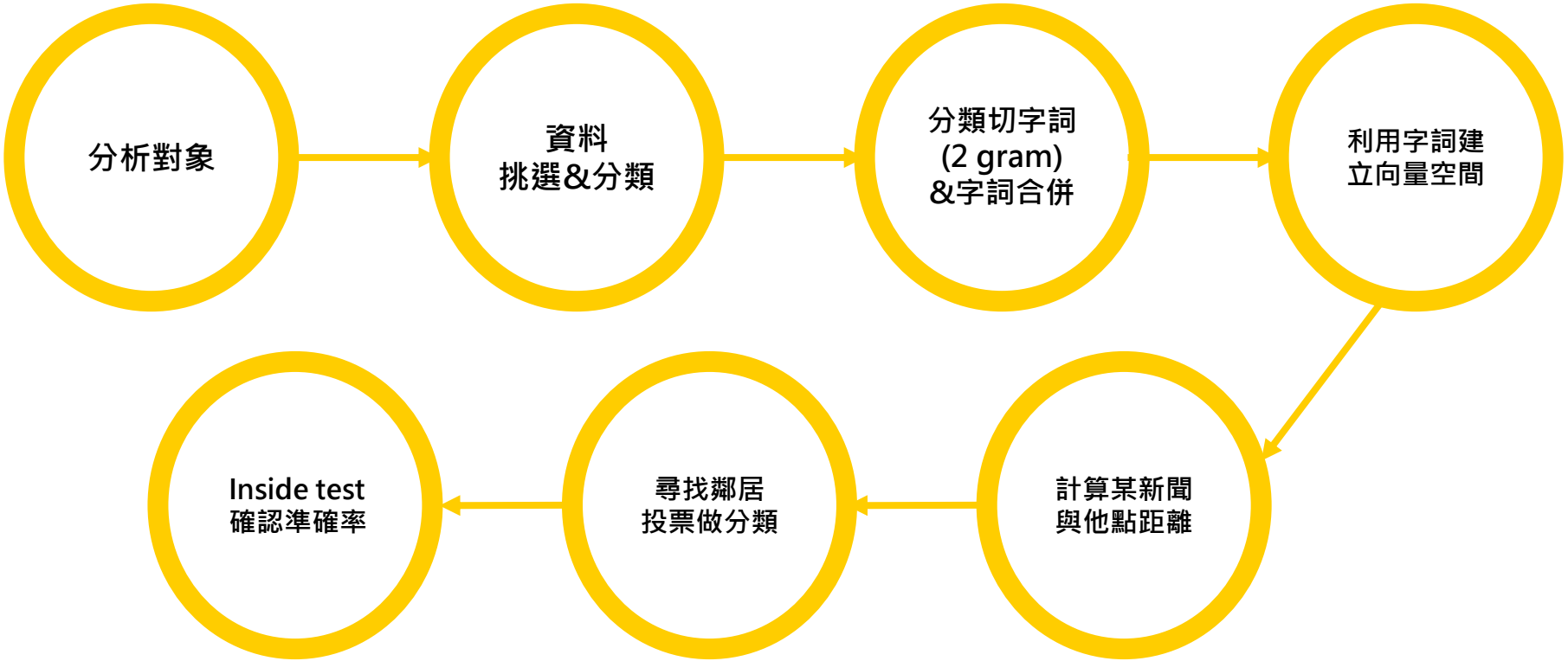
大數據與商業分析

期中報告

Team 1
資管三：b03705002 林軒逸, B03705030 張睿君, b03705009 單開民
國企三：b03704046 余采庭, B03704051 李沛璇,
B03704073 黃心柔, B03704044 何勁儀

大綱


分析對象
資料
挑選&分類
分類切字詞
(2 gram)
&字詞合併
利用字詞建
立向量空間
計算某新聞
與他點距離
尋找鄰居
投票做分類
Inside test
確認準確率

1

分析對象

如何選定分析對象

選擇標準

- 資料量
- 敏感度

選擇結果

- 鴻海
- 友達
- 聯發科

2

資料挑選&分類

GOUP GODOWN訓練集

訓練集

- 以鴻海、友達、聯發科篩選文章
- 每日的每篇文章，對應到**該日七天後均價**
- 沒有股價資料 → NODATA
- COMMON（不顯著）：
  |Price(D+7)-Price(D)| / Price(D) < 5%
- GOUP：Price(D+7) > Price(D)
- GODOWN：Price(D+7) < Price(D)

3

分類切詞與字詞合併

切詞為2 gram與將字詞合併

漲跌訓練詞

- 切出**2 gram**的詞
- 去掉共同出現的字詞
- 合併字詞
- **GOUP類**關鍵詞⇒代表**上漲**的詞
- **GODOWN類**關鍵詞⇒代表**下跌**的詞

4

建立向量空間

利用Python建立維度為dim{GOUP+GODOWN}的向量空間

使用Python

- 去除NODATA新聞
- 合併GOUP、GODOWN關鍵詞 → **List**
- 尋找每篇新聞向量

尋找向量：三種方法

- 0/1法：有該字詞為1，沒有為0
  eg. (0,1,1,0…)
- 次數法：計算該字詞在該新聞出現次數
  eg. (0,3,2,0…)
- tf-idf法：利用該字 df 與 tf 值計算 tf-idf 值
  eg. (-3.333,4.3,3.2,-3.333…)

向量空間

- 計算訓練集向量，記錄於向量空間
  →每篇已確認類別新聞都在向量空間特定一點
- 每篇欲查詢趨勢新聞皆有一個向量
  →代表在向量空間的特定一個點
- 向量空間建立完畢
  →kNN法評估

5

計算某新聞與他點距離

利用數學式計算某點與訓練集各點距離

$$D = \sqrt{(x1 - x2)^2 \quad + \quad (y1 - y2)^2 \quad + (z1 - z2)^2 \quad \ldots}$$

Simple, but useful.

計算距離

- 計算每篇欲歸類新聞與訓練集各點距離
- 預留**前四個**最小距離者

6 尋找鄰居，投票做分類

利用kNN找3鄰居做字詞分類

3-NN

- **前三個**鄰居進行投票
- GOUP、COMMON、GODOWN
- 得兩票者成功歸類

→ 如果各得一票呢？

同票處理

- 利用**第四個鄰居**
- GOUP、COMMON或GODOWN
- 直接歸類於該分類

7

Inside Test 確認準確率

比較三種方法準確率

準確率分析

- 分析資料：訓練集（非NODATA新聞）
- 進行切詞、向量空間與kNN計算
- 結果（預測）vs type（解答）

⇒ 類別相同者正確，反之錯誤

0/1法結果

- 聯發科：80.71%
- 友達：73.91%
- 鴻海：90.5%

次數法結果

- 聯發科：80.71%
- 友達：74.69%
- 鴻海：90.02%

tf-idf法結果

- 聯發科：82.77%
- 友達：74.69%
- 鴻海：90.66%

結論

- **tf-idf法**整體**準確率較高**，但增加幅度不高
- 0/1法與數數法各有高低準確率之公司

8 使用資源

本次作業使用之程式及資源

Excel

- 資料分類

Python

- 分割字詞
- 計算字詞之tf、df值
- 建立向量空間
- kNN (3-NN)

靈魂之窗

- 預期：程式自動化
- 實際：看看法，合併字詞

9 Demo

有時間的話(?

Thanks!

Thanks for Listening !

You can find us at

- BDA TEAM 1