DSP Final Project: PLSA

B03902082 資工三 江懿友

1. PLSA

我的 final project 做的題目是閱讀並且實做有關 PLSA 的論文。Probabilistic Latent Semantic Analysis (PLSA) 是一種基於 LSA 的主題分析方法，不同於 LSA 的線性代數解法，PLSA 使用統計與機率模型來找出文件集合 (corpus) 中隱藏的主題 (latent topic)。

我閱讀的論文有 "Probabilistic Latent Semantic Indexing"[1]、"TOPIC-BASED LANGUAGE MODELS USING EM"[2]、"Unsupervised Learning by Probabilistic Latent Semantic Analysis"[3]。第一篇論文主要是解釋 PLSA 的模型跟數學意義，第二篇跟第三篇則提供了較多實做程式碼的細節。

PLSA 的模型需要三組變數 d, z, w 分別代表 document, latent topic, word (term)，還有兩個機率分佈 $P(z|d), P(w|z)$ 分別代表 "given document d, the probability that it is topic z", "given topic z, the probability to generate term w"。另外對 observation 定義變數 $N(d, w)$ 代表 w 在 d 中出現的次數，而我們想做的是讓 P(d, w) 盡量符合 N(d, w)，因此我們把 likelihood function 定義成 $L = \Sigma_{(d, w)} N(d, w) \log P(d, w)$。根據貝氏定理 P(d, w) = P(d)P(w|d)，因為 P(d) 在 training 過程是定值所以我們關心的只有 P(w|d)，而這邊我們把 P(z|d) 當成一種 topic mixture，用他來"混成"我們的 P(w|d) 機率分佈：$P(w|d) = \Sigma_z P(w|z)P(z|d)$。

而在 training 時我們就套用 EM algorithm 調整 P(w|z) 和 P(z|d) 這兩個機率分佈，詳細的調整方法論文上有寫，因為算式很長這邊就不重複寫了。Train 完以後對於在 corpus 裡面的文件我們可以用 P(z|d) 找出它的潛在類別，而對於每個主題我們可以用 P(w|z) 找出跟這個主題相關性高的詞。而對於不在 corpus 內的文件，也就是一個查詢 (query) 我們也可以把它 fold-in 到 model 裡面，fold-in 的方法是固定住除了

[1] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.3584&rep=rep1&type=pdf

[2] http://http.icsi.berkeley.edu/ftp/global/pub/speech/papers/euro99-emlm.pdf

[3] https://pdfs.semanticscholar.org/dc8f/89865ad9c9b6e643abc296ec5000ccdb16ee.pdf

P(z|q) 外的所有參數後，用 EM 調整 model 以得到 fold-in 過後的 P(z|q) 分佈，也就是 query 的主題分佈 (topic mixture)。

LSA 和 PLSA 最大的不同是他的 objective function。LSA 的作法基本上是用 SVD 來找出最大化 variance 的基底，然後投影到那個基底 span 出來的子空間，因此他的 objective function 是"想辦法最大化 variance"，而這個 variance 的計算方法是使用 l2-norm。而 PLSA 則是在找出最大化 likelihood function 的 $P(w|z)$, $P(z|d)$ 分佈，而這個 likelihood function 是在計算 $N(d, w)$, $P(d, w)$ 兩個分佈的吻合程度，也就是 cross entropy 或是 KL-divergence。論文上的說法是說因為這兩種 "投影" 到低維度子空間的方法不同，所以 PLSA 可以得到更好的分析結果。

2. Implementation

我使用 C++ 寫了一個實做 PLSA 分析法的程式，它會對輸入的文件集合做分析然後輸出一個 model。我使用 Reuters-21578[4] 當作我的文件集合，這個 corpus 含有 21578 篇 1987 年路透社 (Reuters) 刊登的文章。

首先我對 corpus 做預處理，把長度過短的文章去掉，並且把 header 等 metadata 去掉只留標題和內文，這樣子做完剩下 19043 篇文章。

接著我把文章內所有除了英文字母以外的符號去掉，並把長度為一的詞去掉（像是 "a", "I"）。然後我用一個現成的 stop word list 過濾詞彙，之後把 document-frequency > 20%（出現在 20% 的文章） 的詞丟掉，因為這些詞出現太多次了可能對文章分類沒有幫助。最後剩下的詞就用來建構每個文章的 bag of words 也就是 $N(d, w)$。在這步驟因為 document-frequency 太高被丟掉的詞有：said, year, mln, march, dlrs, new, april, reuter, lt, inc, pct, corp, one, company；的確都是些常用字。

接著我就隨機初始化 $P(w|z)$, $P(z|d)$ 開始做 EM，然後把 train 好的 model 輸出到檔案裡。而另一個 testing 用的程式則會讀進剛剛 train 好的 model，這個 testing 的程式是互動式的，我就可以打一些詞或是句子 (query)，它就會輸出 fold-in 過後的結果。下面列出一些我得出比較有趣的結果與發現。

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578/

3. Result

以下都是在 Reuters-21578 上分成 128 個主題，跑 100 個 EM iterations 的結果。我試著重現論文上的結果，這個結果是想要說明 PLSA 可以找出同一個詞彙在不同情境 (context) 下會有的不同語意，方法是對於一個詞 w，把所有主題依照 P(w|z) 排序找出 "最有可能產生 w 的主題"，然後輸出這個主題 "最有可能產生的詞" 當作他的 summary。

以 "campaign" 這個詞為例：

Class 91	Class 114	Class 117	Class 111
Party	Details	Plan	Union
Election	Gave	Meeting	strike
Opposition	Maximum	proposal	workers
Majority	Intervention	annual	port
Poll	Release	plans	wage
Parties	Minimum	told	spokesman
Elections	Ecus	shareholders	government
Labour	Market	special	strikes
conservative	Traders	proposed	brazil
government	export	approve	unions

上表是四個最容易產生 "campaign" 這個詞的類別，每個類別下面有十個字代表這個類別最容易產生的字，依照 P(w|z) 由大到小排序。Class 91 告訴我們 campaign 可能代表政黨選舉時的 "選戰"，114 告訴我們可能代表國際貿易時經濟上的競爭，117 則比較像股票交易相關，而 111 是代表了勞資糾紛財團與公會之間的對立。

"Labour"：

Class 91	Class 111
Party	Union
Election	strike
Opposition	workers
Majority	port
Poll	wage
Parties	spokesman

Elections Labour conservative government	government strikes brazil unions

上表是前兩個容易產生 "labour" 的類別。"labour" 代表著英國的政黨工黨 "labour party"，或是代表勞動階級的 "labour"。

"War"：

Class 100	Class 89
Gulf Iran Iranian Attack iraq oil war bahrain iraqi norway	Talks Pact Agreement Trade Japan Japanese Officials Tariffs Last washington

"War" 則會找到國與國之間的侵略戰爭，像是波斯灣戰爭 (gulf war)；或是找到同樣是國與國但是卻不需要動兵的經濟"戰爭" (pact agreement tariff)。

"Reserves"：

Class 19	Class 113	Class 6
Reserve reserves fed funds federal week supply agreements day york	Gas oil energy natural exploration houston production properties barrels reserves	Gold resources mining tons mine ore ounces ltd columbia mines

"Reserves" 則可能代表美國央行 (fed) 的貨幣存底 (reserves funds)，或是石油、金礦等自然資源礦藏。

至於 Fold-in 的例子就比較普通，就是給它一個句子然後它會給我那個句子的主題分佈。

4. Conclusion

整體而言我覺得這個 project 滿有趣的，輸入不同的關鍵字找所屬類別的過程滿好玩的，還會看到一些 80 年代的舊新聞，像是有關台灣的新聞就有不少關於匯率、進出口的新聞報告。

如果需要重現我的執行結果的話，請打下列指令：

```
make
./parse > corpus #產生 corpus
./train [iteration] [number of classes]
./test #開始測試
```

開始執行 ./test 以後，有下面三種指令可以打：

1. "c [class z]"：輸出 class id P(w|z) 前十大的詞。
2. "w [word w]"：輸出最容易產生 word w 的前五個 class。
3. "q [sentence q]"：用 sentence 做 fold-in，然後輸出 P(z|q) 機率最高的前五個 class。