一、NLP 常见任务

1. 自动摘要

抽取式：关键词提取，常用的（TF-IDF，TextRank，Word2Vec，主题模型等）

生成式：主要依靠神经网络实现，也叫作 encoder-decoder，encoder 将原文本编码成向量，然后 decoder 从该向量提取信息，然后生成文本

2. 指代消解 ： 小明放学了，妈妈去接他

3. 机器翻译

4. 主题识别

5. 文本分类

典型的就是新闻文本分类，里面主要包括分词，去停用词，构建词向量，模型的建立

6. 情感分析

7. 问答系统（query）

二、NLP 常用的处理方式：

1. 词性标注

2. 分词

常用的分词方法

a HMM:能很好处理歧义和未登陆词，单需要大量人工标注，分词速度慢 ICTCLAS

b CRF 分词按照字的词位分类，通常定义字的词位信息如下（B M E S）

c 基于词典的正向最大匹配法:分词快，采用 TRIE 索引树，哈希索引，百度

通常，jieba，gensim，nltk 里面集成了分词，词性标注等功能

3. 实体识别（命名实体识别）

比如医疗领域电子病历的识别，主要任务就是要是识别出某人某个位置某个疾病以及表现。

4. 句法分析

三、举例美团点评的 NLP 框架

四、NLP 中的词

离散的表示字或词：

John likes to watch movies. Mary likes too.

John also likes to watch football games.

词典：

{'John':1,'likes':2,'to':3,'watch':4,'movies':5,'also':6,'football':7.'games':8,'Marry':9,'too':10}

1.One-hot 表示：

John:[1,0,0,0,0,0,0,0,0,0]

…

2.Bag of Words

John likes to watch movies. Mary likes too.[1,2,1,1,1,0,0,0,1,1]

John also likes to watch football games. [1,1,1,1,0,1,1,1,0,0]

one-hot 和 Bag of Words 均没有考虑上下文信息。

3.Bi-gram 和 N-gram（语言模型）

2-gram：

John likes :1

likes to:2

to watch:3

watch movies:4

Mary likes:5

likes too:6

John also:7

also likes:8

watch football:9

football games:10

所以上面两句话表示成：

[1,1,1,1,1,1,0,0,0,0,]

[0,1,1,0,0,0,1,1,1,1]

补充语言模型：

Unigram/1-gram:P(Mary likes too) = P(too|Mary,likes)*P(like|Mary)*P(Mary)

=P(too)*P(likes)*P(Mary)

Bi-gram/2-gram:P(Mary likes too) = P(too|Mary,likes)*P(like|Mary)*P(Mary)

=P(too|likes)*P(likes|Marry)*P(Mary)

离散表示词向量的问题：

1. 词表维度随着语料库增长膨胀
2. 数据稀疏的问题：不好捕捉文本的含义

分布式表示：一个词用附近的其他词表示

1. 通过共现矩阵做 SVD 降维之后的低纬度向量表示

共现矩阵：

a. I like deep learning.
b. I like reading book.
c. I enjoy eating.

counts	I	like	enjoy	deep	learning	reading	book	eating	.
I	0	2	1	0	0	0	0	0	0
like	2	0	0	1	0	1	0	0	0
enjoy	1	0	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0	0
learning	0	0	0	1	0	0	0	0	1
reading	0	0	0	0	0	0	1	0	0
book	0	0	0	0	0	1	0	0	1
eating	0	0	1	0	0	0	0	0	1
.	0	0	0	0	1	0	1	1	0

将共现矩阵作为词向量的缺点：

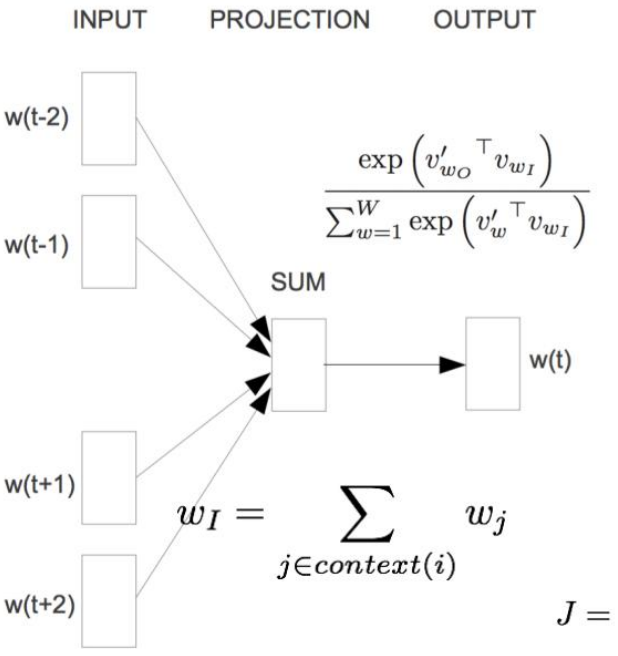
A. 向量维度随着词典大小线性增长

B. 某些情况下还会存在矩阵稀疏的问题

所以获得共现矩阵之后一般会做降维，但降维之后会难以对新词分配词向量。

2.词向量 word2vec:

CBOW（连续词袋）：



Skip-Gram 模型：