

Sentiment analysis using product review data

Beilei Bian

2017/07/01

NLP入门

- 什么是自然语言处理?

自然语言处理的目标：设计算法使计算机能够理解自然语言并执行一些任务。

- Easy

- Spell Checking
- Keyword Search(tf-idf)
- Finding Synonyms

- Medium

- Parsing information from websites, documents, etc.

- Hard

- Machine Translation
- Semantic Analysis
- Coreference(共指解析)
- Question Answering

词向量

The first and arguably most important common denominator across all NLP tasks is how we represent words as input to any and all of our models.

- **one-hot vector:** 为词库中的词建立索引，每个词都表示为一个 $\mathbb{R}^{|V| \times 1}$ 向量，其中 $|V|$ 为词库中词的个数。

$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^a = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{at} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots w^{zebra} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

$$(w^{hotel})^T w^{motel} = (w^{hotel})^T w^{cat} = 0$$

one-hot存在的问题：1.词与词是正交的，无法度量相似度 2.高维且稀疏

词向量

- 基于特征值分解的方法(SVD)

we first loop over a massive dataset and accumulate word co-occurrence counts in some form of a matrix X , and then perform Singular Value Decomposition on X to get a USV^T decomposition.

Use the rows of U as the word embeddings

1. Word-Document Matrix

$\mathbb{R}^{|V| \times M}$ 的矩阵 (large matrix)

缺点：文档多，词的数量多，矩阵过大，存储困难

2. Window based Co-occurrence Matrix

缺点：1.矩阵的维数经常变化 2.矩阵过于稀疏，因为许多词和词不会共同出现 3.高维

1. I enjoy flying.
2. I like NLP.
3. I like deep learning.

The resulting counts matrix will then be:

$$X = \begin{matrix} & \begin{matrix} I & like & enjoy & deep & learning & NLP & flying & . \end{matrix} \\ \begin{matrix} I \\ like \\ enjoy \\ deep \\ learning \\ NLP \\ flying \\ . \end{matrix} & \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \end{matrix}$$

词向量

- Iteration Based Methods

- 统计语言模型：一个词序列的概率

比如： "The cat jumped over the puddle."

一元模型：

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^n P(w_i)$$

但是，下一个词的出现是强依赖于上一个词的。

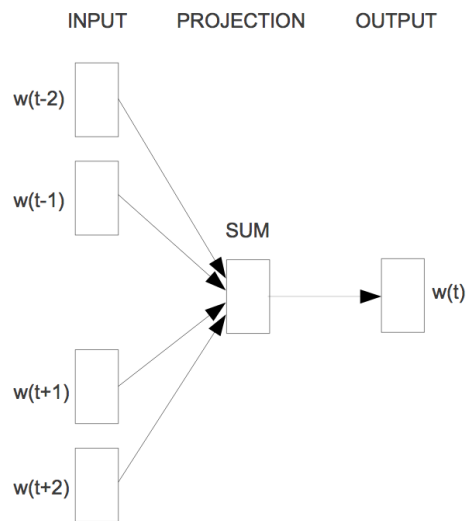
二元模型：

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^n P(w_i | w_{i-1})$$

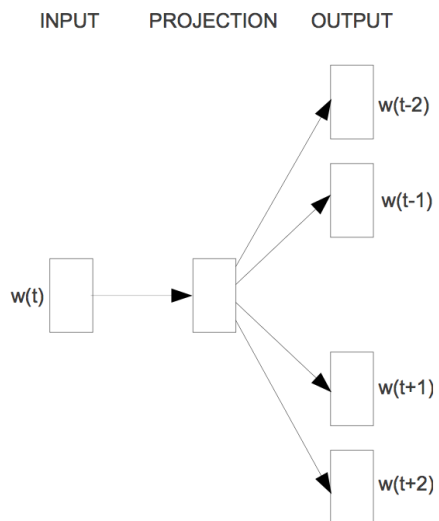
实际应用中最多的的是N=3的情况

词向量

- word2vec
 - CBOW
 - Skip-gram



CBOW



Skip-gram

三层的神经网络：输入层，投影层，输出层

涉及到的数学原理：统计语言模型、二叉树（Huffman树）、Hierarchical softmax、negative sampling、梯度下降法

词向量是帮助构造目标函数的辅助参数，训练完成后，它是统计语言模型的一个副产物。

Data Analysis

以bestbuy电子产品评论为例进行情感分析

Part 1

数据爬取：以Mac为例

Customer Image Gallery

Page 1 of 11

+ Filter Reviews by: Rating, Verified Purchase

Sort by: **Most Helpful**

Search

momdawson1

★ ★ ★ ★ ★ 1

laptop

October 15, 2016

My Laptop is wonderful now, I own 2 other Mac laptops which I bought from apple and never had a problem. I know things happen and thats why I wasn't mad when we opened the laptop the next day and we had all popups and something saying something about a virus, I called and they told me I could return it or have geek squad fix so we took it back but we forgot the cord so when my son brought it back the following day had an appointment with the geek squad and they were charging me \$199 to fix it (a brand new laptop) my son called me and I said then we will just return it, they put a manager on who was very rude told me that we couldn't return it, he said we did something to the computer that we don't know how to use it, he was rude and ignorant said the only thing he could do is sell me the geek squad for a year for \$89 I had no other choice because my son needed it for school so on top of everything else I had to buy the geek squad which I already had the apple plan which he told me wouldn't cover it, I wrote many reviews and a letter to best buy with no response except one that said sorry!! I think its a disgrace and ill never go to best buy again, i think i should have my laptop repaired for free being it was brand new.

No, I would not recommend this to a friend.

Helpful (192) Unhelpful (100) Report

Read Comments (5) Post Comment

Elements Console Sources Network Performance

Filter: All XHR CSS Img Media Font Doc WS Manifest Other

500 ms 1000 ms 1500 ms 2000 ms 2500 ms 3000 ms 3500 ms 4000 ms 4500 ms 5000 ms

Name: reviews.djs?format=em...

Headers: Preview Response Timing

General

Request URL: https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true

Request Method: GET

Status Code: 200 OK

Remote Address: 223.119.236.79:443

Referrer Policy: no-referrer-when-downgrade

Response Headers

Cache-Control: no-cache, no-store

Connection: keep-alive

Content-Encoding: gzip

Content-Language: en-US

Content-Length: 42584

Content-Type: text/javascript; charset=UTF-8

Date: Fri, 30 Jun 2017 06:57:32 GMT

Expires: Thu, 01 Jan 1970 00:00:00 GMT

P3P: CP="Bazaarvoice does not have a P3P policy."

Server: nginx

Vary: Accept-Encoding

X-Bazaarvoice-Region: eu-west-1

X-Content-Type-Options: nosniff

Request Headers

Accept: */*

Accept-Encoding: gzip, deflate, br

1 / 12 requests | 42.0 KB / 7...

Console What's New

Highlights from Chrome 59 update

CSS and JS code coverage

Find unused CSS and JS with the new Coverage drawer.

Full-page screenshots

Take a screenshot of the entire page, from the top of the viewport to the bottom.

Block requests

URL	Type	Total Bytes	Uncompressed
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	360,000	230
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	214,482	171
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	231,281	150
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	145,862	107
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	128,754	104
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	148,513	90
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	122,003	81
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	88,980	40
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	23,947	21
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	21,440	20
https://bestbuy.ugc.bazaarvoice.com/3545w/6443834/reviews.djs?format=embeddedhtml&page=2&scrollToTop=true	JS	44,444	2

找到资源所在的地址：

Request URL:<https://bestbuy.ugc.bazaarvoice.com/3545w/6443034/reviews.djs?format=embeddedhtml&page=1&scrollToTop=true>

找出URL的规律，对整个页面进行下载并解析。

用到的工具：

- urllib
- json
- lxml.html
- CSSSelector
- BeautifulSoup

Part 2

数据清洗

- regular expression

```
import re
review_letters = re.sub("[^a-zA-Z]", " ", review_text)
```

- 分词 / 字母大小写问题

```
words = review_letters.lower().split()
```

- nltk去停顿词

```
from nltk.corpus import stopwords
stops = set(stopwords.words("english"))
words = [w for w in words if not w in stops]
```

Part 3

建立特征

- 方法一: Bag of words (工具: sklearn)

词袋法: 不考虑语法, 不考虑词的顺序, 只考虑词的多样性。常用于文档分类。

每个词出现的频率作为特征输入分类器。(也可以用于图像分类: Bag of words for computer vision <https://github.com/bikz05/bag-of-words>)

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(analyzer = "word", tokenizer = None, preprocessor = None)
train_data_features = vectorizer.fit_transform(clean_train_reviews)
train_data_features = train_data_features.toarray()
```

- 方法二: word2vec (工具: gensim)

```
from gensim.models import word2vec
tokenizer = nltk.data.load('tokenizers/punkt/english.pickle')
all_sentences = []
for sum in data["review"]:
    all_sentences += sum_to_sent(sum, tokenizer)
model = word2vec.Word2Vec(all_sentences, size=300, min_count = 1, window=5)
model.save(model_name)
```

Part 4

建立分类模型(sklearn)

- Random Forest

```
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(n_estimators = 100)
forest = forest.fit(trainDataVecs, train["star"])
result = forest.predict(testDataVecs)
```

- Naive Bayes
- SVM

Part 5

模型评估

将数据集随机分为训练集用来建模，测试集用来测试。

评估方法：

- Accuracy
- TPR/TNR
- ROC/AUC

```
from sklearn import metrics  
print(metrics.classification_report(test['star'], result))
```

Deep learning

- CNN CNN用于文本分类的开山之作：Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

CNN最初用于图像分类，在自然语言处理中，可以把word embedding看作image

- RNN RNN引入了定向循环，可以处理序列数据，网络对前面的数据进行记忆并应用于当前输出的计算中。应用最广泛的为：LSTM

一些资料

关于word2vec: <http://blog.csdn.net/itplus/article/details/37969635>

NLP tutorial: <http://cs224d.stanford.edu/>

数学之美

常用工具:

- NLTK
- gensim
- Stanford NLP
- openNLP