中文的文字處理

授課老師：邱淑怡

DATE:2025/10/07

2 OUTLINE

- Pandas載入資料
- 文字清理
- 詞頻分析
- 視覺化

3 讀取檔案

- pd.read_csv(): 讀取csv檔
- pd.read_excel(): 讀取xlsx檔

4 只保留文字部分

冷冷的天來杯熱騰騰的豆漿配上燒餅油條或者來份蛋餅是不少人心目中的早餐完美組合但常常這樣吃要小心熱量掌控有營養師整理出常見的中式早餐熱量排行其中用白糯米包入肉鬆菜脯和油條的飯糰是中式早餐熱量王一個200公克的飯糰熱量高達600大卡至少要騎超過2個小時的單車才能消耗掉熱量 飯糰一個熱量約600大卡榮登中式早餐熱量王 營養師夏子雯表示飯糰本身是糯米飯量很多裡面包的餡料像是油條肉鬆煎蛋每一個無形當中都是熱量的疊加這樣吃完的話起碼會占到一天總熱量攝取的13 還有像是燒餅油條一份量高達554大卡蔥油餅蔥抓餅大肉包的熱量也有400大卡左右另外想要加點配料光是一條經過反覆油炸後的油條熱量就要將近300大卡想要吃的營養一點再加顆雞蛋熱量也有約100大卡 夏子雯說一顆雞蛋的熱量就是75大卡再加上油下去烹調雞蛋很吸油75再加上一份油脂45大卡的話可能有110大卡的熱量這是最基本的 營養師建議要吃中式早餐可以選純澱粉的饅頭或蘿蔔糕分成兩餐吃再搭配蛋白質食物至於飲品一定要挑無糖才不會吃完早餐就熱量爆表

```
text = re.sub(r"[^\u4e00-\u9fa5a-zA-Z0-9\s]", "", text)
```

中文　　英文與數字

5 JIEBA斷詞

- 支援三種分詞模式：
  - 精確模式，試圖將句子最精確地切開，適合文本分析(參數:cut_all=False)
  - 全模式，把句子中所有的可以成詞的詞語都掃描出來, 速度非常快，但是不能辨別詞義(參數:cut_all=True)
  - 搜尋引擎模式，在精確模式的基礎上，對長詞再次切分，提高召回率，適合用於搜尋引擎分詞
- 支援繁體分詞

前10個詞： ['冷冷的', '天來杯', '熱騰騰', '豆漿', '配上', '燒餅', '油條', '或者', '來份', '蛋餅']

6 COUNTER記數: 計算 TF(TERM FREQUENCY)

- Counter 屬於collection模組，是dict 的子類別，用來統計可雜湊物件的數量。這在需要計算元素出現次數時非常便利，例如統計字串中每個字母出現的頻率

TF-IDF關鍵詞選取

8 語法

- vectorizer = TfidfVectorizer(tokenizer=tokenize_and_clean):將文字轉為tf-idf特徵向量
- tfidf_matrix = vectorizer.fit_transform([docs[0]]):處理文字，依照tokenizer清除文字並且去除重複字詞，docs的資料型態必須是list of string
- feature_names = vectorizer.get_feature_names_out():轉換為文字

```
TF-IDF 關鍵字：
[['熱量', '早餐', '糰', '飯', '中式', '包入', '或者', '心目', '常常', '小心']]
```

9 詞性

- words = jieba.posseg.cut(news)

標籤	含意	標籤	含意	標籤	含意	標籤	含意
n	普通名词	f	方位名词	s	处所名词	t	时间
nr	人名	ns	地名	nt	机构名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	LOC	地名	ORG	机构名	TIME	时间

10

文字雲

語法

- from wordcloud import WordCloud:先下載wordcloud模組
- wc = WordCloud(

  font_path=font_path, # 指定中文字型路徑，Windows 可用 'msjh.ttc'

  background_color=background_color,

  max_words=max_words,

  width=800,

  height=400

).generate(text):建立文字雲物件，text必須為string

12 結果

- 字越大出現次數越高

13 連結

- https://github.com/okqji3ng2l/gdg
- https://colab.research.google.com/drive/1aMgLen3EhY8JH3SCIuyKOe_VXFspMwVE?usp=sharing