Natural Language Processing Summer 2021 #3

Chi-Jen Wu

General Information

- Time
  - Thu 09:10 - 12:00 and Thu 14:10 - 17:00
- Course Assistants
  - 陳秉宏, 沈育賢, 黃教華
  - m0829014@cgu.edu.tw, b0629056@cgu.edu.tw, b0629048@cgu.edu.tw
- Classroom:
  - Google Meet
  - https://csiecgu.slack.com
    - nlp
- Office Hour:
  - 1, 2, 3

Grading

- 35% Homework / Exercises
- 40% Final project
- 25% Participation/Quizzes

About final project

- Using github, only and **處理中文資料** only
- 兩種 個人 或兩人組隊
  - 個人
    - 選取 Kaggle 上的data set
    - 分析訓練模型並上傳至kaggle 上 評分
  - 兩人組隊
    - 自行爬網頁資料 標註
    - 分析訓練模型得出精準度

期末分組和報告順序

- 請到下面google sheets 填寫期末分組和報告順序
- 請於 今天 8/5 23:59 填寫完畢
- 謝謝

https://docs.google.com/spreadsheets/d/1ODGHrH66LeRWk-G5pAgRIHwsDL79jwYOfc7I2sVvXp8/edit?usp=sharing

Topics

- An introduction to NLP
- Language modeling
- Representation learning
- Web crawling and indexes
- Word Embeddings
- Text classification
- Sequence modeling
- Machine learning models
- Deep learning models

An introduction to NLP

NLP 相關工作

自然語言處理科學家(NLP)

儲存

竹 竹間智能科技有限公司
台北市內湖區

前往應徵：104人力銀行

6 天前 12 分鐘

1. 豐富的自然語言處理(NLP)實際項目工作經驗，包括文本分類 (text classification)、信息抽取 (information extraction)、知識庫構建 (knowledge graph construction)、主題詞標引 (keyword tagging)、自動摘要 (automatic summarization)等；熟悉機器學習(machine learning)的各類算法，並有實際項目使用經驗。
2. 能夠利用自然語言處理的理論和方法研發分詞 (word segmentation)、詞性標注 (POS tagging)、命名實體識別 (named entity recognition)等基礎算法；善於跟蹤自然語言處理業界最新動態，進行自然語言處理相關的其他算法預研。有搭建知識圖譜 (knowledge graph)和基於雲端的機器學習系統 (cloud-based machine learning)者優先。
3. 精通C, C++, Java, 或Python中的一門或多門語言，精通數據結構和算法設計，熟悉Linux / Unix系統和Shell編程; 有MapReduce或Hadoop等海量處理經驗優先。
4. 優秀的分析問題和解決問題的能力，對解決具有挑戰性問題充滿激情；對數據敏感，有強烈的好奇心，喜歡折騰數據並從數據中發現價值。
5. 較強的溝通能力和邏輯表達能力，具備良好的團隊合作精神和主動意識，良好的自我驅動和學習能力。

NLP 相關工作

自然語言處理研發工程師 Software Engineer - Natural Language Processing (AI)

儲存

華碩電腦股份有限公司(ASUS)
台北市北投區

前往應徵：1111人力銀行

3 天前

榮獲1111人力銀行「2019科技業幸福企業大賞」員工最幸福「電腦 / 消費性電子類別」前二十大企業

【工作內容】

Your responsibilities may involve one or more of the following tasks:

1. Design Chinese/English Syntactic & Semantic Grammar.
2. Design knowledge representations (Lexicon, Ontology, Logical Form, Knowledge Graph, and Common Sense Knowledge Rules).
3. Develop rule-based or machine-learning-based NLP pipeline tools (Segmentation, NER, Syntactic & Semantic Parsers, Co-reference Resolution, Discourse Analysis, and Logical Form to Knowledge-Graph Conversion).
4. Develop Knowledge Graph based Inference Engines for Question Answering, Free Chatting, and Persona Assistant.
5. Develop automatic Domain Language Model and Ontology Generation tools.
6. Develop automatic Knowledge Acquisition tools including Common Sense Crowd Sourcing tools.

【發展願景】

We are looking for a Natural Language Processing (NLP) Engineer to help us improve our NLP tools and create a framework for developing next-generation NLP/AI applications. NLP Engineer's...

AI自然語言處理工程師(NLP)

儲存

台泥企業團_臺泥資訊股份有限公司
台北市

前往應徵：104人力銀行

7 天前

【工作內容】

「環保、能源、水泥」是台泥三大核心事業，企業團旗下包含14家子公司，蘊含各式產業以及各個管理面向豐富的文本資料以及數據資料，AI自然語言工程師的首要任務就是利用這一些資源，幫助台泥企業團打造符合循環經濟的全新無人島。在島上你必須要用最新最好的NLP工具製作出各式道具，幫助企業團完成每一個與環境平衡共生的任務關卡。集合吧! NLP工程師們，我們需要您能具有下面的技能:

1.擁有中文/英文自然語言處理相關工作經驗，例如: 文本分類 (Text classification)、自動摘要 (Automatic summarization)、信息抽取 (Information extraction)、情感分析 (Sentiment Analysis)、中文分詞 (Word Segmentation)、詞性標記 (POS Tagging) 或是知識庫構建 (Knowledge graph construction) 等。

2.熟悉NLP的架構/方法，例如: Bert, Transformer, encoder-decoder, Attention, Autoencoder 或是RNN等。

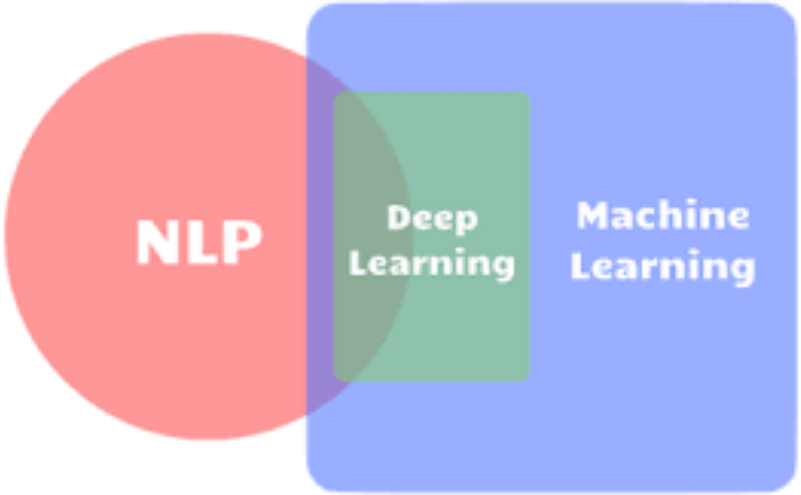
3.熟悉NLP相關工具/套件，例如: CkipTagger, Jieba, funNLP, huggingface/transformers, genism 或是OpenCC等。

4.精通下列一門或多門程式語言，例如: Python (TensorFlow, Pytorch, Keras, Pandas, Numpy, Scipy, Sklearn, Matplotlib, Seaborn) 或C++語言等。

5.具備獨立作業及溝通協調能力並有良好的團隊合作精神和主動意識。

Two Approaches to NLP

- Machine Learning
  - 處理語料
  - 特徵工程
  - 分類器
- Deep Learning
  - 處理語料
  - 設計模型
  - 模型訓練



處理語料

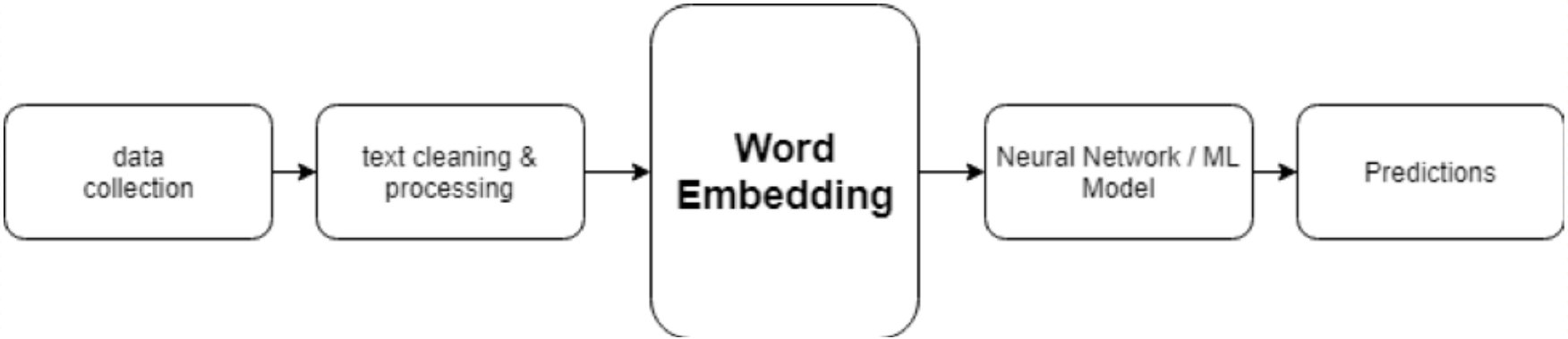
- 英文
  - 分詞 (Tokenization)
  - 詞性提取 (Stemming)
  - 詞性還原 (Lemmatization)
  - 命名實體識別 (Named Entity Recognition)
- 中文
  - 分詞 (Tokenization)
  - 詞性標註 (Parts of Speech)
  - 命名實體識別 (Named Entity Recognition)

NLP 程式工具大部分都會提供這些功能

Word Embeddings

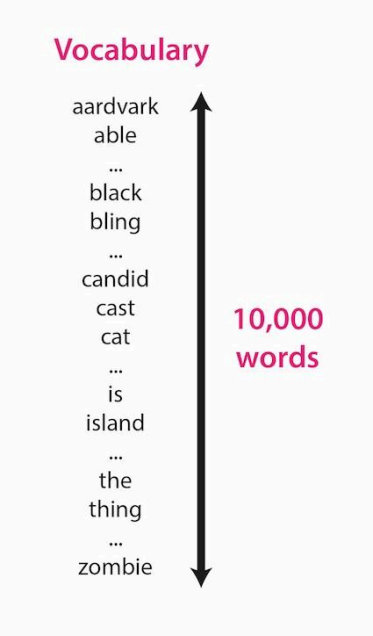
- 將字詞看作文本的最小單元
- 一個維數為所有詞的數量的高維空間映射到一個較低維數的連續向量空間中
- Embedding (映射 嵌入)
  - the representation of words
- 目前主流的方式 都以Neural Network (NN)為主
  - 之前介紹得 n-gram 就慢慢比較少人用
  - TF-IDF
- 目前大概分為兩大類 embedding 方法
  - Frequency based embedding
  - Predicted Based Embedding

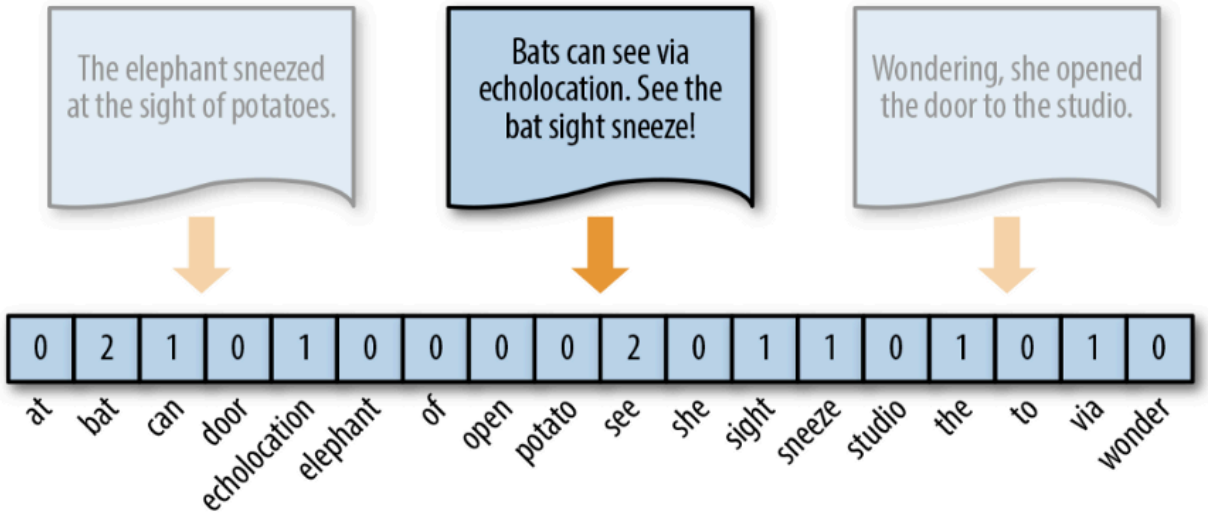
NLP flow


data collection
text cleaning & processing
Word Embedding
Neural Network / ML Model
Predictions

One Hot Encoding

- Traditional Context-Free Representations





Frequency based Embeddings

- Count Vector (每個詞的權重用其出現的次數來表示)
  - 一個文件情況

	about	all	cent	cents	money	new	old	one	two
doc	1	1	3	1	1	1	1	1	1

In theory (a)

Index	0	1	2	3	4	5	6	7	8
doc	1	1	3	1	1	1	1	1	1

In practice (b)

Frequency based Embeddings

- Count Vector
  - 多個文件情況

字詞

文件	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

Sparse matrix

Frequency based Embeddings

- TF-IDF Vector
  - 字詞的重要性與它在檔案中出現的次數成正比，同時與它在語料庫中出現的頻率成反比

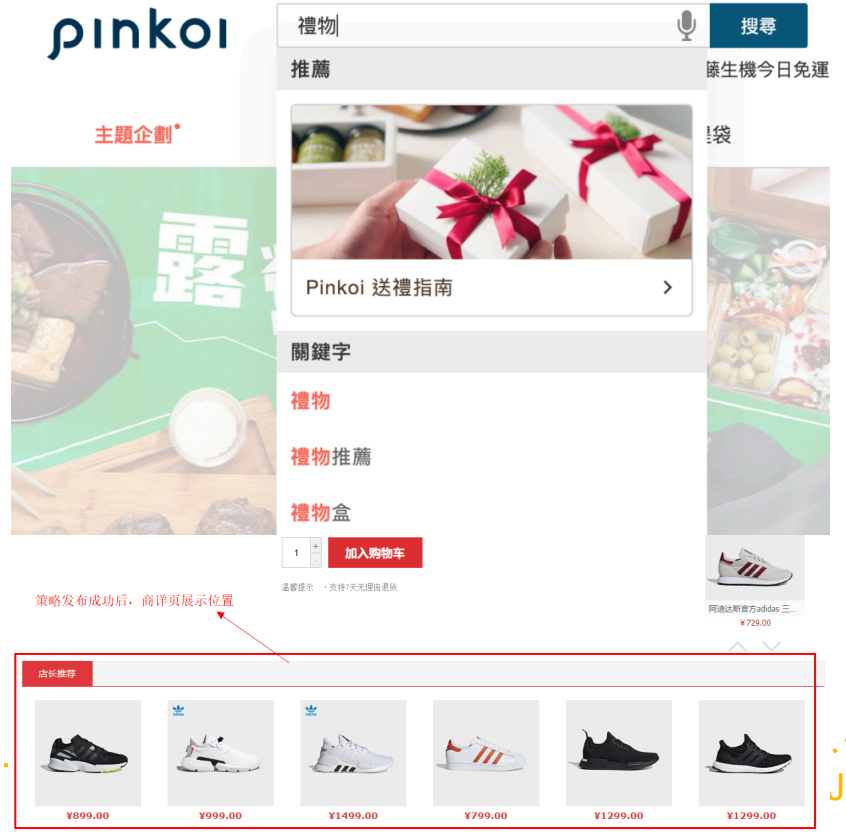
文件	the	red	dog	cat	eats	food
1. the red dog →	0.2	0.2	0.2	0	0	0
2. cat eats dog →	0	0	0.2	0.2	0.2	0
3. dog eats food →	0	0	0.2	0	0.2	0.2
4. red cat eats →	0	0.2	0	0.2	0.2	0

字詞

Sparse matrix

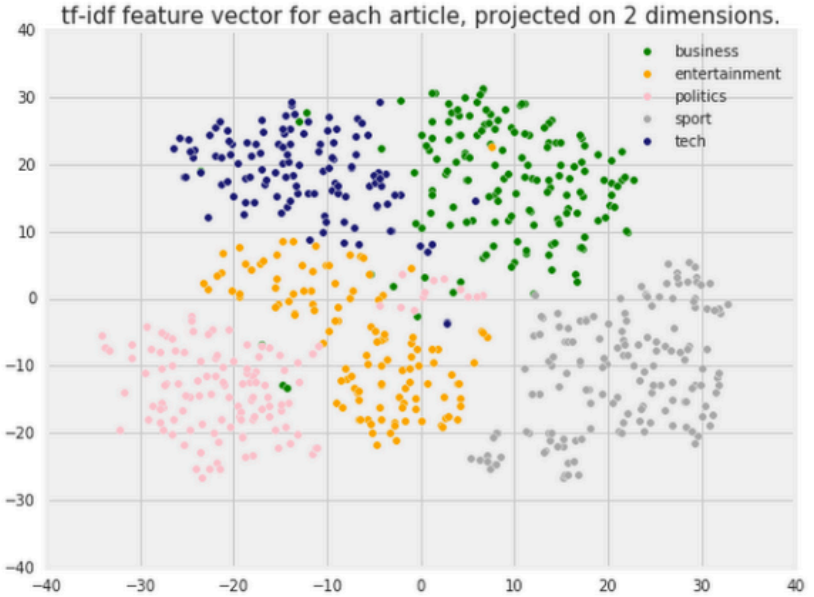
文件分類- NLP應用

- Text classification
  - Document classification
  - **電影分類**
  - 相似商品(**電影**)推薦
- Information Retrieval
  - 尋找相似的文件
  - Search engine



文件分類 by TF-IDF

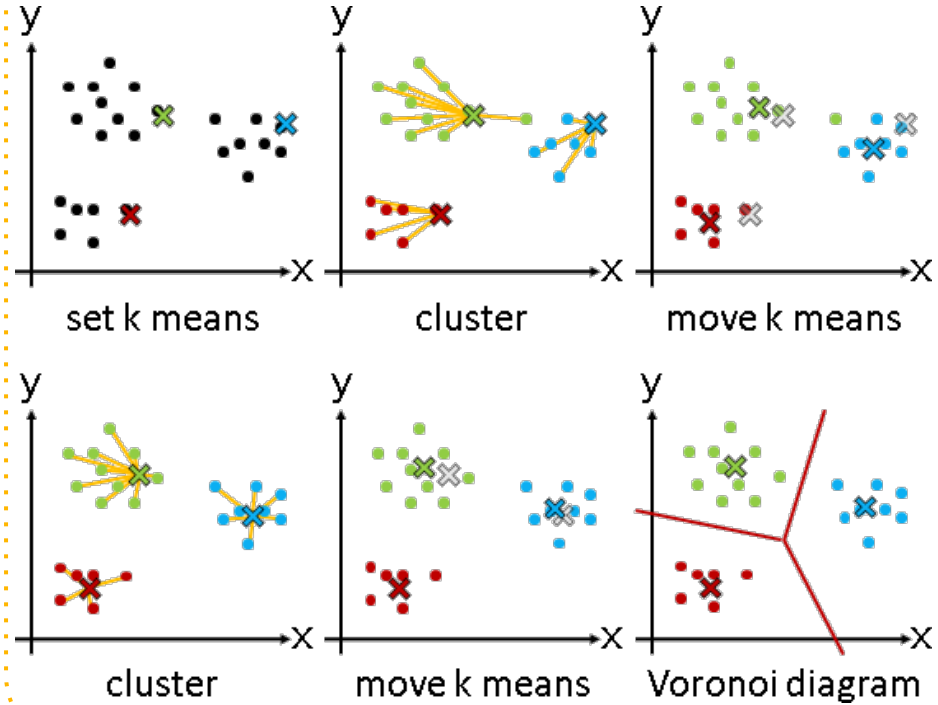
- 利用文件的高tf-idf的字詞分群
  - Topic Modeling
  - 每個文件的主題
  - 利用主題關鍵字詞分群
  - Unsupervised learning
    - K-means algo
  - Supervised learning
    - SVM, KNN



Unsupervised learning

- K-means algorithm (算是最簡單的NLP算法之一)
  - 物以類聚
  - k是分成幾群, means就是每一群群心
  - 群心初始位置 (亂給, 所以群心初始位置會影響結果)
  - 時間複雜度 O(NKT) ， N 是數據數量， K 是群集數量， T 是重複次數

K-means algorithm



- 1. 決定把資料分成k群
- 2. 在二維平面上隨機選取 k 個點
- 3. 對每個資料都計算與這k個cluster中心距離（歐基李德距離Euclidean distance），取比較近的距離，並把它歸為該群
- 4. 接著重新計算 cluster 中心的位置
- 5. 一直重複3、4，直到所有群 cluster 中心沒有太大的變動

http://shabal.in/visuals/kmeans/5.html

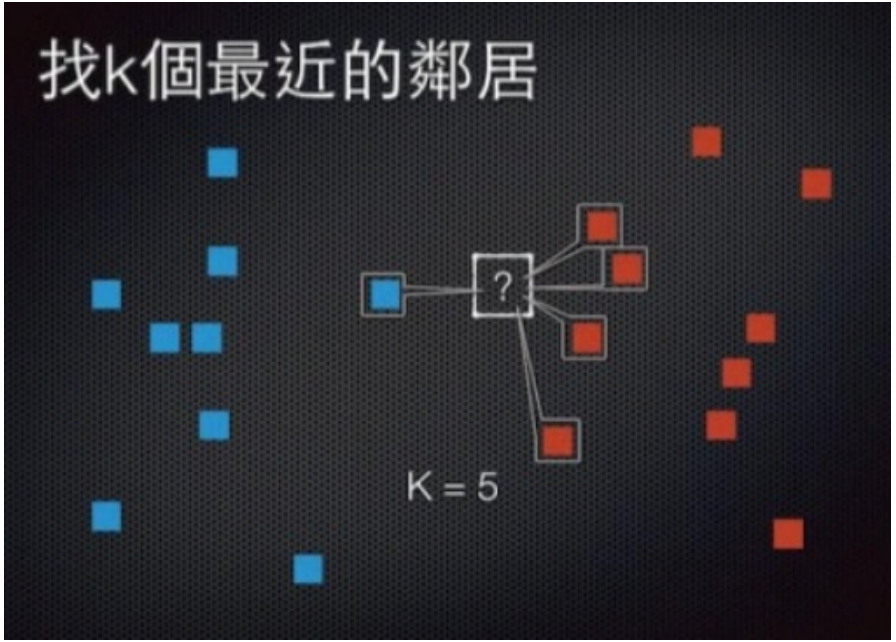
Supervised learning

- kNN algorithm (也是最簡單的NLP算法之一)
  - k-nearest neighbors algorithm
  - 也是物以類聚,
  - **一群已經分好類別的資料** (像是yahoo電影分類)
  - **再加入未分類的資料**
  - 計算資料的"距離", 可能是座標距離或是字詞重疊程度等等 (要自行決定)

k-nearest neighbors algorithm

假設欲預測點是 i

找出離 i最近的 k 筆資料多數是哪一類，預測 i 的類型



How to select k?

How to select k?

About K

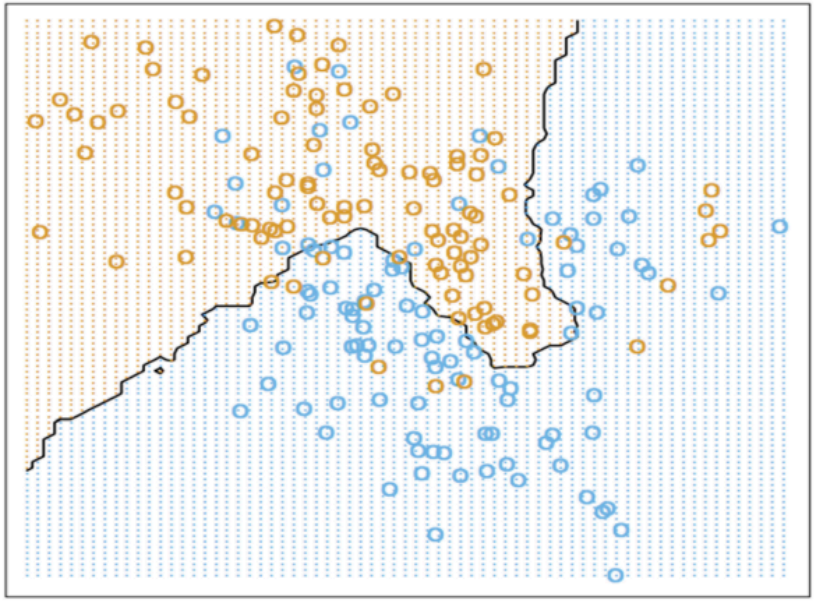
不能是偶數

不能太小 (1)

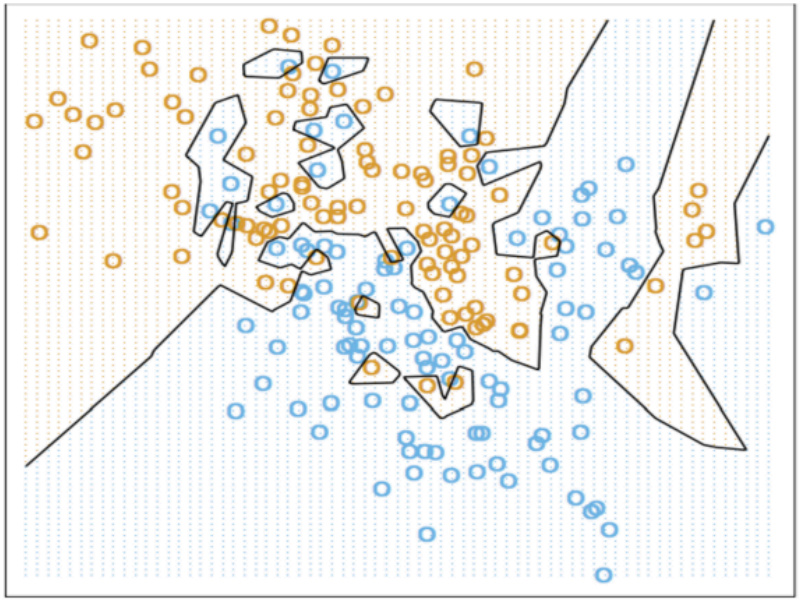
不能太大 (n)

通常 3, 5, 7

How Important is the K in KNN Algorithm



15-nearest neighbors



1-nearest neighbors

underfitting

Overfitting

剛剛好 HW# 2 有分好的電影分類

爬取的資料

1. 動作　2. 冒險

無分類就是 NA

劇情介紹

漫威影業2020年的首部重頭戲，是粉絲請求多時，終於實現的《黑寡婦》個人電影，黑寡婦在漫威電影宇宙中舉足輕重，10多年來參與了7部電影。現在，娜塔莎羅曼諾夫，也就是大家所熟知的黑寡婦，終於踏上個人的旅程，讓粉絲看到她不為人知的一面。

故事時間點設定在《美國隊長 3：英雄內戰》之後，娜塔莎因為幫助了美國隊長而踏上流亡之路，當她發現一個與過去有關的陰謀時，她必須全球追蹤，回頭面對她神祕的間諜生涯，同時逃過反派「模仿大師」的追殺。

HW#3 電影分類實作

1. 已擁有爬取 **6,000** 筆以上電影資訊
   - 1.1 名稱, 分類, 劇情介紹, 上映日期
   - 1.2 分類 **要保留排序**
   - **1.3 有些電影分類為NA,** 不要理他
2. 透過分詞工具, TF-IDF和KNN algo 分類電影
   - 2.1 可以使用現成程式套件 (scikit, etc)
3. 要求 (以第一類別為基準, 忽略其他類別)
   - 3.1 5500+筆已知分類電影, 另500筆為驗證資料
   - 3.2 分類正確數/500 為你的算法精準度 (評分標準)

HW#3 Yahoo! 電影好心分類設計

4. 請 Commit 到 github
5. 請記得要commit >= 5次
6. 不要改github branch name
7. 8/16 12:00 前 push github
8. 遲交或補交 扣20分
9. 一個function 不能超過50行 (超過一行扣一分)

TF-IDF and KNN

a

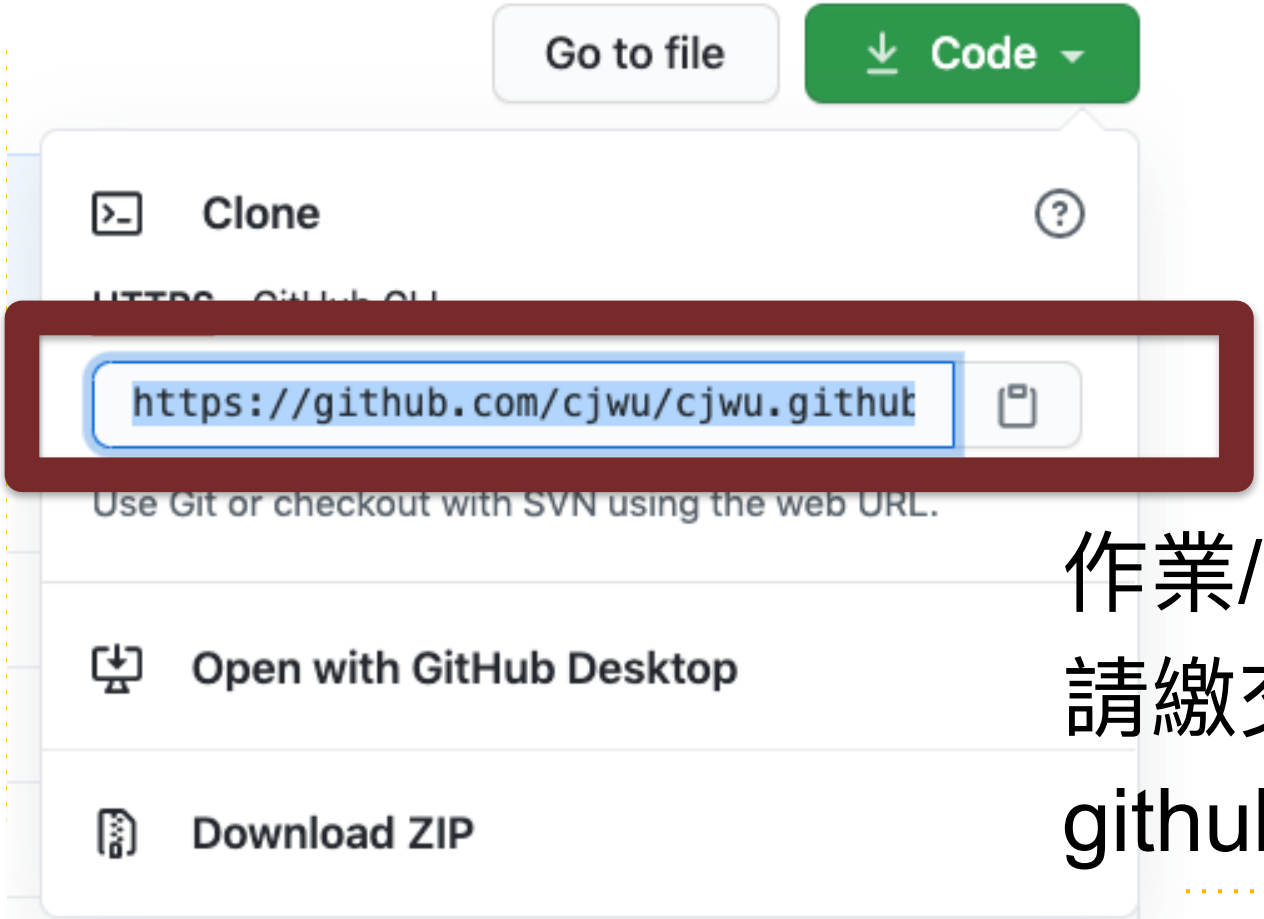
詞彙	文件1	文件2	…	文件 *D*
文字 1	0.48	0.03	…	0.00
文字 2	0.00	0.37	…	0.38
文字 3	0.05	0.08	…	0.22
…	…	…	…	…
文字 *T*	0.12	0.19	…	0.00

假設kNN是算距離

計算 文件x 和 文件y 的距離

d(x, y) =
Sqrt(
  For r = 1 to T
    $(a_{rx} - a_{ry})^2$
)

$$d(x,y) = \sqrt{\sum_{r=1}^{T}(a_{rx} - a_{ry})^2}$$



作業/隨堂練習
請繳交你自己
github上的這個Link

Thanks!

Open for any questions

CJ Wu

cjwu@mail.cgu.edu.tw