



文本分析與程式設計



I have no idea ●

2020.12.22

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：營養或飲食媒體





Target Problem and Solution

食材使用率
名字出現的頻率

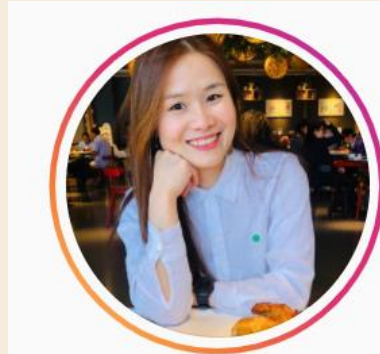
- 是否包含特定品牌、商家
- 可能的或普遍的烹飪方式
- 可能的形容詞
- 潛在的熱潮迷思

Data

Instagram：食譜部落客、
example: betty.bento

Why instagram: 年齡層、

其他可能性
Facebook:
Youtube



betty.bento

Follow

1,112 posts

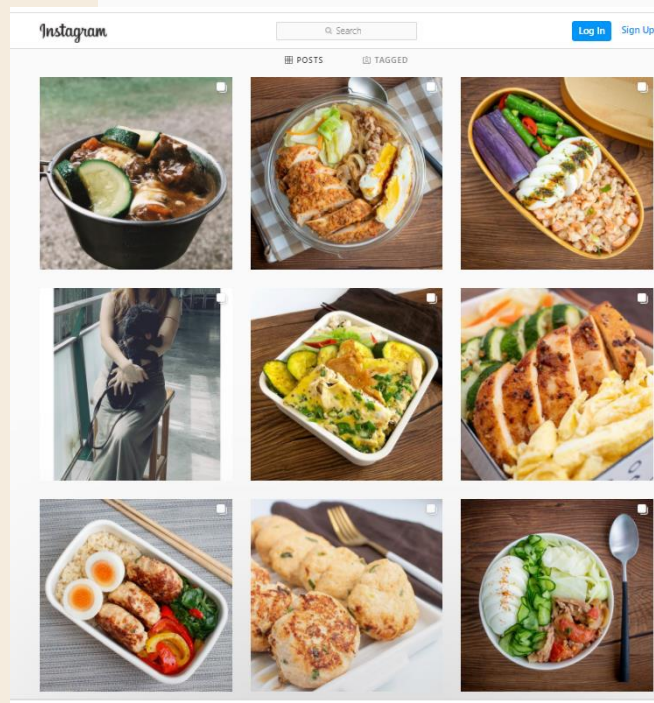
306k followers

609 following

貝蒂做便當

我寫了三本瘦身料理書👍

開啟輕鬆下廚、健康瘦身生活。



新出版：《愛妻瘦身便當【減醣成功三部曲】》

2018年出版《#愛妻瘦身便當》

2019年出版《愛妻無壓力瘦身便當》

❤️各大網路書店、實體書店都可以買到喔～

inktr.ee/betty0620

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

questions

- 要怎麼挑選分析對象？

1. 資料上：香港人

2. 可信度評估依據：追蹤數高的？按讚數？

Peter_12/22: 如果食材很少見但追蹤人數多而與事實不相符？

- 要怎麼有效率的看每一篇貼文？

- 要怎麼處理中英雙語食譜？

- 要怎麼總整處理結果？



12/22回應

1. 會挑選複數位 + h a s h t a g s

But Hashtags 不一定一致

3. 如果是貼文為單位，也就是一篇食譜有提到該項食材。
其實就不太會有追縱數的問題

And 要怎麼驗收常見？ 如果用按讚數就會跟追蹤數有關

5. 如果有一個明確的對比？ 另一個介面：紙本？

6. 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

7. 抓資料的具體方式：時間範圍、怎麼抓



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

