Programming for Social Scientists: Text Mining III

黃瀚萱

Last Week Update

```
pair_mean_distances = Counter()



for (w1, w2, distance), c in word_pair_distance_counts.most_common():

    pair_mean_distances[(w1, w2)] += distance * (c / word_pair_counts[(w1, w2)])
```

```
for (w1, w2), distance in pair_mean_distances.most_common(20):

    print("%s\t%s\t%f\t%d" % (w1, w2, distance, word_pair_counts[(w1, w2)]))
```

$$\bar{d} = \frac{\sum_{i=1}^{n} d \times c_d}{n} = \sum_{i=1}^{n} \frac{d \times c_d}{n}$$

Agenda

- Part of Speech
- POS Tagging
- Counting
- POS-Aware Collocation Mining
- Chinese Processing

Part of Speech

- Classification of words into a number of syntactic or grammatical categories.
    - Noun: People, animals, concepts, and things.
    - Verb: Action
    - Adjective: Properties of Nouns

The dog eats the big hotdog.

DT NN VBZ DT JJ NN .

Noun and Pronouns

- Refer to entities in the world.
    - People, animals, and things.
- In English, only one inflection of the noun
    - Plural form vs Singular form
- Gender inflection in the third person singular pronoun
    - he, she, and it

Determiners

- Describing the particular reference of a noun.
  - Articles
    - a/an: Indicates the person/thing was not previously mentioned.

A boat on the sea with clouds. A fisher stands checking her equipments.

    - the: Already made reference to the noun, or if the reference is clear from context.

A boat on the sea with clouds. The fisher stands checking her equipments.

Verbs

- Verbs are used to describe actions.
- Usually the most important word in a sentence.

Form	Regular	Irregular
root / base	walk	write
Third singular present	walks	writes
Gerund / present particle	walking	writing
Past tesne	walked	wrote
past/passive particle	walked	written

Adverbs

- Adverbs modify a verb in the same way as the adjectives modify nouns.
    - Place: The guide finds a restaurant locally
    - Time: She often visits to hospital.
    - Manner: He grabbed her roughly.
    - Degree: I completely forgot that it’s his birthday today.

Others

- Prepositions are mainly small words that prototypically express spatial relationships.
  - to, of, on, in…
- Conjunctions
  - Coordinating conjunctions: and, but, or, …
  - Subordinating conjunctions: because, although, that…

https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Fundamental Chinese Processing

- Sentence segmentation
- Simplified / Traditional conversion
- Word segmentation
- POS tagging

Simplified / Traditional Conversion

标准间太差 房间还不如3星的 而且设施非常陈旧.建议酒店把老的标准间从新改善.

標準間太差 房間還不如3星的 而且設施非常陳舊.建議酒店把老的標準間從新改善.

One to Many Ambiguous

美发现号航天飞机，
上头发奖金；
头发应该剪了，
后天，皇后

美發現號航天飛機，
上頭發獎金；
頭髮應該剪了，
後天，皇后

Conversion with HanziConv

```
>>> from hanziconv import HanziConv

>>> HanziConv.toSimplified('後天')

'后天'

>>> HanziConv.toTraditional('后天')

'後天'

>>> HanziConv.toTraditional('桌子上面')

'桌子上麵'

>>> HanziConv.toSimplified('乾燥')

'干燥'

>>> HanziConv.toTraditional('干燥')

'幹燥'
```

Conversion with OpenCC

```
>>> import opencc
>>> opencc.convert('後天', config='zht2zhs.ini')
'后天'
>>> opencc.convert('后天', config='zhs2zht.ini')
'後天'
>>> opencc.convert('桌子上面', config='zhs2zht.ini')
'桌子上面'
>>> opencc.convert('上了一碗面', config='zhs2zht.ini')
'上了一碗麪'
>>> opencc.convert('乾燥', config='zht2zhs.ini')
'干燥'
>>> opencc.convert('干燥', config='zhs2zht.ini')
'乾燥'
```

Chinese Word Segmentation

标准间太差，房间还不如3星的，而且设施非常陈旧。建议酒店把老的标准间从新改善。

标准间 / 太差 / ， / 房间 / 还 / 不如 / 3 / 星的 / ， / 而且 / 设施 / 非常 / 陈旧 / 。 / 建议 / 酒店 / 把 / 老 / 的 / 标准间 / 从新 / 改善 。

Ambiguity in Chinese Word Segmentation

- Word (詞) is a basic semantic unit in language processing.
- Unlike English, there is no explicit delimiter between Chinese words.
- The segmentation is ambiguous.
  - 日 文章 魚 怎麼 說
  - 日文 章魚 怎麼 說
- CWS attempts to find the most possible segmentation given a piece of Chinese text.

Machine Learning based CWS

- Training a machine learning model based on labeled data.
  - Labeled data: a corpus in which words are separated with a delimiter.
- To add or not to add a delimiter between a successive character pair, depending on its context.
  - Sequence modeling
- Most training data is in the general domain.

Sequence Lebeling Approach

- Label each Chinese character in the sentence with one of four labels
  - L: left boundary of a word
  - R: right boundary of a word
  - M: middle character in a word
  - S: a word with single character
- Sequence labeling model is trained to perform the labeling
  - HMM、CRF

日	文	章	魚	怎	麼	說
L	R	L	R	L	R	S

Stanford CoreNLP

- A more powerful toolkit for multiple language processing, including Chinese.
  - Sentence segmentation
  - Chinese word segmentation
  - Part-of-speech tagging
  - Syntactic parsing

https://stanfordnlp.github.io/CoreNLP/
Demo: https://corenlp.run/

Assignments

- Using mutual information to mine the collocations from “../text_mining_2/corpus.txt” with following POS patterns:
  - N / N
  - NNP / N
  - J / N
- Using mutual information to mine the Chinese collocations from blessing.txt.