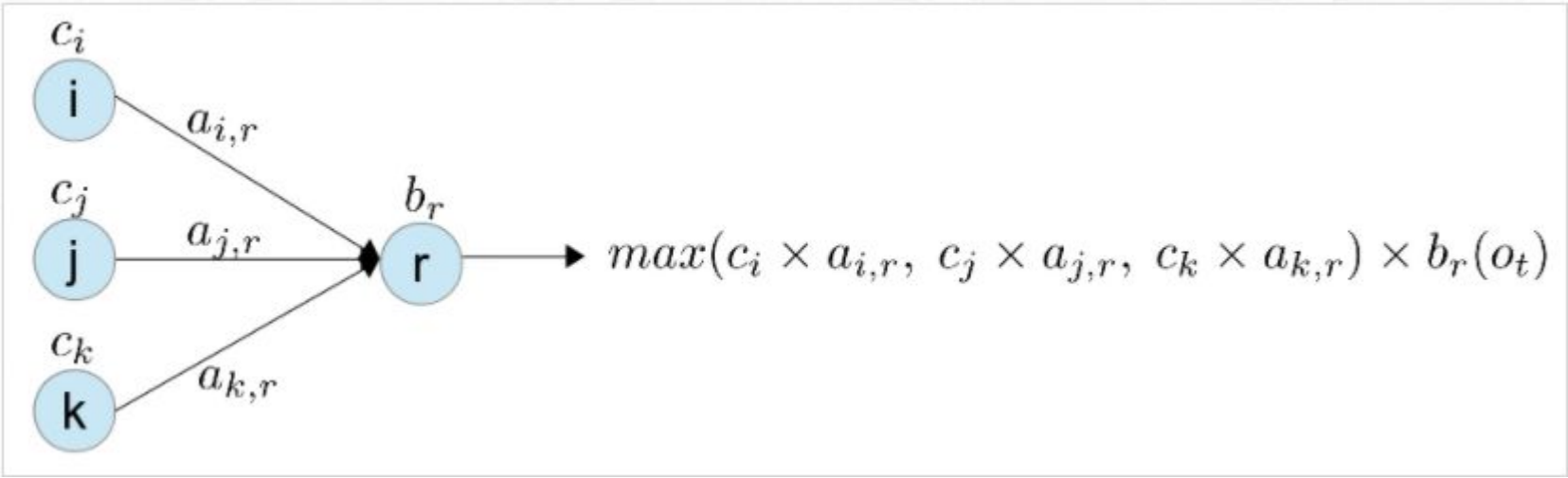
- 利用 **pypinyin** 的套件, 將 **Rime** 詞頻表中的詞, 轉換成注音

```
>>> from pypinyin import pinyin, lazy_pinyin, Style
>>> pinyin('中心', style=Style.BOPOMOFO)
[['ㄓㄨㄥ'], ['ㄒㄧㄣ']]
>>>
```

Viterbi

- 暴搜缺點太慢, 7 個字以上就沒跑出來
- 核心概念利用 **Dynamic Programming** 進行加速

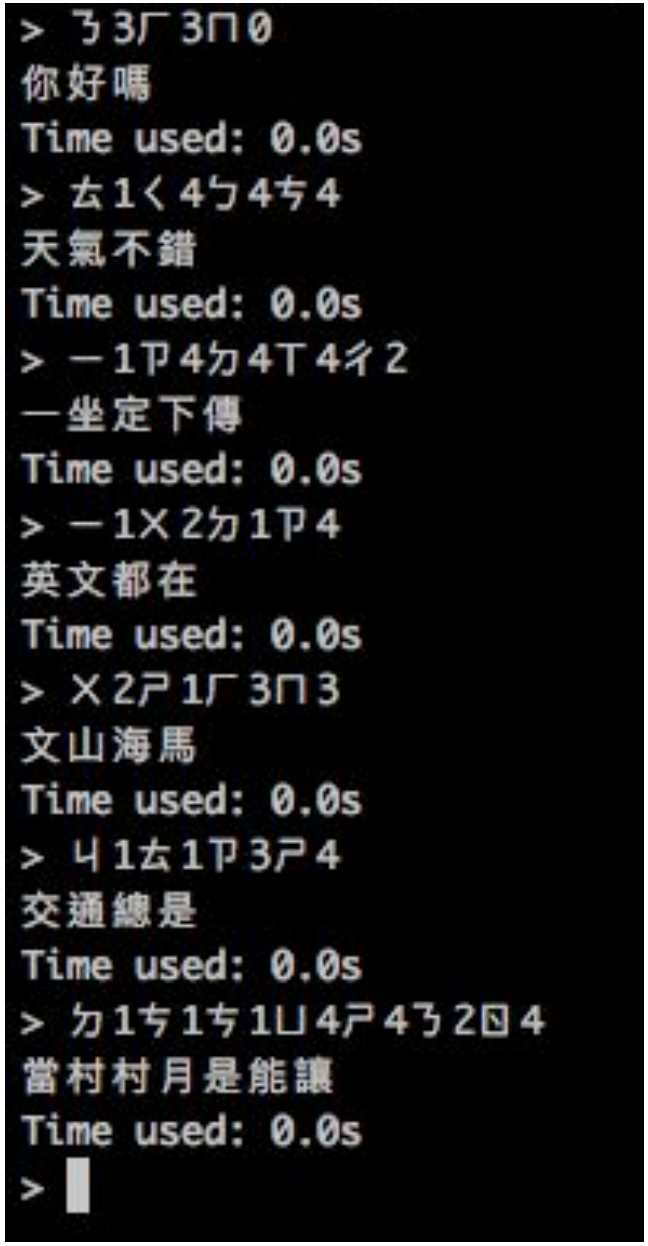

c_i
i
$a_{i,r}$
b_r
r
$c_i \times a_{i,r} \times b_r(o_t)$
c_j
j
$a_{j,r}$
b_r
r
$c_j \times a_{j,r} \times b_r(o_t)$
c_k
k
$a_{k,r}$
b_r
r
$c_k \times a_{k,r} \times b_r(o_t)$

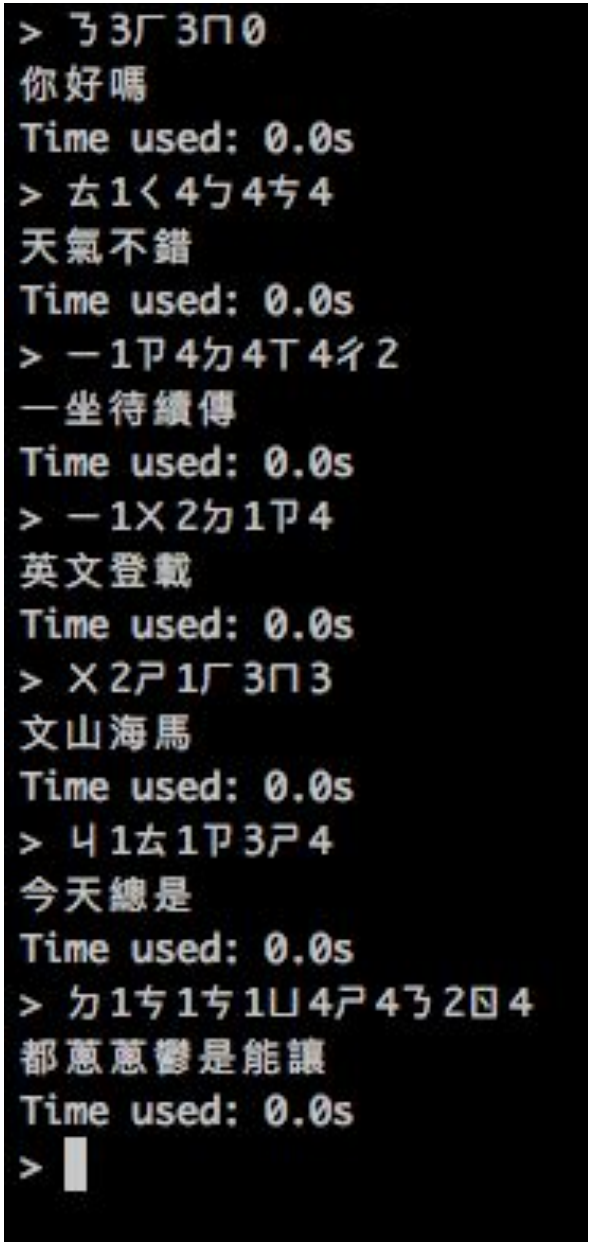

c_i
i
$a_{i,r}$
c_j
j
$a_{j,r}$
b_r
r
$max(c_i \times a_{i,r},\ c_j \times a_{j,r},\ c_k \times a_{k,r}) \times b_r(o_t)$
c_k
k
$a_{k,r}$

Viterbi 概念

測資

- 你好嗎：ㄋ3ㄏ3ㄇ0
- 天氣不錯：ㄊ1ㄑ4ㄅ4ㄘ4
- 一座地下城：ㄧ1ㄗ4ㄉ4ㄒ4ㄔ2
- 英文單字：ㄧ1ㄨ2ㄉ1ㄗ4
- 吳昇好猛：ㄨ2ㄕ1ㄏ3ㄇ3
- 今天早上：ㄐ1ㄊ1ㄗ3ㄕ4
- 當蔥蔥遇上牛肉：ㄉ1ㄘ1ㄘ1ㄩ4ㄕ4ㄋ2ㄖ4


> ㄋ3ㄏ3ㄇ0
你好嗎
Time used: 0.0s
> ㄊ1ㄑ4ㄅ4ㄘ4
天氣不錯
Time used: 0.0s
> ㄧ1ㄗ4ㄉ4ㄒ4ㄔ2
一坐定下傳
Time used: 0.0s
> ㄧ1ㄨ2ㄉ1ㄗ4
英文都在
Time used: 0.0s
> ㄨ2ㄕ1ㄏ3ㄇ3
文山海馬
Time used: 0.0s
> ㄐ1ㄊ1ㄗ3ㄕ4
交通總是
Time used: 0.0s
> ㄉ1ㄘ1ㄘ1ㄩ4ㄕ4ㄋ2ㄖ4
當村村月是能讓
Time used: 0.0s
>


> ㄋ3ㄏ3ㄇ0
你好嗎
Time used: 0.0s
> ㄊ1ㄑ4ㄅ4ㄘ4
天氣不錯
Time used: 0.0s
> ㄧ1ㄗ4ㄉ4ㄒ4ㄔ2
一坐待續傳
Time used: 0.0s
> ㄧ1ㄨ2ㄉ1ㄗ4
英文登載
Time used: 0.0s
> ㄨ2ㄕ1ㄏ3ㄇ3
文山海馬
Time used: 0.0s
> ㄐ1ㄊ1ㄗ3ㄕ4
今天總是
Time used: 0.0s
> ㄉ1ㄘ1ㄘ1ㄩ4ㄕ4ㄋ2ㄖ4
都蔥蔥鬱是能讓
Time used: 0.0s
>

Model 3 結果

Demo

感想

感想 (1/2)

- Pickle 很好用
  - 相當於 JAVA 中的 Serialization
  - 以二進位儲存 Python 變數
  - 讀取速度非常快
- readline
  - 讓 python 的 input() 變成像 terminal 一樣
  - 可以
    - Up arrow Key: 上次輸入的內容
    - Ctrl + U, Ctrl + L 都有效

感想 (2/2)

- 可惜做出來的 **Model** 都沒有很成功
  - 應該從自己處理 **ngram** 開始
  - 不要用別人的詞頻, 不知道別人是怎麼產生的
  - 教育部字典真的很爛唉
  - 想不到要再怎麼改進, 可以讓他變得更好

Q & A