文本分析與程式設計

I have no idea

2020 Session Expired

2021.01.12

Target Problem and Solution

食材使用率

名字出現的頻率

- 是否包含特定品牌、商家
- 可能的或普遍的烹飪方式
- 可能的形容詞

- 潛在的熱潮迷思

常見
(觀眾?)

vs

常用
(發文數量?)

如果是
影響力大的人發文,
看到的人當然多;

但如果是著重在使用,
那麼應該改成"常用"

What and Why

社交媒體上
一定時間範圍內的常用食材

- 證實“常見” 又 “常用”
- 稍稍了解飲食趨勢或習慣
- 觀察可能的迷思或熱潮

可能的應用：
(1)觀察營養、飲食媒體
(2)飲食習慣研究

其它應用

⑴不同食譜類別：
甜點、健身、養身

⑵新食譜與食譜延伸：
菜單開發、計算食譜營養成分

⑶ 市場：
食材供應商、農業……

questions : 自己想到的問題

- **<u>要怎麼挑選分析對象?</u>**

1. 語言 : 繁體中文使用者, 不包含香港人
2. **可信度評估依據 : 追蹤數高的? 按讚數？**

- **<u>要怎麼有效率的看每一篇貼文?</u>**
- 要怎麼處理中英雙語食譜?
- 要怎麼總整處理結果?

questions　感謝大家的提問

(1)如果食材很少見但追蹤人數多而與事實不相符?

ans 如果是以貼文為單位? 也就是各篇食譜提到該項食材。

(2)食材是甚麼?像章魚小香腸與小香腸一不一樣?

章魚小香腸可能是兩種食材也可能是香腸造型,
如果是香腸, 那一樣會列入小香腸的計算。

(3) 基本調味料, 也會一併處理嗎?

會, 常用到的食材也會一併列入統計,
但這些必備品應該可以另外分類

(4)不會直接用文字標在圖片上的

12/22回應及發現更多的問題

1. **挑選複數位 ＋ hashtags**
   但是 Hashtags 不一定一致

2. **要怎麼驗收常見?**
如果用按讚數, 就會跟追蹤數有關

3. **如何明確對比**, 另一個介面 : 紙本？

4. **要找誰的貼文？** 因為難免會參考影響力。
如果直接用 ＃食譜#減肥食譜 就怕漏掉, 會漏掉多少?
Instagram influencer :
並沒有類別廚藝, 在生活知識之類的下面

5. **抓資料的具體方式**：**時間範圍、怎麼抓**

Target Problem and Solution

食材使用率

名字出現的頻率

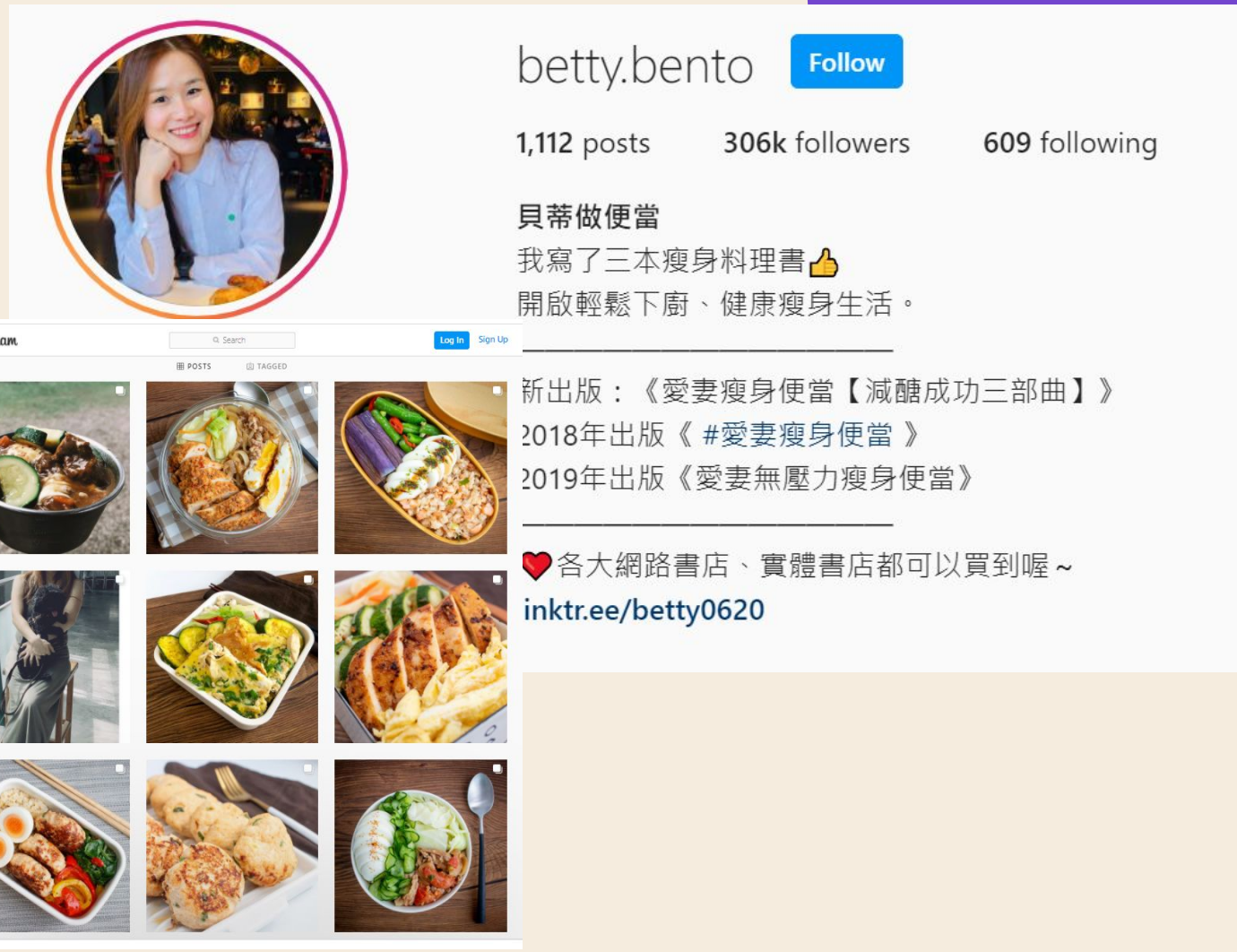
常用的食材

Data

Instagram：食譜部落客、
example： betty.bento

Why instagram： 年齡層、

其他可能性
Facebook、Youtube



原定Procedure

文本類別

- **分類文字/圖片/影像**
- Pytesseract
- 識別的正確率

文本分析

- 抓出每篇食材
- 統計結果

1

2

3

4

篩選目標

- 選帳號
- 標註# hashtags

前處理

- 每篇資料語言
- 斷詞

實際步驟

1229

new Procedure

文本類別

- 文字
  格式相似

文本分析

- 抓出每篇食材
- 統計結果

1 2 3 4

篩選目標

- 選帳號
- (非香港繁體字)
  標註# hashtags

前處理

- ~~每篇資料語言確認~~
- 擷取下來
- 抓食材

步驟	細項	注意事項
挑選帳號	1. 貼文呈現類型 : 文字 2. Hashtags 食譜 (熱門帳號)	@betty.bento 發文頻率
挑選範圍	1. 台灣 2. 以文字呈現	過濾香港人 過濾圖片及影片
挑選貼文	1. 2020 Dec. 正餐	內文格式 文字、圖片、影片

範例文本：emoji

1. 前言

常有網友問貝蒂，妳很愛櫛瓜哦！🤔

沒錯，我超愛櫛瓜料理的😂

只要在市場有看到櫛瓜的蹤影，我一定會順手買幾條回家備著；櫛瓜相較於其實葉菜體積大水份多，隨便炒一盤看起來的份量就很十足，更棒的是熱量低、富含營養……優點多多。

如果你在市場也有看到新鮮的櫛瓜，一定也要順手買幾條回家，一起煮起來👍

與大家分享簡易酸甜夠味的《茄汁櫛瓜炒肉》食譜

2. 食材

【材料，約3人份】
梅花豬肉片…300g
櫛瓜…2條(共350g)
熟白芝麻…適量

【醃料】
薑泥…1/2小匙
醬油…1大匙
料理米酒…2小匙
白胡椒粉…少許

【醬汁，調成一碗】
番茄醬…2大匙
飲用水2大匙
醬油1大匙
糖1/2小匙

[調味料]
油…少許

3.【作法】

①肉片加入醃料拌勻醃製約10分鐘;櫛瓜切成半圓型;醬汁預先調成一碗備用。
②鍋內倒入少許油，將醃妥的肉片入鍋以中火快炒，炒至半熟時起鍋備用。
③原鍋(免洗)，再倒入少許油，將櫛瓜入鍋煸炒至熟。
④作法2的半熟肉片倒回鍋中與櫛瓜一起拌炒。
⑤倒入醬汁，以中小火煨煮至略收汁後，撒入白芝麻，拌勻即完成。

【料理筆記】
✏️櫛瓜可依自己喜歡的熟度來決定拌炒的時間。

步驟	細項	注意事項
文本處理	抓名詞: Articut pos regular expression 計算次數: 詞頻分析	
結果應用	將食材數據抓出，統計結果	“常用”:統整頻率 基本調味料

範例:茄汁？炒肉？

常有網友問貝蒂，妳很愛櫛瓜哦！🤔

沒錯，我超愛櫛瓜料理的😂

只要在市場有看到櫛瓜的蹤影，我一定會順手買幾條回家備著；櫛瓜相較於其實葉菜體積大水份多，隨便炒一盤看起來的份量就很十足，更棒的是熱量低、富含營養……優點多多。

如果你在市場也有看到新鮮的櫛瓜，一定也要順手買幾條回家，一起煮起來👍

與大家分享簡易酸甜夠味的《茄汁櫛瓜炒肉》食譜

【材料，約3人份】
梅花豬肉片…300g
櫛瓜…2條(共350g)
熟白芝麻…適量

【醃料】
薑泥…1/2小匙
醬油…1大匙
料理米酒…2小匙
白胡椒粉…少許

【醬汁，調成一碗】
番茄醬…2大匙
飲用水2大匙
醬油1大匙
糖1/2小匙

［調味料］
油…少許

食物的狀態： 片、泥	食材本身差距： 番茄 vs 番茄醬
食材的特徵： 料理、飲用 白、	食材明確程度： 醬油、糖、油

【作法】
①肉片加入醃料拌勻醃製約10分鐘；櫛瓜切成半圓型；
醬汁預先調成一碗備用。
②鍋內倒入少許油，將醃妥的肉片入鍋以中火快炒，炒至半熟時起鍋備用。
③原鍋(免洗)，再倒入少許油，將櫛瓜入鍋煸炒至熟。
④作法2的半熟肉片倒回鍋中與櫛瓜一起拌炒。
⑤倒入醬汁，以中小火煨煮至略收汁後，撒入白芝麻，拌勻即完成。

【料理筆記】
✏️櫛瓜可依自己喜歡的熟度來決定拌炒的時間。

小結--處理要注意:

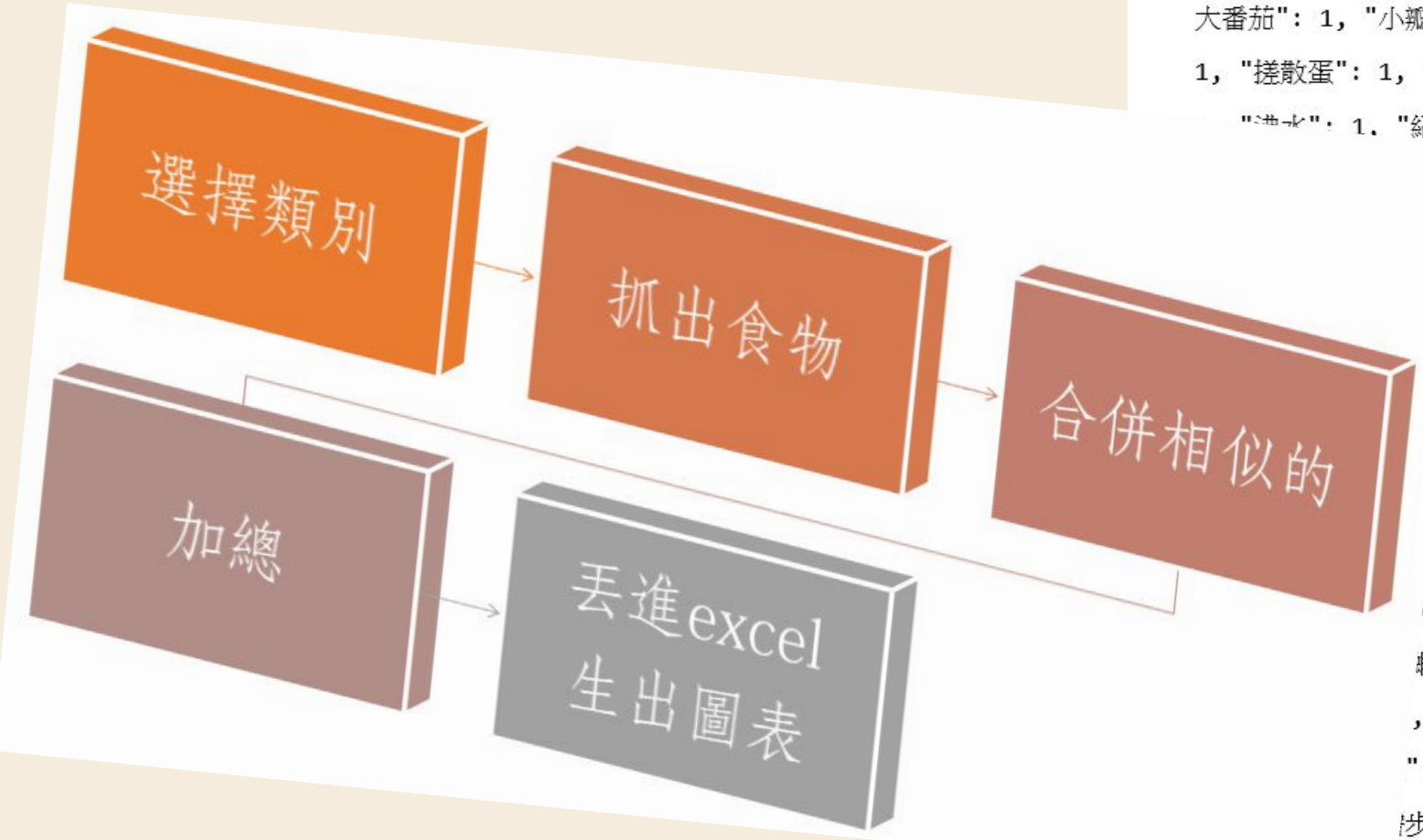
問題	細項	目前想法
食材名稱 與動作	炒肉	如果被斷開？ peter 12/29 給解
食材名稱 與狀態	肉片 薑泥 熟白芝麻	regular expression
食材名稱 與類似食材	油、醬油;	合併算總數

Code & Materials

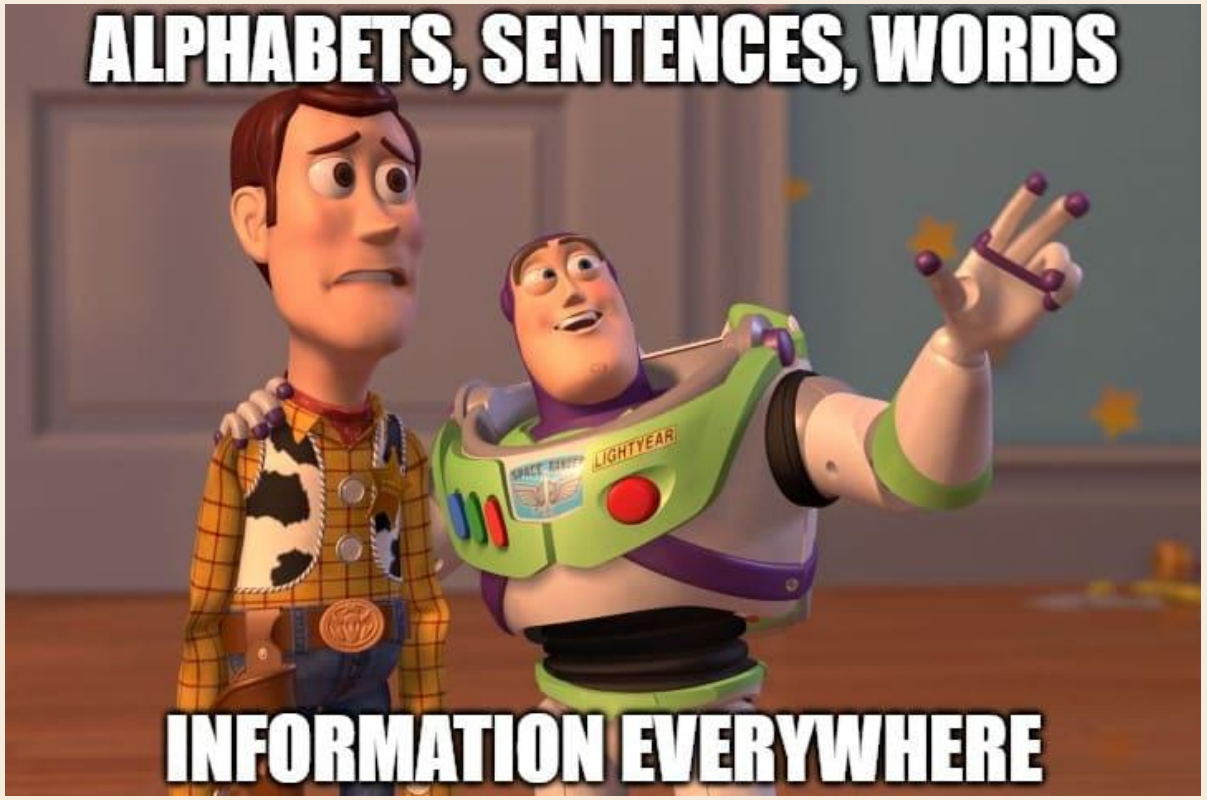
Github page:
https://github.com/KTJL/NTNUTextProcessing-I_have_no_idea

Result　json dict

工人智慧



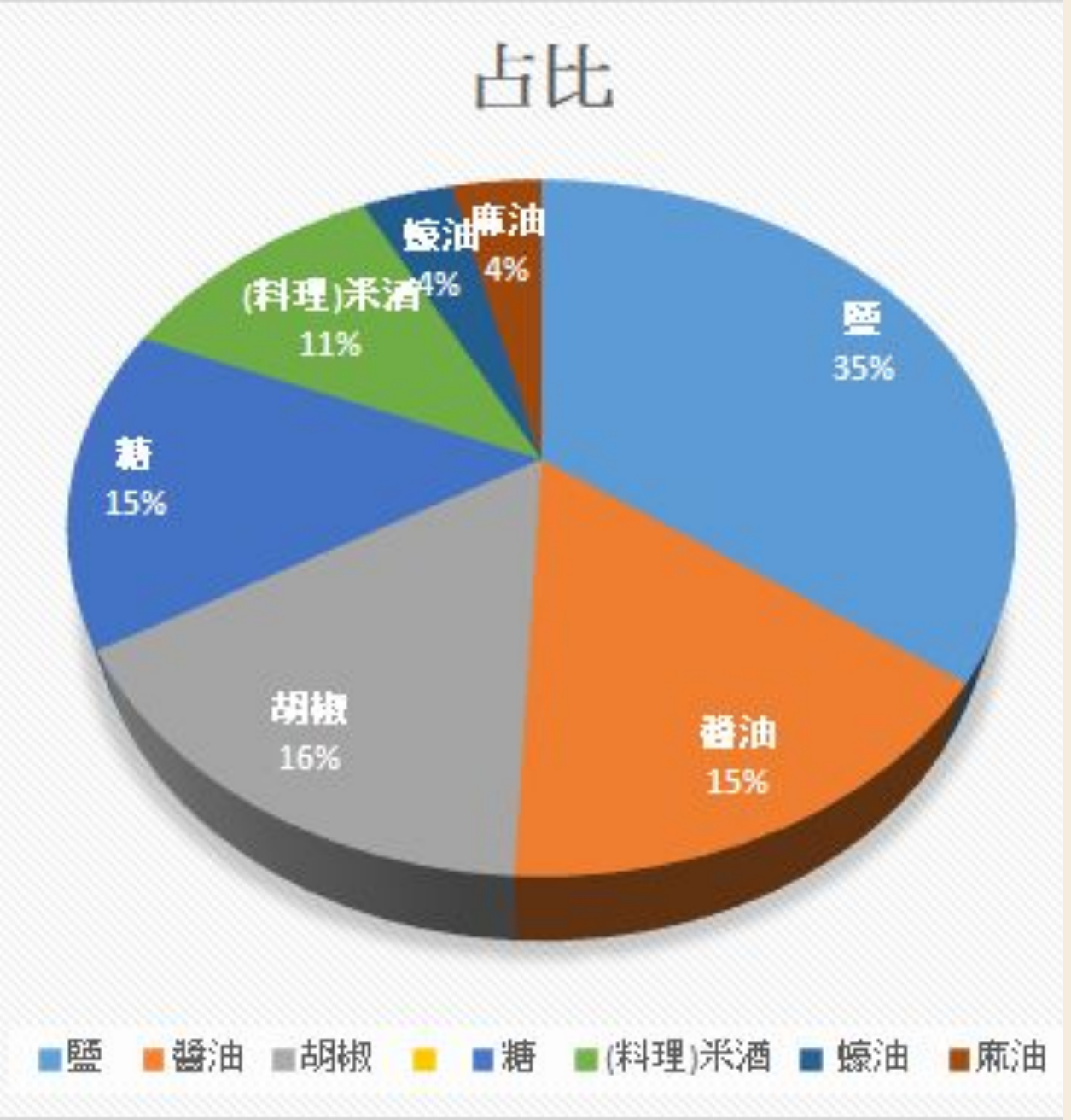
工麵": 1, "滑": 1, "吸麵": 1, "聲": 1, "哨子": 1, "🐐": 1, "小知識": 2, "屑
子": 1, "剁碎": 1, "草果": 1, "花椒…": 1, "成": 1, "🅱": 1, "好勁道麵條": 1,
"手工": 1, "製麵": 1, "好品質": 1, "原料": 1, "黃金": 1, "麥核心精華": 1, "
口口": 1, "杉木室": 1, "根根": 1, "影片": 3, "粉絲頁名稱": 2, "尹 yin": 2, "
大番茄": 1, "小瓣": 2, "個": 1, "豆瓣醬": 1, "3 香粉": 2, "小匙": 1, "菜鍋":
1, "搓散蛋": 1, "炒熟起鍋": 1, "肉鋪": 1, "洋蔥丁": 1, "番茄丁": 1, "扇形":
1, "紹子醬": 1, "🐿": 1, "條時間": 1, "冷水": 1, "🐿蛋液": 1, "🐿
1, "5 香粉": 1, "成番": 1, "茄罐": 1, "刀具": 1, "🍎蘋果
": 1, "奔牧場": 1, "running": 1, "成功率": 1, "蛋花湯😊":
秘訣": 1, "吐司": 1, "麵包": 1, "沙拉": 1, "蘋果": 1, "蜂
"蛋黃": 1, "體": 1, "首": 1, "儂": 1, "年紀😊": 1, "小簡
, "環境": 1, "可愛阿": 1, "阿": 1, "老板娘": 1, "缺": 1, "
1, "伙食": 1, "金盞花": 1, "葉黃": 1, "素": 1, "紅藻": 1,
": 1, "手機": 1, "阿🙈": 1, "🙈 ": 2, "SGS": 1, "生素":
藥": 1, "芬": 1, "普尼": 1, "沙": 1, "桿菌": 1, "檢出": 1,
檢": 1, "🐭": 1, "山林": 1, "老闆娘": 1, "🐮": 1, "優質阿":
皮": 1, "🔍草": 1, "⬛草": 1, "闆娘": 1, "？": 1, "層次分明
, "特": 1, "？阿": 1, "常溫": 1, "鮮度": 1, "陽光": 1, "蛋殼
": 1, "裂": 1, "㊙": 1, "大氣": 1, "飯殺手": 1, "蘋果蜂蜜咖哩
步驟": 1, "點": 1, "➡": 1, "妃蘋": 1, "果 ": 1, "ORO": 1, "



鹽巴": 19, "午餐": 33, "晚餐": 24, "夕食": 21, "早餐": 22, "食 ": 21, "クッキング ラム": 21, "チャーハン": 1, "卵チャーハン": 1, "friedrice": 1, "yahoo": 19, "級廚師": 18, "醬油蛋": 1, "中華料理": 1, "文鱼": 1, "饭 ": 1,

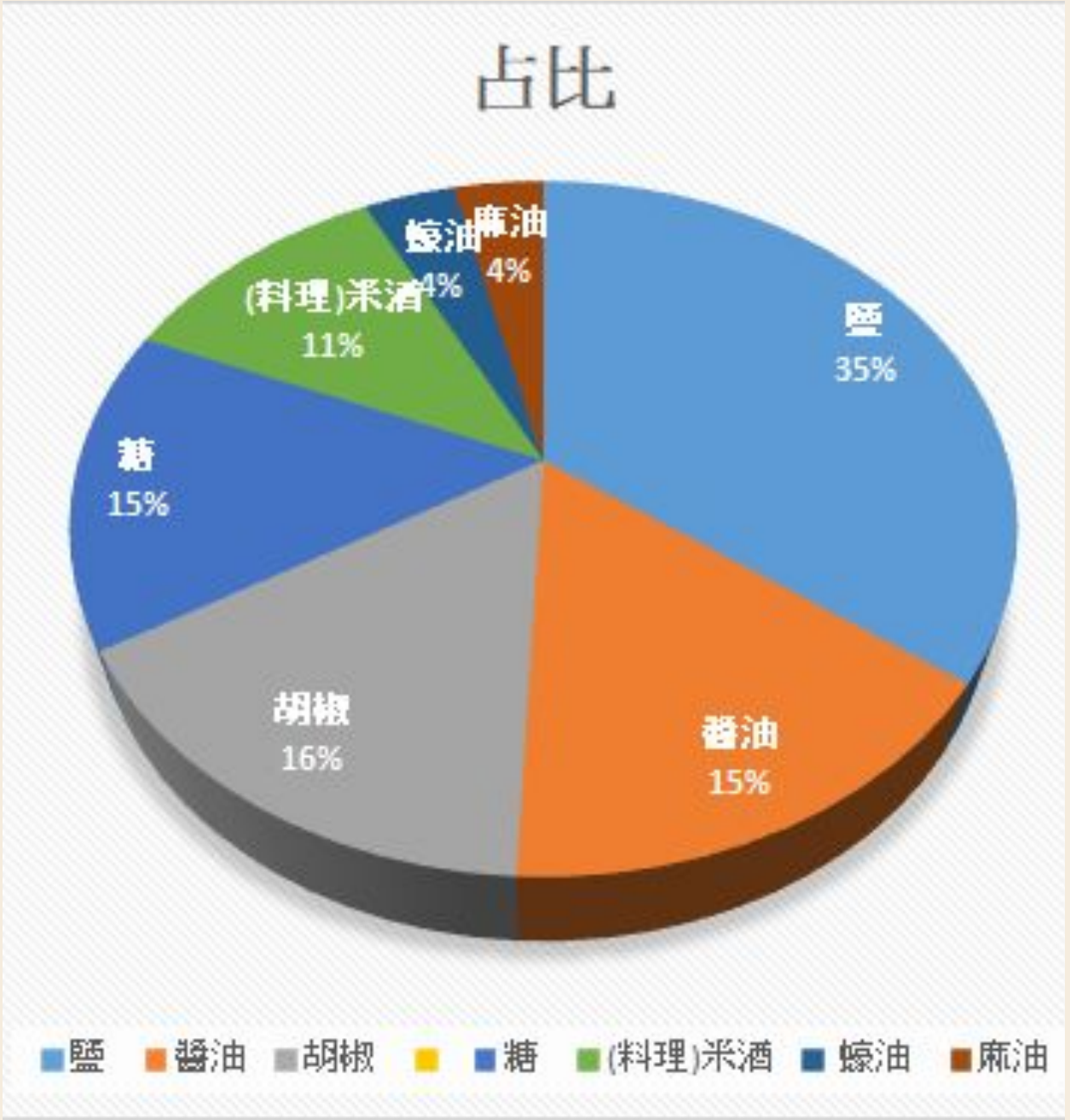
Result　調味料

1. 鹽 (94%)
   -- 鹽(巴)、海鹽(30%)
2. 醬油 (42%)
   胡椒 (42%)
   -- 胡椒、胡椒鹽(6%)
   、胡椒粉(12%)、
   胡椒粒(4%)
3. 糖 (40%)
   -- 糖、紅糖(14%)、黑糖(2%)
4. (料理)米酒 (30%)
5. 蠔油 (10%)
   麻油 (10%)



Result　調味料

1. 鹽 (94%)
   -- 鹽(巴)、**海鹽(30%)**
2. 醬油 (42%)
   胡椒 (42%)
   -- 胡椒、胡椒鹽(6%)
   、**胡椒粉(12%)**、
   胡椒粒(4%)
3. 糖 (40%)
   -- 糖、**紅糖(14%)**、黑糖(2%)
4. (料理)米酒 (30%)
5. 蠔油 (10%)
   麻油 (10%)

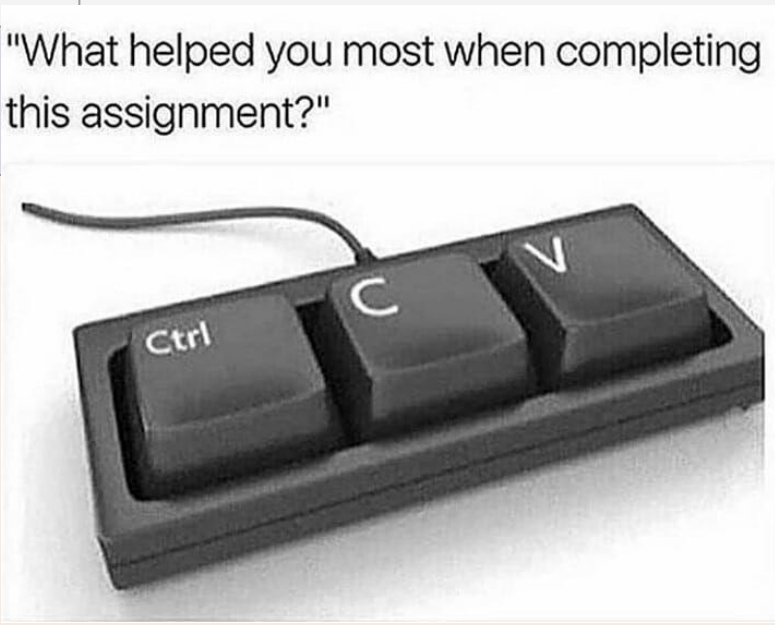


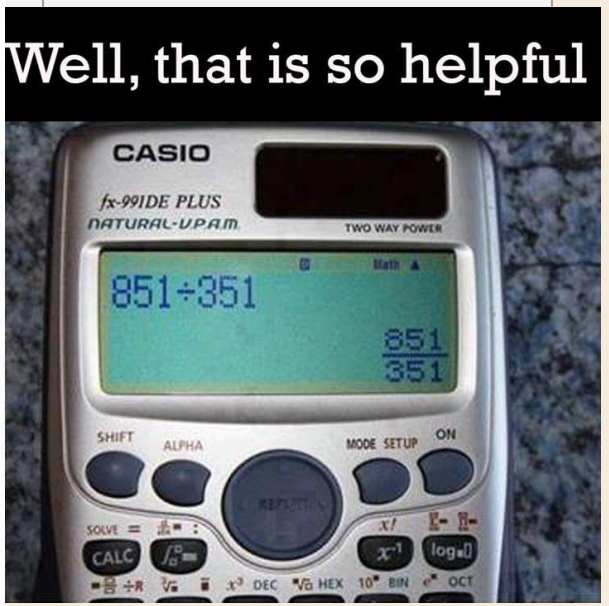
checklist

1. 文本的可信度
   - 有限的對象
2. 抓出目標
   - 僅抓了食材
   - 沒有處理其他烹飪、形容、廠商等資訊
3. 統計結果
   - 在貼文中出現的頻率
4. 解釋
   - 看得出常用嗎?

回到之前的發想及問題

原先的問題			
• 要怎麼挑選分析對象?	可信度評估依據?	• 要怎麼有效率的看每一篇貼文?	• 要怎麼總整處理結果?
少量	主觀假設		





回到之前的發想及問題

原先的問題			
• 食材是甚麼？	基本調味料的處理?	• 怎麼驗收常見？	• 怎麼進行大量、長期對比分析？
飯 : 食材 炒飯 : 料理	把別稱細項也計入總數, 並額外列出	這個食材在選定貼文中出現的頻率	

-I_have_no_idea / result.

繼續發想

1. **IG爬蟲**
   --自動擷取IG貼文
2. **Articut Premium**
   --無時間字數限制
3. **食材數據Data base: 政府開放式平台**
   --自動提出食材
4. **(硬體設備$)**
   --資料太龐大

1. betty.bento：一個月以上
2. hashtags 食譜：一週看能抓__篇

3. 貼文：txt
4. articut + regex

5. result: 單一作者

貼文:檔名
西元年+月+日

1. 一天一個
2. 一週一個檔

工作分配

- 陳辰:
  主題內容發想、簡報製作、內容討論、資料分析
- 劉鎧葶:
  內容討論、文本蒐集、code、資料分析
- 溫嘉煒:
  內容討論、資料分析

提問及回饋

thank (v) you (pronoun) very (adv) much (adv)

文本分析與程式設計

I have no idea

感謝peterwolf 的用心 & 助教們的熱心!!
辛苦了(╥﹏╥)
也謝謝同學們

12/22 小結論與代辦事項

定義常見與常用
抓資料的方式

Reference

圖片變文字：pytesseract　https://reurl.cc/VXrxnb

Comments from Peter, 12/22

Github 上要有page

正規化 instagram hosts
的食材的影響力與followers 的數量

note 1229
對於平均大四的 "詩"

perter wolf :

主詞 : 我
形容詞

找現成的詩句，分析詩人喜歡的習慣及用詞，就能用這個架構生出類似的詩。
function word : 如、像、般

content word :

noun水珠
noun noun : 對岸的柳枝
verb 掠過、靠近