Text Mining and Data Viz

2018-05-12

leoluyi@iii

Slides http://pcse.pw/6WHWJ

關於我

- 呂奕 Leo Lu
- 台大工管
- 目前於金融業服務
- Build data products
  - ETL
  - Models
  - Text mining
  - Viz
  - ...

Text Minning

流程 與 工具們

舊時代的工具

vs.

新世代的工具

以前我們都用外國人寫的東西

tm + tmcn
Rwordseg

流程

Get data → Tokenize → Embedding → Viz → Model

Get data

PTT 是宅宅的好朋友

每天都有很多很多的~~廢文~~語料

自己的爬蟲自己寫

```
devtools::install_packages(
    "leoluyi/PTTr")
```

Cleaning and preprocessing text

留下資訊，去掉雜訊

Tokenize

Transform whole text into parts

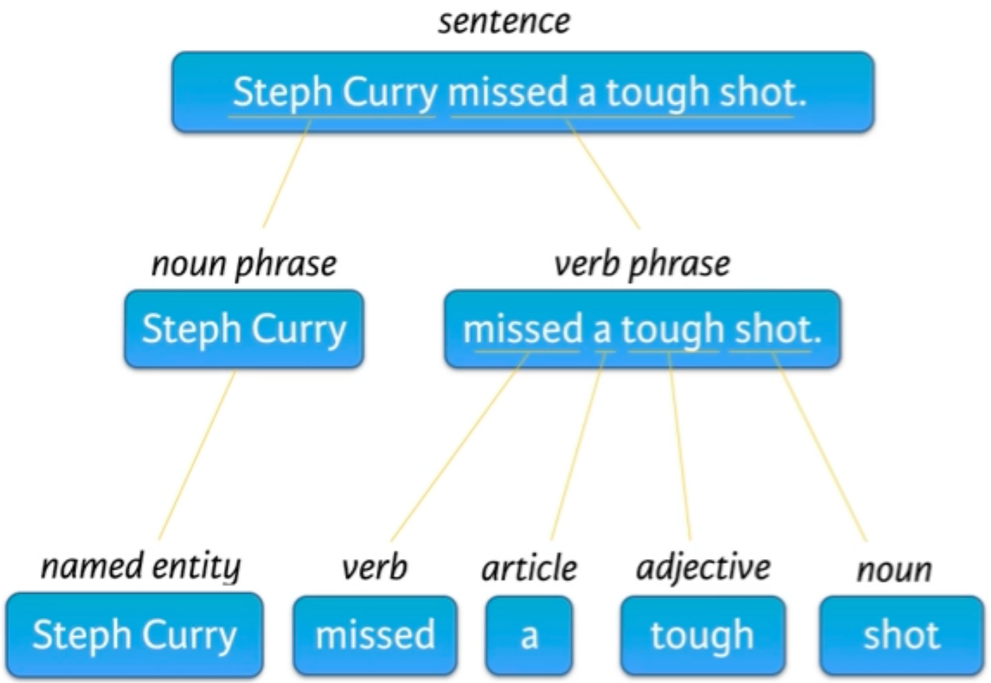
For English

- normalization
  - stemming (詞幹提取)
  - lemmatization (詞型還原)
- POS tagging
- ...

中文似乎比較簡單

- 斷詞
- 不斷詞
- **POS tagging**
- **...**

Semantic Parsing vs. Bag-of-Words



R tools

- stringr
- jiebaR

Embedding

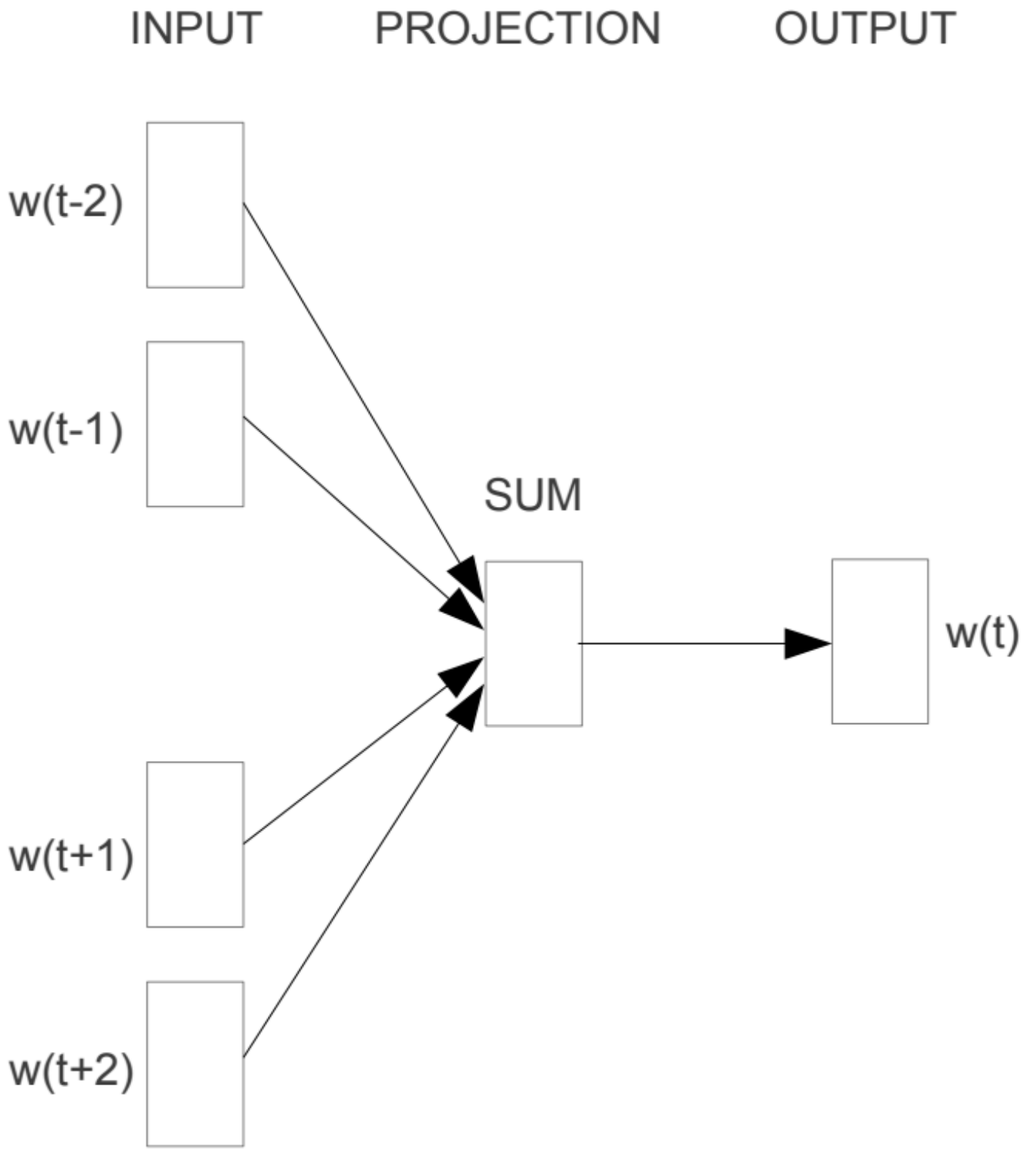
(Encode, Feature Extraction)

Embedding

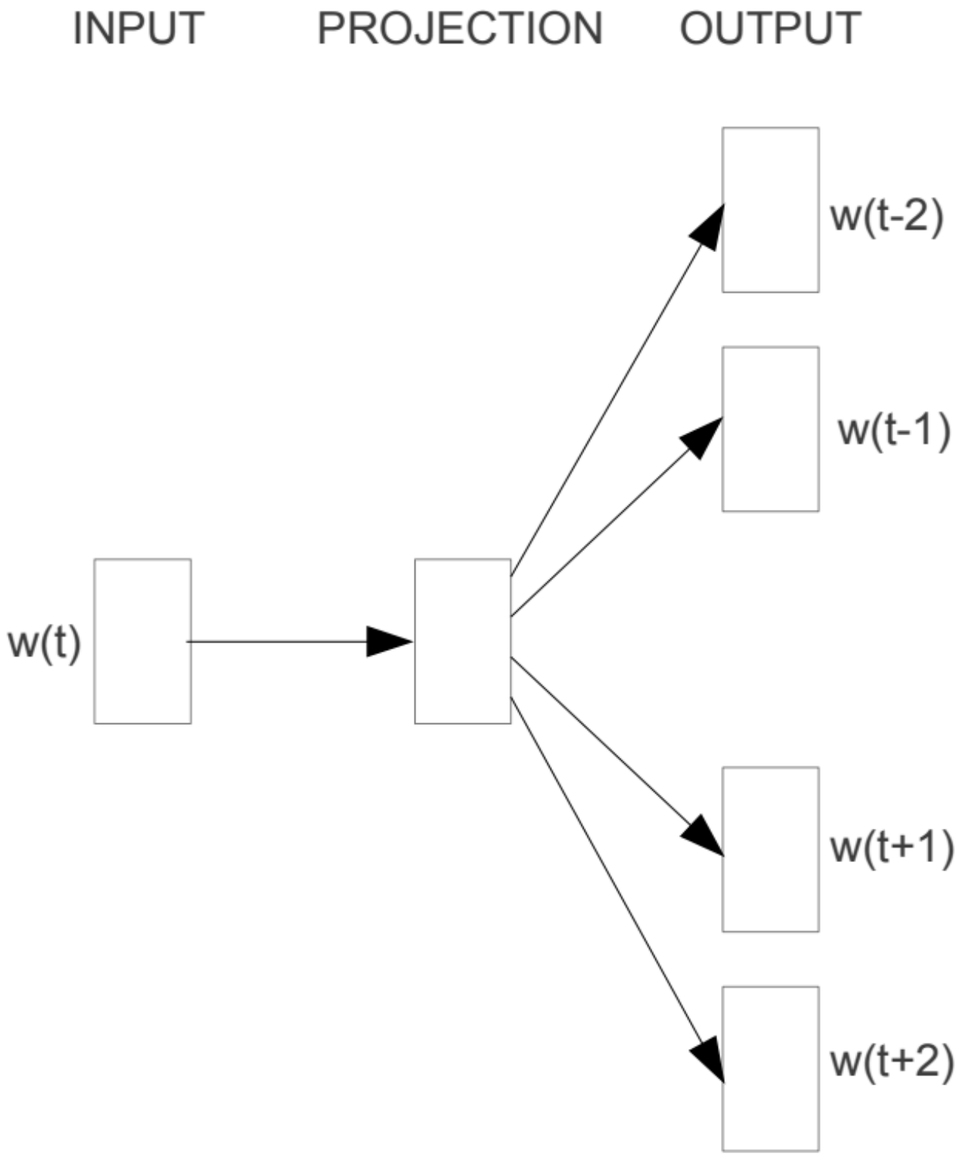
In a nutshell, Word Embedding turns *text* into *numbers*.

- Embedding Layer[1]
- Word2Vec
- GloVe
- doc2vec
- sense2vec

[1] https://machinelearningmastery.com/what-are-word-embeddings/



CBOW



Skip-gram

Demo

Information Retrieval

Visualize

- Dimension Reduction
  - t-sne
  - PCA
- Clustering
- Interactive or static plots

Visualize

- `tsne::tsne()`
- `prcomp()`

Model

Tasks

- Classification
  - 文本分類
- Clustering
  - 找尋相似文本
- Generative models
  - 文本自動生成

用到最後都會想要寫自己的 toolkit

- Sparse Matrix manipulation
- Informaiton retrieval tools
- ...

Summary

1. Problem definition & specific goal: Get Curious About Text
2. Finding Your Data
3. Preprocessing Your Data
   - → Removing stopwords, Stemming, Segmentation, ...
4. Feature Extraction
   - → Document-Term Matrix: `tm`, `text2vec`
   - → Named Entity Recognition, POS tagging
   - → Word embeddings: word2vec, GloVe
5. More Text Mining Skills
   - → sentiment analysis
   - → `topicmodels`, `LDAViz`: LDA
6. More Than Words - Visualizing Your Results

數據科學

呂奕 leoluyi@github

https://leoluyi.github.io