Introduction to Massive Data Analysis Term Project

第 16 組 賴怡惠、楊季蓁

Topic

Pattern Grammar

Motivation

找到pattern grammar可以降低寫作時產生的錯誤，例如文法的錯誤、搭配詞用法的錯誤。

文法錯誤如：

“ We discussed about the issues.”

應為

“ We discussed the issues.”

discuss 經過運算，若沒有discuss about的rule，則我們可知discuss不加 about !

又如

“ He commented the situation.”應為“ He commented on the situation.”, comment 這個詞的 pattern 為 comment on .

這些是我們寫作時經常誤用的情況，故可利用pattern grammar讓使用者知道常見用法。

DataSet

1. raw data

UM-Corpus.en.200k

All that remains for me to do is to say good-bye.

He is not concerned with the difficultied of the factory at all.

after tagging

2. hadoop input

UM-Corpus.en.200k.tagged

record1:

[('all that', 'remains', 'for', 'me', 'to', 'do is', 'to', 'say', 'good-bye', '.'), ('all that', 'remain', 'for', 'me', 'to', 'do be', 'to', 'say', 'good-bye', '.'), ('PDT DT', 'VBZ', 'IN', 'PRP', 'TO', 'VB VBZ', 'TO', 'VB', 'NN', '.'), ('I-NP H-NP', 'H-VP', 'H-PP', 'H-NP', 'H-TO', 'I-VP H-VB', 'H-TO', 'H-VP', 'H-NP', 'O')]

record2:

[('he', 'is not concerned', 'with', 'the difficulties', 'of', 'the factory', 'at all', '.'), ('he', 'be not concern', 'with', 'the difficulty', 'of', 'the factory', 'at all', '.'), ('PRP', 'VBZ RB VBN', 'IN', 'DT NNS', 'IN', 'DT NN', 'IN DT', '.'), ('H-NP', 'I-VP I-VP H-VP', 'H-PP', 'I-NP H-NP', 'H-PP', 'I-NP H-NP', 'I-ADVP H-ADVP', 'O')]

Implement

共運用兩組 mapper 與 reducer 實作

1. 第一組

- mapper :

  將句子切均分給 32 個 reducer。因 mapper 的數量是依據 input 大小，所以我們無法控制 mapper的數量。 而 reducer 的數量我們是可以控制的，因句子做 parse 運算花滿多時間的，所以我們把這部分交給 reducer ，使用 32 個 reducer 。

- reducer ：將所得的句子做 parse 與 tag 運算後 output 如下圖

  1. sentence 轉成 ngram n = 2~9
  2. ngram 轉成 pattern

     由於我們的input已經tag過每個單字的詞性，但所謂的詞性不只有名詞、動詞、形容詞、副詞還有數字、定冠詞之類的。

     我們根據已經寫好的pattern表，如以下所示，再根據input tag轉換成pattern表中的表示法。

     最後再把ngram(for n = 2 ~ 9)轉成pattern表示法的pattern

grammar送進verbpat、nounpat、adjpat測試，若轉換後的pattern存在pattern表內，則顯示此pattern是存在的！故，我們是利用已知的domain knowledge判斷此pattern是否「重要」！只有這些重要pattern才是我們要找的！

```
pgPreps = 'in_favor_of|_|about|after|against|among|as|at|between|behind|by|for|from|in|into|of|on|upon|over|through|to|toward|towards|
otherPreps ='out|down'.split('|')
verbpat = ('V; V n; V ord; V oneself; V adj; V -ing; V to v; V v; V that; V wh; V wh to v; V quote; '+\
           'V so; V not; V as if; V as though; V someway; V together; V as adj; V as to wh; V by amount; '+\
           'V amount; V by -ing; V in favour of n; V in favour of ing; V n in favour of n; V n in favour of ing; V n n; V n adj; V
           'V n wh; V n wh to v; V n quote; V n v-ed; V n someway; V n with together; '+\
           'V n as adj; V n into -ing; V adv; V and v').split('; ')
verbpat += ['V %s n' % prep for prep in pgPreps]+['V n %s n' % prep for prep in verbpat]
verbpat += [pat.replace('V ', 'V-ed ') for pat in verbpat]
nounpat = ('N for n to v; N from n that; N from n to v; N from n for n; N in favor of; N in favour of; '+\
           'N of amount; N of n as n; N of n to n; N of n with n; N on n for n; N on n to v'+\
           'N that; N to v; N to n that; N to n to v; N with n for n; N with n that; N with n to v').split('; ')
nounpat += nounpat + ['N %s -ing' % prep for prep in pgPreps ]
nounpat += nounpat + ['ADJ %s n' % prep for prep in pgPreps if prep != 'of']+ ['N %s -ing' % prep for prep in pgPreps]
adjpat = ('ADJ adj; ADJ and adj; ADJ as to wh; '+\
          'ADJ enough; ADJ enough for n; ADJ enough for n to v; ADJ enough n; '+\
          'ADJ enough n for n; ADJ enough n for n to v; ADJ enough n that; ADJ enough to v; '+\
          'ADJ for n to v; ADJ from n to n; ADJ in color; ADJ -ing; '+\
          'ADJ in n as n; ADJ in n from n; ADJ in n to n; ADJ in n with n; ADJ in n as n; ADJ n for n'+\
          'ADJ n to v; ADJ on n for n; ADJ on n to v; ADJ that; ADJ to v; ADJ to n for n; ADJ n for -ing'+\
          'ADJ wh; ADJ on n for n; ADJ on n to v; ADJ that; ADJ to v; ADJ to n for n; ADJ n for -ing').split('; ')
adjpat += [ 'ADJ %s n'%prep for prep in pgPreps ]
pgPatterns = verbpat + adjpat + nounpat
reservedWords = 'how wh; who wh; what wh; when wh; someway someway; together together; enoguh enough; amount amount; that that'.split(
pronOBJ = ['me', 'us', 'you', 'him', 'them','her']
```

output

同一句話，因為會extract ngram(for n = 1~9)，所以同一個片段可能出現很多次，如BE的example只差一個vocabulary。

```
BE      V to v  has been to look
BE      V to v  has been to look on
LOOK_ON V       look on
LOOK_ON V with n        look on with folded arms
SIT     V adv   sit around
```

2. 第二組

由以上mapper、reducer我們已經得到每個head word的「重要pattern」，但是無法保證raw data的用法正確，也無法保證tagging詞性的系統正確性！所以找到重要的片段，不代表這些重要片段對於不同pattern都是常見片段，因為pattern表列出的pattern是所有V、N、ADJ會出現的pattern，此階段我們要找出根據某個head word，屬於它的特定rule!

- mapper ：

將上圖的 out ，根據head word字首hashing給26個reducer做各個英文字母的排序和過濾example pattern。

不在mapper做太多運算，只把句子分給reducer是因為我們不能控制mapper數量，可是卻能控制reducer數量，所以大部分的運算於reducer完成，以達到加速。

- reducer :

濾掉不常見的 pattern 和 一些無意義的單詞 ，結果如下。

無意義的像是有'but', 'could', 'can', 'should', 'and', 'might', 'would', 'may', 'will', 'not', 'also', 'or' … ，這些連接詞與助動詞，我們覺得寫出來的例句滿冗長，因此把它濾掉。

另外，我們檢查沒有亂碼的 examples ，才保留，但亂碼的head word其實也可以視為不常見的rules，所以只需要設一個threshold，若連pattern examples的數量都不超過這個threshold，即會濾掉亂碼。

濾掉不常見的 pattern ，我們用 counter 來計算我們在 dataset 中 most common 的 patter 留下。

output

```
AGE     N of amount     the age of 18 | the age of 16 | the age of 12
AGREE   V n     agree that balance | agree that | agreed last year
AGREE   V to n  agree to anything | agree to something | agreed to a looser clause
AGREE   V to v  agreed to take | agreed to go | agreed to let
AIM     V at n  aimed at tweens | aims at the deliverance | still aimed at good humour
AIM     V to v  aimed to reclaim | aims to explore | aim to bring
AIR     N to v  the air to rest | air to enter | the air to escape
ALIGN   V n     align themselves | aligning ball bearing | aligning ball
ALLEVIATE       V n     alleviate pain | alleviate that pain | alleviates dryness
ALLOCATE        V n     allocate resources | allocates bandwidth | allocating resources
ALLOW   V for n allowed for capital formation | allow for the access | allow for these
different classes
ALLOW   V n     allows users | allow users | allow yourself
ALLOW   V n to v        allow the user to enter | allows men to work out | allows men to
work
ALLOW   V to v  was allowed to sit | are allowed to live | was allowed to go
```

其中我們發現若pattern example含dream，都會讓片段變美！

```
[y32jjc00@hcgwc112 final]$ grep 'dream' result
DARE    V to v  daring to step | dared to be | dare to dream
FEED    V n     feed the hungry dream | feeds hundreds | feeding material part
PURSUE  V n     pursue knowledge | pursuing a girl | pursue realistic dreams
```