多媒體程式設計
文字資料處理

Instructor: 馬豪尚

中文文本資料前處理

- 斷詞
  - N-gram
  - 基於詞典
  - 統計模型(HMM)

中文斷詞

- 常遇到的難題
  - 歧異詞
    - 「我們 / 在野 / 生動 / 物 / 園 / 玩」
    - 「我們 / 在 / 野生 / 動物園 / 玩」
  - 新詞識別
    - 特有名詞(ptt梗、溫拿)
    - 人名地名

中文斷詞

- 正向最大批配法
  - 會有一個詞典，將句子在詞典中由前向後比對，一一比對最長詞的匹配結果
  - 「我們 / 在野 / 生動 / 物 / 園 / 玩」
- 逆向最大批配法
  - 會有一個詞典，將句子在詞典中由後向前比對，一一比對最長詞的匹配結果
  - 「我們 / 在 / 野生 / 動物園 / 玩」
- 正反都做完之後取最長詞
  - 「我們 / 在 / 野生 / 動物園 / 玩」
  - 「我們 / 在野 / 生 / 動物園 / 玩」

中文斷詞

- 基於詞典的斷詞方式，只要詞典中沒有收錄句子中的詞，那可能效果會非常差
- 大部份比較好的斷詞系統都是使用全切分方法，切分出與詞庫匹配的所有可能，然後再運用統計模型決定最好的切分結果

中文停用詞

- 在資訊檢索中，停用詞是從計算的角度來討論，會造成計算上的負擔或是降低搜尋準確度的詞，都可以當作停用詞。
- 在中文裡「停用詞」跟代名詞、助動詞、介系詞、連接詞等和文法、句法有關的「功能詞」相似。
- 是否去除停用詞要取決於處理文本的目的或應用
- 一般來說在中文的自然語言處理不一定需要去除停用詞

Jieba斷詞套件

- Jieba其實是簡體中文版本的中文斷詞系統
- 演算法大概可分成三個部份
  - 第一個部分是建立 Trie DAG 資料結構，快速算出全切分法所有合法的切分組合。
  - 採用了動態規劃查找最大概率路徑, 找出基於詞頻的最大切分組合。
  - 最後一步再使用 HMM 模型計算來辨識新詞。

Jieba斷詞套件

- 全模式 ：把句子中所有的可以成詞的詞語都掃描出來
- 精確模式 ：將句子最精確地切開，適合文本分析
- 搜索引擎模式 ：在精確模式的基礎上，對長詞再次切分，適合用於搜索引擎分詞
- paddle模式 :利用PaddlePaddle深度學習框架，訓練序列標註（雙向GRU）網絡模型實現分詞。同時提供詞性標註功能。目前paddle模式支持jieba v0.40及以上版本

安裝套件 → pip install jieba

paddle模式需安裝

pip install paddlepaddle-tiny==1.6.1

Jieba官方github https://github.com/fxsjy/jieba

Jieba斷詞套件

- 斷詞函數jieba.cut(text, cut_all=True, HMM=False, use_paddle=True)
  - 方法接受四個輸入參數:
    - 需要分詞的字符串
    - cut_all 布林參數用來控制是否採用全模式
    - HMM 布林參數用來控制是否使用 HMM 模型
    - use_paddle 布林參數用來控制是否使用paddle模式下的分詞模式，paddle模式採用延遲加載方式

Jieba斷詞套件-全模式

- seg_list = jieba.cut(text, cut_all=True)
  - 返回的結構都是一個可疊代的 generator，可以使用 for迴圈來獲得分詞後得到的每一個詞語
- seg_list = jieba.lcut(text, cut_all=True)
  - 返回的結構為一個list，可以使用 for 迴圈來獲得分詞後得到的每一個詞語
- Example output: ['我', '們', '在野', '野生', '動', '物', '園', '玩']

Jieba斷詞套件-精準模式

- seg_list = jieba.cut(text, cut_all=False)
  - 返回的結構都是一個可疊代的 generator，可以使用 for迴圈來獲得分詞後得到的每一個詞語
- seg_list = jieba.lcut(text, cut_all=False)
  - 返回的結構為一個list，可以使用 for迴圈來獲得分詞後得到的每一個詞語
- Example output: ['我們', '在', '野生', '動物園', '玩']

Jieba斷詞套件-paddle模式

- seg_list = jieba.cut(text, use_paddle=True)
  - 返回的結構都是一個可疊代的 generator，可以使用 for迴圈來獲得分詞後得到的每一個詞語
- seg_list = jieba.lcut(text, use_paddle=True)
  - 返回的結構為一個list，可以使用 for迴圈來獲得分詞後得到的每一個詞語

Jieba斷詞套件-搜尋引擎模式

- seg_list = jieba.cut_for_search (text)
  - 返回的結構都是一個可疊代的 generator，可以使用 for 循環來獲得分詞後得到的每一個詞語
- seg_list = jieba.lcut_for_search (text)
  - 返回的結構為一個list，可以使用 for 循環來獲得分詞後得到的每一個詞語
- 該方法適合用於搜索引擎構建倒排索引的分詞，會將有可能可以成詞的詞都分出來，粒度比較細

Jieba斷詞套件-詞典

- 載入自定義的詞典
  - jieba.load_userdict(file_name)
- 詞典格式
  - 一個詞佔一行
  - 每一行分三部分：詞語、詞頻（可省略）、詞性（可省略），用空格隔開，順序不可顛倒。
  - file_name 若為路徑或二進制方式打開的文件，則文件必須為 UTF-8 編碼。
  - Example: T恤 28 N

Jieba斷詞套件-詞典操作

- 直接加詞入現有詞典
  - Jieba.addword(word, freq=None, tag=None)
- 刪除詞典內的詞
  - Jieba.del_word(word)
- 修改詞典內詞的頻率
  - Jieba.suggest_freq(segment, tune=True)

練習1

- 用jieba套件建立一個中文文字斷詞器
  - 資料清理
  - 處理繁體中文常用詞典
  - 斷詞
  - 詞頻統計

Jieba-TFIDF關鍵詞

- 載入
  - import jieba.analyse
- 函式
  - jieba.analyse.extract_tags(sentence, topK=20, withWeight=False, allowPOS=())
    - sentence: 待提取的文本
    - topK: 返回幾個TF / IDF權重最大的關鍵詞，默認值為20
    - withWeight: 是否一併返回關鍵詞權重值，默認值為False
    - allowPOS: 僅包括指定詞性的詞，默認值為空，即不篩選

Jieba-TextRank關鍵詞

- 載入
  - import jieba.analyse
- 函式
  - jieba.analyse.textrank (sentence, topK=20, withWeight=False, allowPOS=('ns','n','vn','v'))

Jieba- 詞性標註

- 載入
  - import jieba.posseg
- 函式
  - pseg.cut(sentence)
  - 回傳一個以(word, flag) 二元組為組成的generator ，可以使用 for 迴圈來獲得分詞後得到的每一個詞語和詞性

Jieba Tokenize

- 返回詞語在原文的起止和結束位置
- 函式
  - jieba.tokenize(u'字串')
  - 字串前面要加一個u (因為只能用unicode模式)
  - 開檔編碼一定要是utf-8

Jieba Tokenize 搜尋引擎模式

- 返回搜尋引擎模式詞語在原文的起止和結束位置
- 函式
  - jieba.tokenize(u'字串', mode='search')
  - 字串前面要加一個u (因為只能用unicode模式)
  - 開檔編碼一定要是utf-8

CKIP Tagger套件

- 由台灣中研院資訊所、語言所於民國 75 年成立的中文語言言小組所開發。
- 此套件為一具有新詞辨識能力並附加詞類標記的選擇性功能之中文分詞系統。
- 分詞依據為此一詞彙庫及定量詞、重疊詞等構詞規律及線上辨識的新詞，並解決分詞歧義問題。
- 含有詞類標記，可附加文本中切分詞的詞類解決詞類歧義並猜測新詞之詞類。

CKIP Tagger套件

- 載入模組
  - from ckiptagger import data_utils, construct_dictionary, WS, POS, NER
- 載入模型
  - ws = WS("./data")
  - pos = POS("./data")
  - ner = NER("./data")

CKIP Tagger套件

- 斷詞
  - ws_results = ws([text])
- 詞性標註
  - pos_results = pos(ws_results)
- 命名實體識別
  - ner_results = ner(ws_results, pos_results)

練習2

- 讀取給定文件
  - 建立一個中文文本關鍵詞抽取器
  - 建立一個中文文本詞性標註器

Reference

- https://github.com/fxsjy/jieba jieba官方文件
- https://ckip.iis.sinica.edu.tw/project/ws ckip lab中文詞知識庫小組