文本分析與程式設計
Week02

本課程由卓騰語言科技贊助

文本分析任務

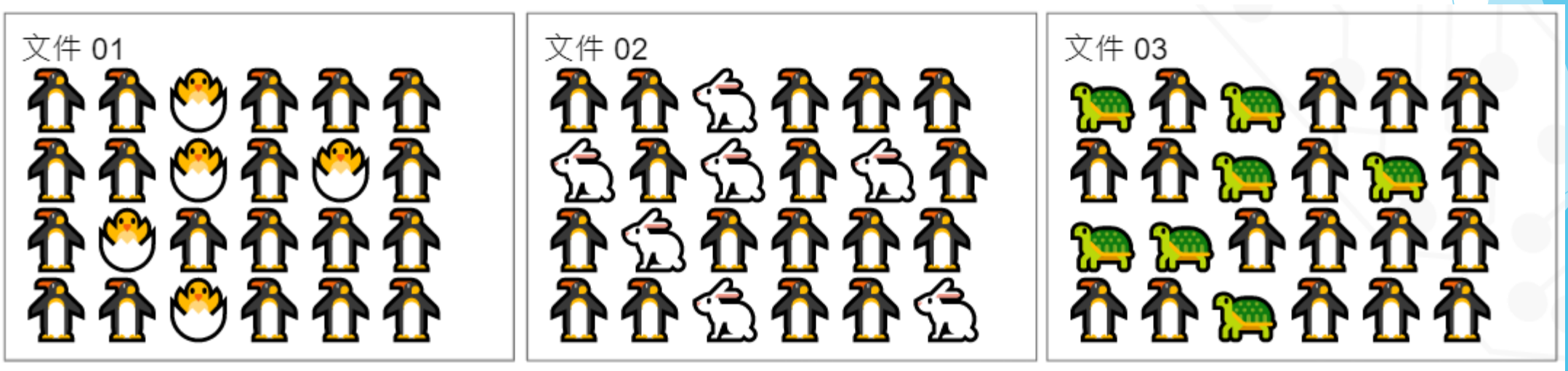
- 文本分析是常見的自然語言處理(Natural Language Processing, NLP) 任務。
- 本節將示範文本分析中三種常見的取特徵詞手法
  - 一是取 TF-IDF 特徵詞。
  - 二是依詞性取特徵詞。
  - 三是依「人、事、時、地、物」取特徵詞。

什麼是特徵詞

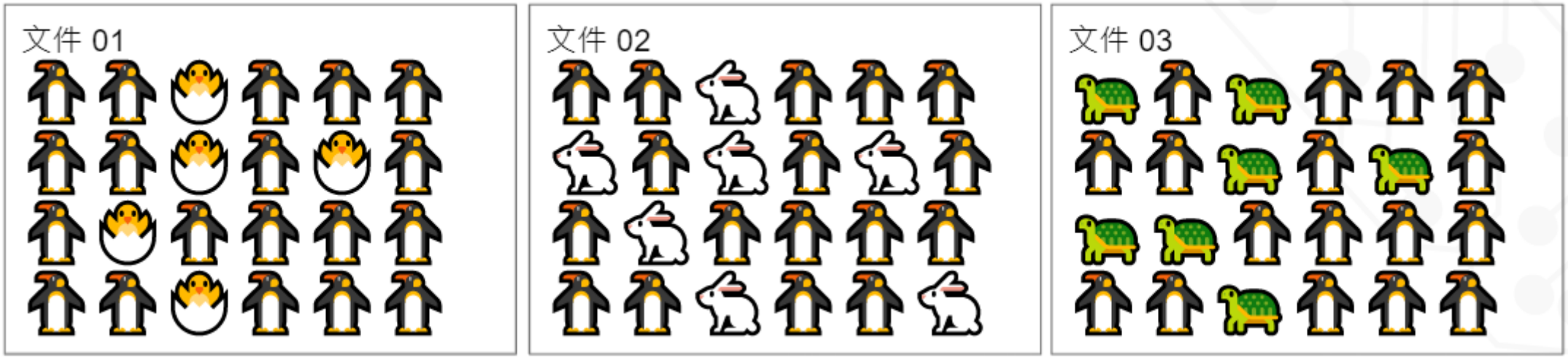
- 特徵詞可以做為代表文本內容的一種參考維度。
- 換句話說，一篇文本之中，那些字詞是可以拿來區辨不同的文件

什麼是特徵詞

- 以下面這三個由emoji 的文件來舉例，請問什麼圖案可以拿來區分這三個文件呢？
- 也就是說，哪一個「詞」是以下文件的「特徵詞」呢？

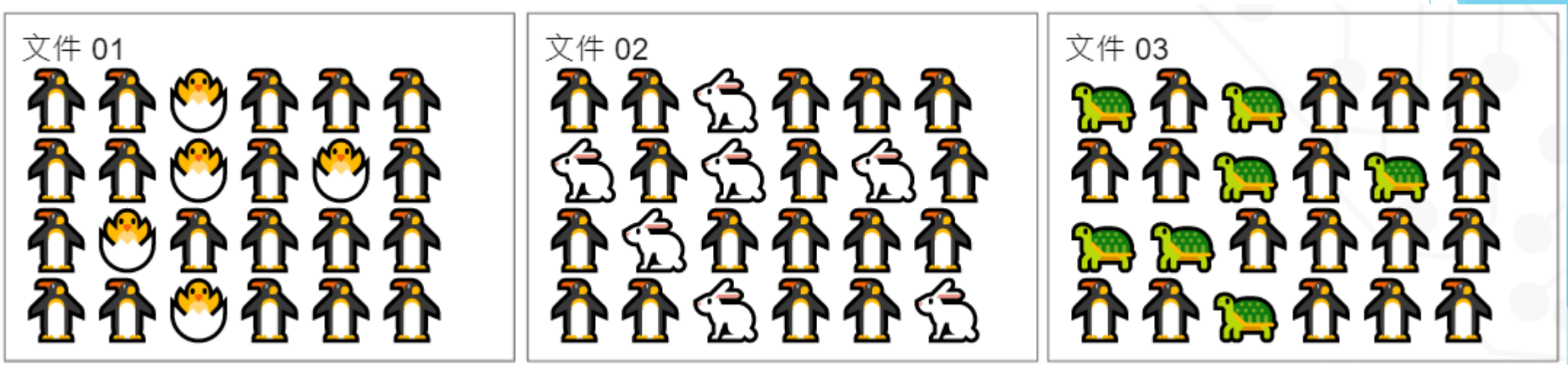


什麼是特徵詞



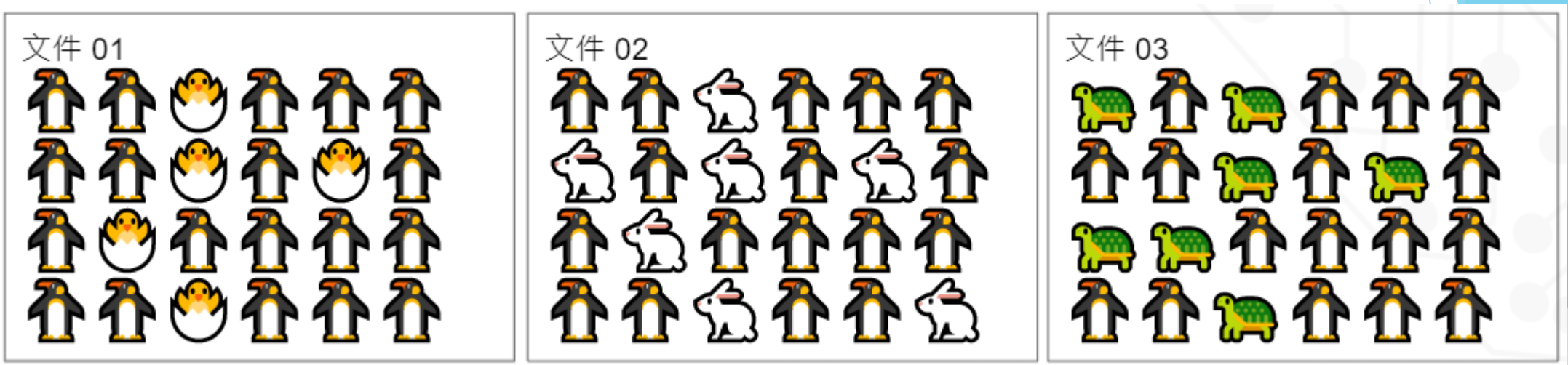
- 我們會說：
- 第一個文件的特徵詞是
- 第二個文件的特徵詞是
- 第三個文件的特徵詞是

什麼是特徵詞



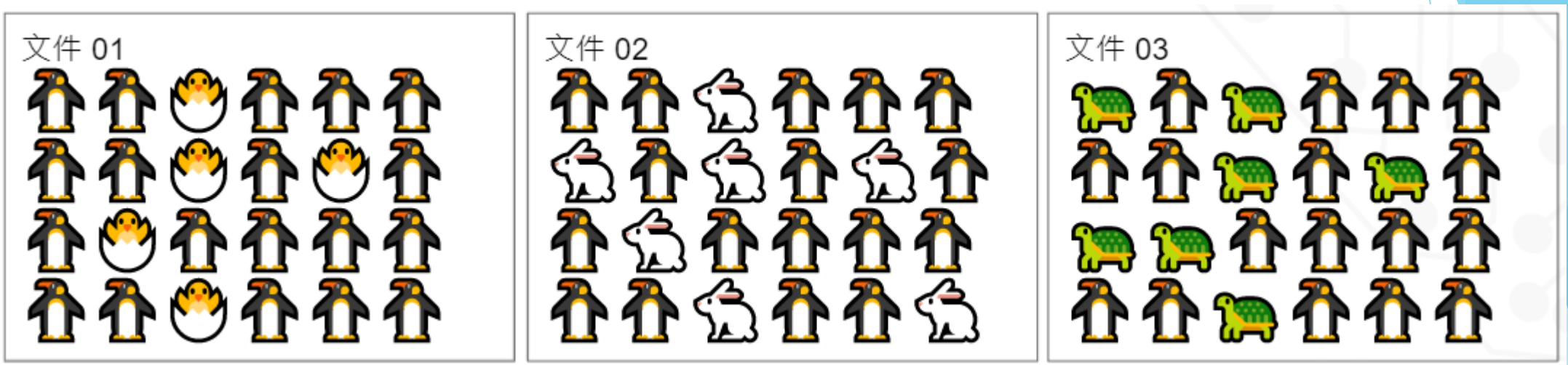
-    是特徵詞，因為他們是分別這三個文件中比較特別的存在

什麼是特徵詞



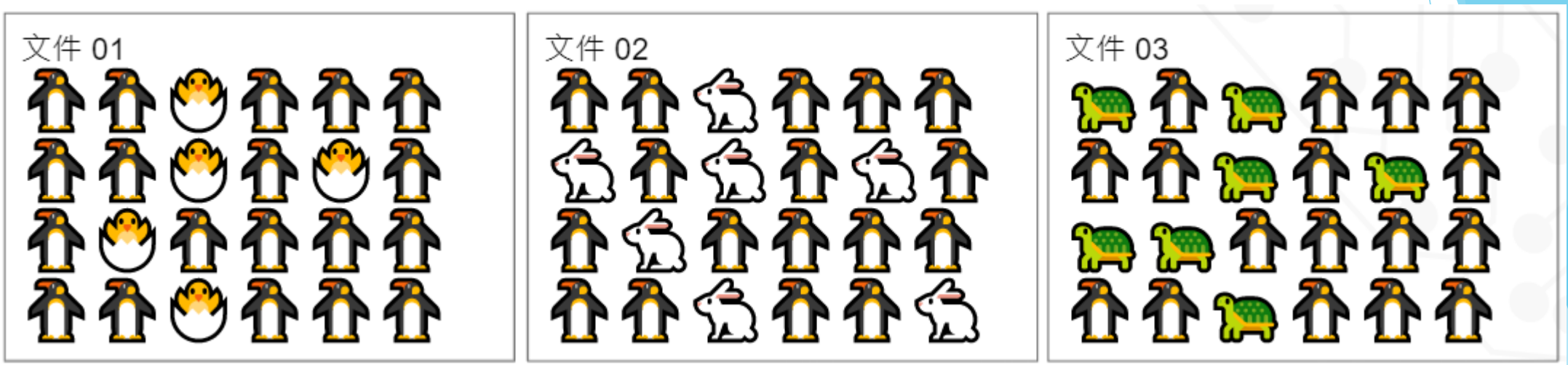
- 特別的是，當人類在做判斷的時候，我們不會去拿每個文件中數量最多的  來當作特徵。
- 反而我們會用  這三個數量比較少的當作該文件的特徵，因為我們覺得這是這些文件「特別」的地方。

什麼是特徵詞



- 所以從以上的例子，我們可以知道一個詞出現的「頻率最高」不能當作判斷特徵詞的唯一指標。
- 那麼，該怎麼找到屬於人類判斷的「特徵詞」呢？

特徵詞取得方法



- 我們有三種方法
  - 取 TF-IDF
  - 看詞性
  - 看「人、事、時、地、物」

用TF-IDF來取得特徵詞

什麼是TF-IDF

- TF-IDF 是一種統計方法，用在計算某一詞在一篇文章中的重要性。
- 而TF 和 IDF 各有自己的意思

什麼是TF-IDF

- TF 是 Term Frequency 的縮寫，意思是某個字在本文件裡出現的加權後的頻率
- 所以計算方法可以從計算每一個字的數量開始
- 我們會以這個emoji 文件為例

什麼是TF-IDF

- TF 的計算方法如下，以右邊的emoji 文件為例

1. 計算每個詞在本文件中的頻率
   例如：

    有19個
    有5個

2. 再來以下面算式計算他們的比重
   這個比重就是 TF
   例如  的TF 是 19 /(19+5) 大約是 0.792

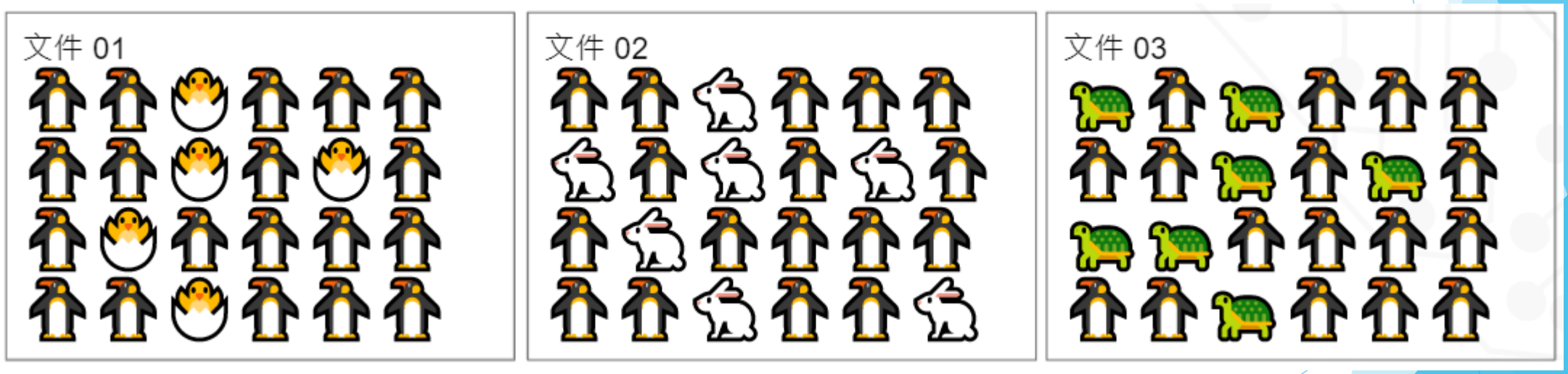
    的TF 是5/(19+5) 大約是 0.2

什麼是TF-IDF

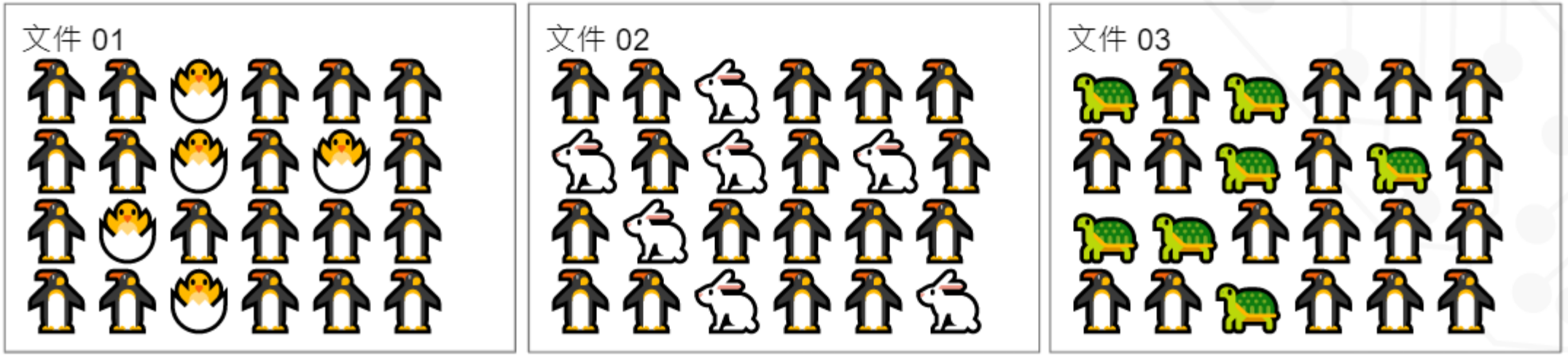
- 從上頁的例子，我們可以知道企鵝的TF比較大
- 所以從TF 的角度來說，企鵝比較重要

什麼是TF-IDF

- IDF 是 Inverse Document Frequency的縮寫，也就是某個字在所有的文件中出現的頻率
- 我們再來看看這個emoji 文件

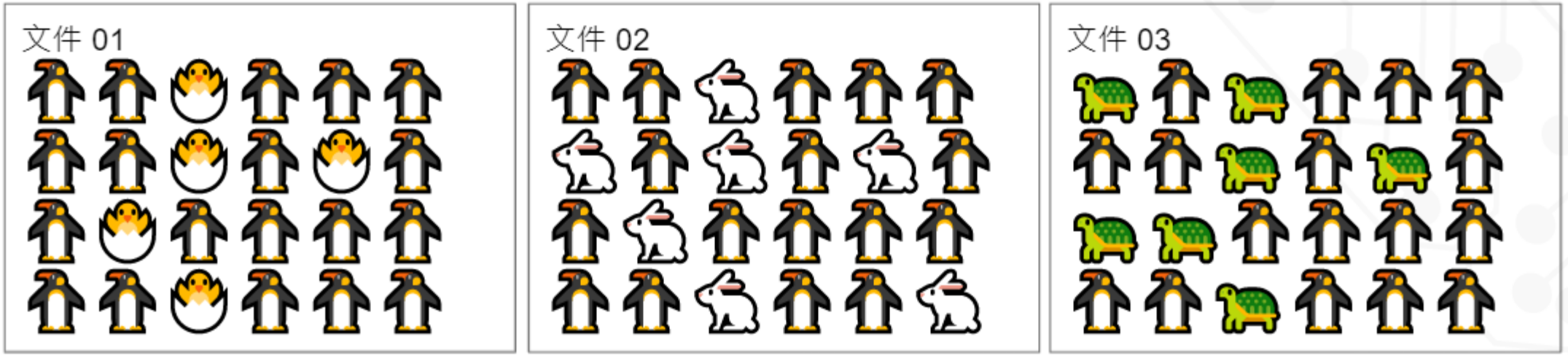


什麼是TF-IDF



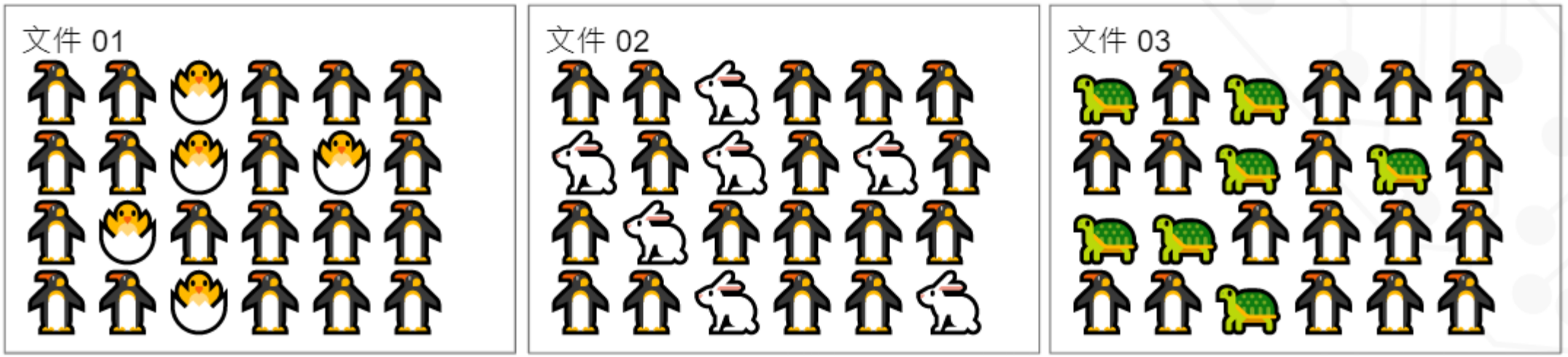
- 計算IDF時，我們要先計算這些內容
  1. 我們需要知道我們有多少文件 ➔ 目前我們有三個文件
  2. 某個字出現在文件中的次數
     例如
     ➔ 🐧 出現在這三個文件中
     ➔ 🐣🐰🐢 各出現在一個文件中 (我們用 🐣 當例子)

什麼是TF-IDF



- 計算IDF時，我們要先計算這些內容
  3. 接下來就是將剛剛的文件數取log
  4. 最後按照這個算式：log(文件數/該單詞出現在幾個文件)
     - ➔ 所以 🐧 是log(3/3) = 0
     - ➔ 🐥 是 log (3/1) 大約等於 0.477

什麼是TF-IDF

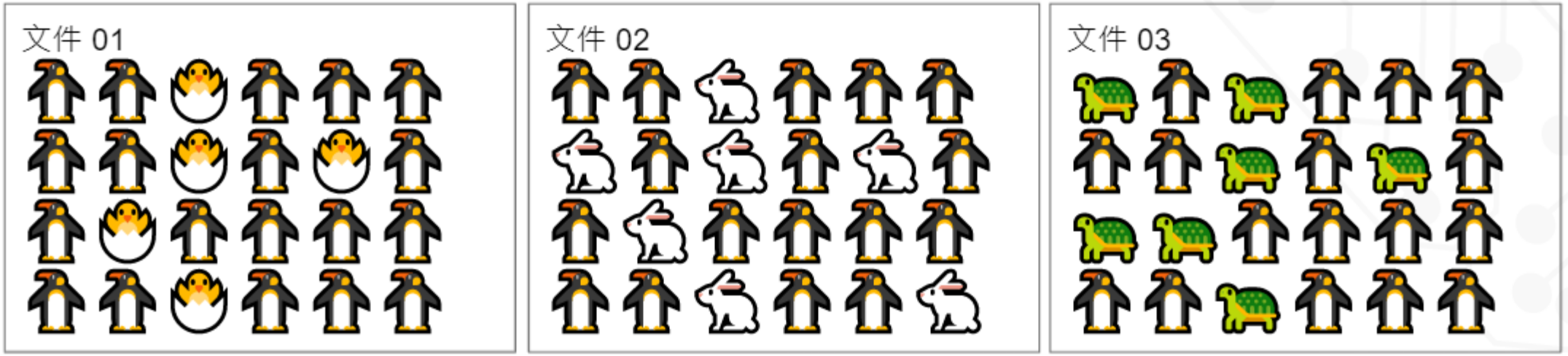


- 計算之後我們發現  的算出來比較大，然後  比較小
- 這個代表什麼呢？
- 這是代表  在IDF的角度中，比較重要

課間練習1

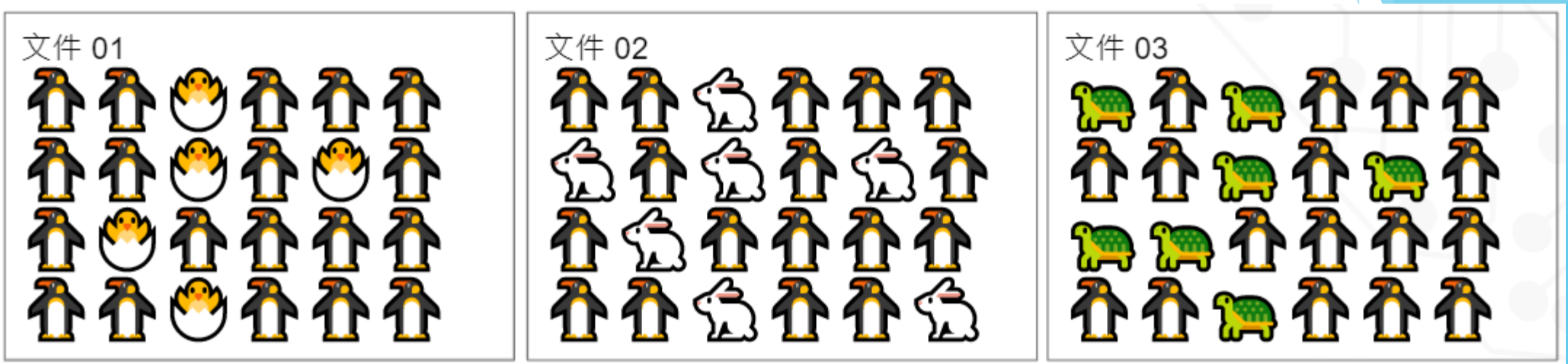
- 看完TF 和 IDF 的內容，你覺決TF 和 IDF 哪個比較接近我們人類思考之中的特徵詞呢？

什麼是TF-IDF



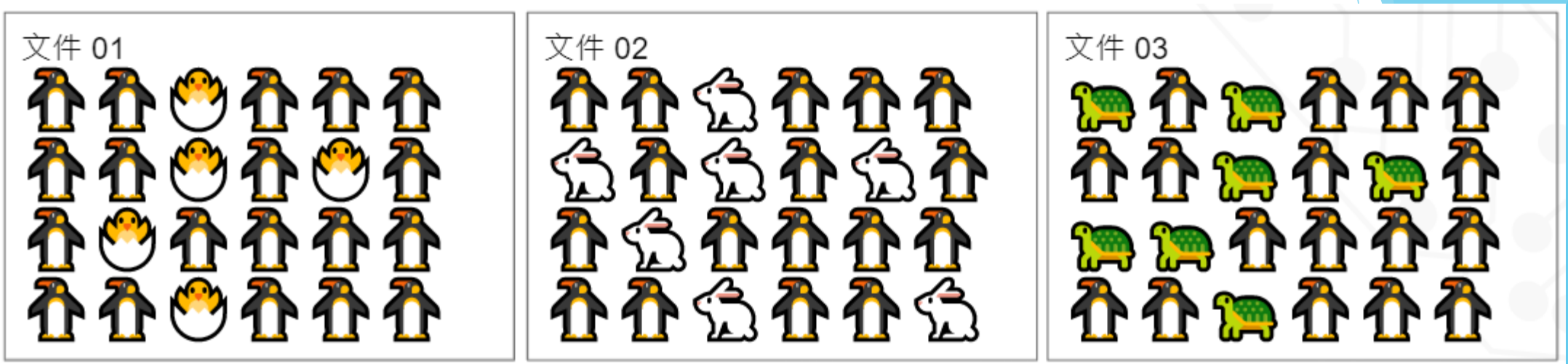
- 最後，我們可以將我們計算出來的TF 和 IDF 相乘
- 例如 是0.792 * 0 = 0
- 而 是 0.2 * 0.477 = 0.0954 ➔ 比較重要
- 0 和 0.0954 就是所謂的「 TF-IDF 權重值」

TF-IDF 可以告訴我們什麼？



- 現在我們知道TF 和 IDF 背後的原理，那TF-IDF 可以告訴我們什麼呢？

TF-IDF 可以告訴我們什麼？



- TF-IDF 可以告訴我們文件數量和詞的頻率之間的關係
- 所以我們可以發現TF-IDF 可以幫助我們過濾常見的詞。常見字詞為什麼要過濾呢？我們先來做課間練習二來想想看

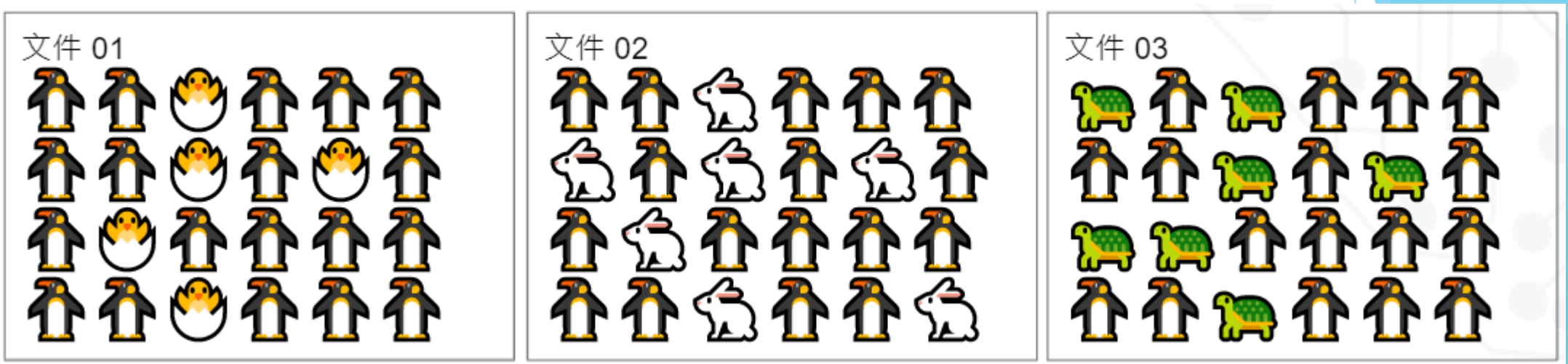
課間練習2

- 請看下一頁的新聞，並利用前一堂課學習到的 Articut 斷詞方法斷詞，並計算頻率，什麼字詞頻率最高？高頻率的那些字是可以代表那篇新聞的字詞嗎？

課間練習2

- （中央社記者葉冠吟台北27日電）入圍坎城影展、改編自村上春樹短篇小說的日本電影Drive My Car，由西島秀俊、岡田將生主演，台灣演員袁子芸也參演，更透露西島秀俊本人相當親切，還有許多暖男舉動。袁子芸曾參演戲劇「泡沫之夏」、「未來媽媽」，Drive My Car是她的首部日本電影，飾演一名母語是中英雙語的女演員，在角色背景設定上與她自身有許多相似之處。袁子芸表示，開拍第一天自己很緊張，而且第一場戲就與岡田將生有近距離接觸。不過當時午餐放飯時間延遲，拍戲時大家都有點餓，沒想到在最安靜的瞬間，岡田將生的肚子竟然咕嚕叫了一聲。袁子芸回憶，當下兩人都很專業的繼續演出，等導演一喊卡之後，才忍不住大笑出來，連岡田將生都笑著表示不好意思。不過也因為這個意外插曲，沖淡第一天拍攝的緊張和陌生感，瞬間拉近演員之間的距離。關於男主角西島秀俊，袁子芸看過許多他的作品，原以為他是比較有距離感的大明星，沒想到西島秀俊在讀本過程就相當親切，還有許多暖男舉動。袁子芸分享，有天海外演員殺青後，相約要去慶祝，但因西島秀俊隔天一早還有拍攝，無法同行，他便私下交代製片想請大家吃飯，麻煩製片先代為招待。近日電影傳出入圍坎城影展好消息，袁子芸也相當開心，她表示劇組群組簡直要暴動。雖然她很希望能和劇組一起慶祝，卻因大家都分散不同地方，又正值疫情期間，只能透過通訊軟體互相道賀，期待Drive My Car 能在坎城影展奪下好成績。由柏林影展銀熊獎得主、導演濱口龍介執導的Drive My Car，描述一名想尋找亡妻生前外遇對象的舞台劇演員，藉由與沉默寡言女司機的對話，描繪出各自的故事與情感，電影8月20日將於日本上映。（編輯：張雅淨）1100627
- The news is from https://www.cna.com.tw/news/amov/202106270091.aspx

TF-IDF 可以告訴我們什麼？



- 通過剛剛的課堂練習2，還有以上我們很熟悉emoji文件，以及我們前面討論的討論詞，我們會發現反而最常出現在文件中的詞，通常都並不是最可以突出文件特點的詞，反而只有出現在某些文件的才是特徵詞
- 如果還想更多了解TF-IDF 可以參考以下內容：https://bit.ly/tfidf_explain

如何利用TF-IDF 來取出特徵詞

- Articut 內建的 analyse 工具包裡的 extract_tags() 函式來取得經TF-IDF計算的特徵詞
- 延續上週課程中的棒球與籃球的例子

```
baseball_TFIDF = articut.analyse.extract_tags(baseballResultDICT)
print(baseball_TFIDF)

basketball_TFIDF = articut.analyse.extract_tags(baseballResultDICT)
print(baseball_TFIDF)
```

如何利用TF-IDF 來取出特徵詞

- 棒球語料的 TF-IDF 存在 baseball_TFIDF 裡，印出後內容如下：

執行結果

['\n', '大都會', '馬林魚', '投手', '敲出', '安打', '2', '1', '巴斯', '局面', '康福托', '比賽', '被', '週三', '紐約', '此戰', '使用', '車輪戰', '4名', '輪番', '上陣', '壓制', '打線']

- 籃球語料的 TF-IDF 存在 baseball_TFIDF 裡，印出後內容如下：

執行結果

['\n', '兩', '米契爾', '康利', '罰', '勝利', '太陽', '攻勢', '爵士', '籃板', '3', '助攻', '11分', '投', '一', '昨晚', '紐約', '西區', '霸王']

TF-IDF 的計算結果：觀察

- 從兩篇文章的 TF-IDF 特徵詞裡，如果沒有相關的背景知識知道「大都會」、「馬林魚」是棒球大聯盟的球隊，而「爵士」是 NBA 的球隊名稱的話，其實很難確認究竟哪些詞彙是可以做為「棒球類」或是「籃球類」的文本分類特徵詞。

TF-IDF 的計算結果：觀察

- 但我們可以確認的是「投手」、「安打」、「打線」這三個名詞，應該是只有棒球類的文本才會用到。而「籃板」這個名詞，則是籃球類的文本才會用到。同理，我們也能觀察到「敲出」、「保送」、「打擊」這幾個動詞同時出現的文章，有很大的機率是在描寫棒球類的文本。而「助攻」、「上籃」則常見於描寫籃球比賽過程的動詞。

TF-IDF 的計算結果：觀察

- 這個觀察最重要的是讓我們發現「詞性」(前文提到的「名詞」、「動詞」) 是可以做為抽取特徵詞的一種方法。接下來，我們利用 Articut 的詞性篩選工具來取出「名詞」和「動詞」。

用詞性來取得特徵詞

名詞比較-輸入

- 我們使用 Articut 內建的 .getNounStemLIST() 函式來取得文本中各句的名詞。

```
baseballNounLIST = articut.getNounStemLIST(baseballResultDICT)
print(wordExtractor(baseballNounLIST))

basketballNounLIST = articut.getNounStemLIST(basketballResultDICT)
Print(wordExtractor(basketballNounLIST))
```

名詞比較-輸出

- 棒球報導文本的名詞

執行結果
['主審', '優勢', '分', '勝利', '安打', '局面', '意識地', '手', '打線', '打者', '投手', '此戰', '比賽', '滿壘', '球', '球隊', '登板', '眼', '角滑球', '觸身球', '身肘', '身體', '車輪戰', '追平比數', '這球', '陽春砲', '領先']

- 籃球報導文本的名詞：

執行結果
['中', '人', '勝利', '單場', '場', '布克', '延長賽', '戰', '攻勢', '機會', '次', '波格丹諾維奇', '籃板', '階段', '霸王']

- 從名詞裡看出來，有些詞彙 (例如「勝利」) 是重覆的，要得到更好的效果的話，還可以把重覆的詞彙去除以便讓不同的類別都具有各自獨立的特徵詞表。

動詞比較-輸入

- 我們使用使用 Articut 內建的 .getVerbStemLIST() 函式來取得文本中各句的名詞。

```
baseballVerbLIST = articut.getVerbStemLIST(baseballResultDICT)
print(wordExtractor(baseballVerbLIST))

basketballVerbLIST = articut.getVerbStemLIST(basketballResultDICT)
print(wordExtractor(basketballVerbLIST))
```

動詞比較-輸出

- 棒球報導文本的動詞

執行結果
['上場', '上陣', '伸進', '使用', '保送', '再見', '判定', '到', '壓制', '失', '帶', '形成', '打出', '打擊', '投', '投出', '拿下', '接下來', '推出', '敲', '有下', '看', '碰觸', '讓', '贏得', '越來', '輪到', '進入', '關', '隨', '靠近', '面對']

- 籃球報導文本的動詞：

執行結果
['上籃', '分', '力圖', '助攻', '合計', '回應', '得分', '得手', '打出', '抄截', '投', '拿到', '施展', '比', '犯規', '狂轟', '用', '留給', '罰', '讀秒', '追分', '追平', '造成', '錯失', '鎖定', '開始', '隨']

- 從動詞裡看出來，有些詞彙 (例如「打出」、「投」、「隨」) 是重覆的，要得到更好的效果的話，還可以把重覆的詞彙去除以便讓不同的類別都具有各自獨立的特徵詞表。

用「人、事、時、地、物」取特徵詞

取得人事時地物的工具

- 如果你想知道的特徵詞可以以「人事時地物」來分類，那可以嘗試使用以下的工具來幫助你
- 「人」、「事」、「地」、「時」和「物」因為通常都是名詞，所以可以使用下面的工具
  - 可以用剛剛取名詞的.getNounStemLIST()
  - 可以用 articut.getContentWordLIST()，這個是用來取得content word
- 不過更詳細的內容，下週會再介紹

什麼是content words

- Content words，或是又稱「實詞」，指的是句子中有重要意義的意思。
- 例如以下句子 (取次於 Snow White 的維基條目)
- "Snow White" is a 19th-century German fairy tale that is today known widely across the Western world.
- 其中實詞就包括 Snow White, German, fairy tale
- 而相對於實詞，如果只有文法上功能的就是虛詞

什麼是content words

- 而相對於實詞，如果只有文法上功能的就是虛詞
- 例如在剛剛的句子中
  "Snow White" is a 19th-century German fairy tale that is today known widely across the Western world.
- a, that, the 就是虛詞

- 取自於https://en.wikipedia.org/wiki/Snow_White

取得人事時地物的工具

- 如果你想知道的特徵詞可以以「人事時地物」來分類，那可以嘗試使用以下的工具來幫助你
- 「地」
  - Articut 也可以直接擷取地方的名稱
  - 可以用 articut.getLocationStemLIST()

課間練習3

- 請使用籃球和棒球來練習
- articut.getLocationStemLIST() 和用 articut.getContentWordLIST() 和前面取名詞和動詞的方法是一樣的，請嘗試自己摸索使用看看。
- 觀察看看 content word (內容詞) 和 location word (地方名稱) 是否具有比「動詞」或「名詞」更好/更差的特徵表現能力。說說看你的觀察。
- 請直接使用之前的basketball 和 baseball 的 resultDICT