# **Report of Deep Learning for Natural Language Processing**

曹思远

[3203998114@qq.com](mailto:3203998114@qq.com)

## **Abstract**

本次作业是基于中文语料库，利用Word2Vec神经语言模型，将经过jieba分词后的中文词语进行训练得到词向量。后经过四种方法来验证向量的有效性。1.计算词向量之间的语义距离（如：郭靖-黄蓉，杨过-小龙女）；2. 测试相关词对的有效性（如：杨过-小龙女=黄蓉-？）；3.利用K-means来验证某一类词语的聚类效果（如：各部小说中的主人公）；4.两个相关段落之间的直接语义关联。

# **Introduction**

作业要求：利用给定语料库（金庸小说语料库），基于 1～2 种神经语言模型（如：Word2Vec，LSTM，GloVe 等模型）来训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联、或者其他方法来验证词向量的有效性。

# **Methodology**

为了验证根据中文语料库得到的词向量的有效性，采用gesim库提供的word2vec来得到词向量，后通过四种方法来验证词向量之间的有效性。

1.通过计算相关词对之间的语意距离（比如：郭靖-黄蓉，杨过-小龙女），观察得到的余弦量大小来判断相关性的好坏。

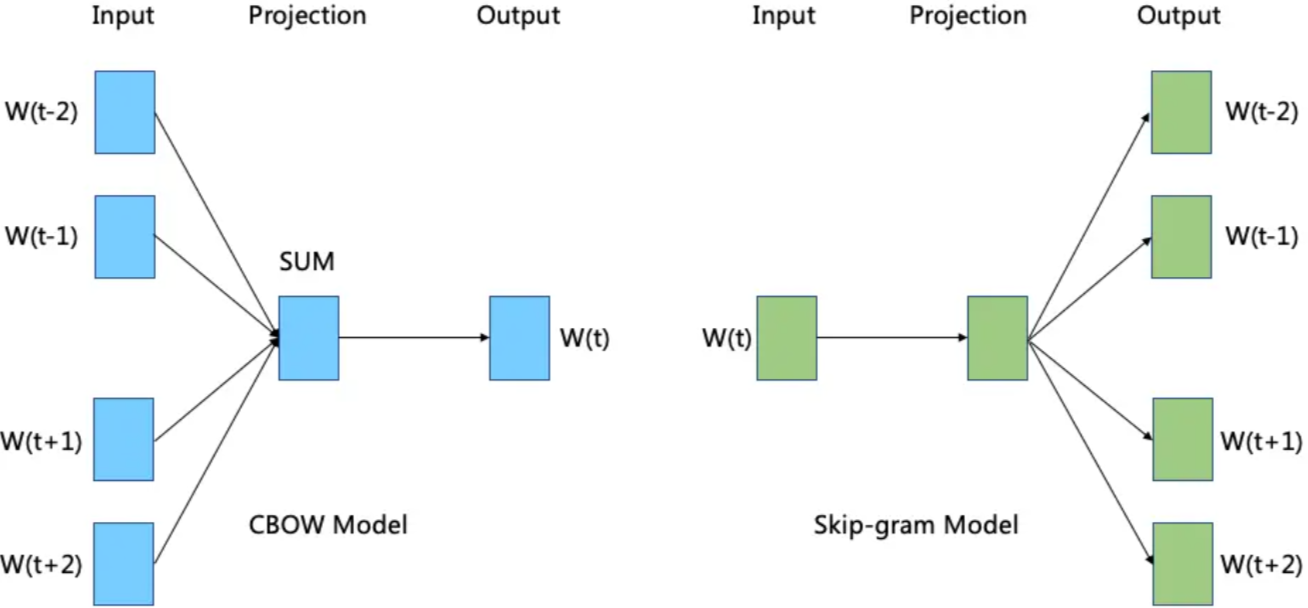
2.做类似国王-男人=皇后-女人词类比测试，如：杨过-小龙女=黄蓉-“郭靖？”，来验证词向量的有效性。

3.验证某一类词语的聚类效果，选择各部小说中的主要人物，再通过K-means进行聚类，后通过TSNE将高维词向量降维到2维进行可视化，观察各人物的聚类情况来验证词向量的有效性。

4.提取同一小说中两个相似段落，并计算词向量的余弦相似度来验证词向量的有效性

## **M1:word2vec模型介绍**

Word2Vec是由 Google于2013年发布的一种用于自然语言处理的模型。它是一种分布式表示学习技术，可以将词语转换为低维稠密向量，这些向量可以捕捉词语之间的语义关系。Word2Ve模型有两种主要的架构：Skip-Gram和Continuous Bag of Words (CBOW)。Word2Vec 的核心思想是利用词语的上下文信息来学习词语的向量表示。模型会通过大量的文本数据，基于词语在句子中的出现模式，来训练一个神经网络，从而学习到每个词语的向量表示。如下图是这两个模型的示意图。



其具备训练速度块，效果好的优点以及能捕捉到词语之间的与语义关系。但是有无法处理登录词和对于长短语义依赖的捕捉能力有限的缺点。

## **M2:实施过程**

（1）数据提取：将金庸小说中文语料库进行提取，根据stopwords去除停词并利用jieba库进行分词。

（2）word2vec模型训练：利用gensim提供的word2vec进行训练，参数设定vector\_size=100，window=5， min\_count=5，epoch=50，negative=10。

（3）词对之间的语义距离：选择神雕侠侣小说中的主人公“郭靖-黄蓉”、“杨过-小龙女”来验证。

（4）词对关系验证：做类似国王-男人=皇后-女人的词类比测试，选择“杨过-小龙女=黄蓉-？”观察输出是不是郭靖来验证。

（5）Kmeans聚类分析：选择各部小说中的主要人物，再通过K-means（参数为7）进行聚类，后通过TSNE将高维降维到2维进行可视化（不同类别使用不同颜色dot），观察各人物的聚类情况来验证词向量的有效性。

（6）段落之间的语义关联：选择倚天屠龙记小说两段相近的语段，分别进行提取词向量后均值，再计算其余弦相似度来验证词向量的有效性

# **Experimental Studies**

（1）词对之间的语义距离试验结果：



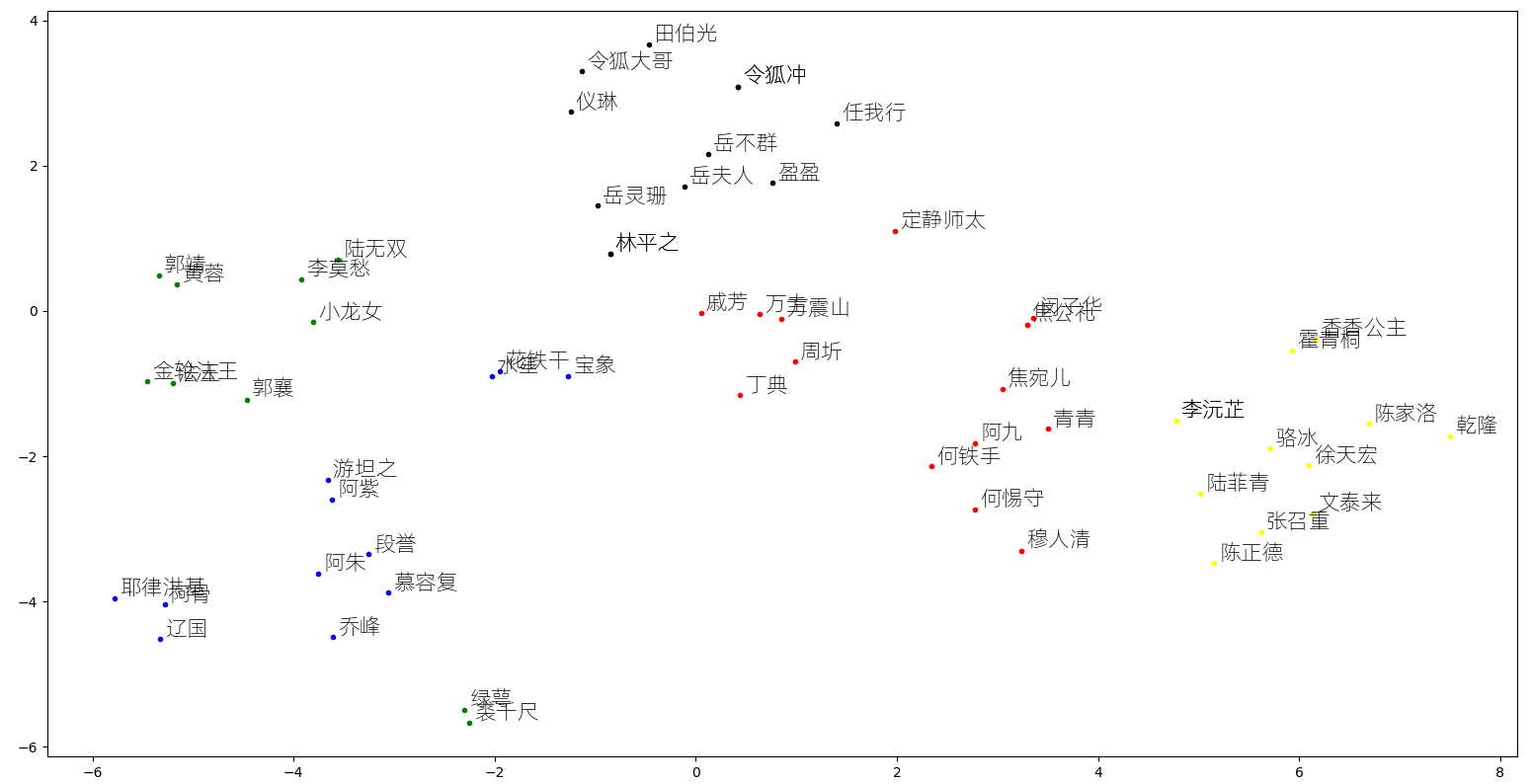
可以看到“郭靖-黄蓉”词对相似度为0.76、“杨过-小龙女”词对相似度为0.81，两对为夫妻或情侣，拥有高相似度是合理的。

（2）词对关系验证试验结果：



可以看到“杨过-小龙女=黄蓉-？”的词类比测试结果是郭靖，是合理的两对为夫妻或情侣。

（3）Kmeans聚类分析试验结果：



通过选择七部小说中的主人公，如陈家洛、袁承志、杨过、张无忌、萧峰、令狐冲和狄云，再通过训练得到的词向量模型提取出相关性前10的人物，共70各人物作为Kmeans数据，经过训练后再通过TSNE降维到2维进行可视化（显示前5类）。从试验结果可以看出聚类效果还是可以的，基本能做到相关人物位于一块。

（6）段落之间的语义关联的试验结果：



提取倚天屠龙记中的两段：

“这般于一刹那间化刚为柔的急剧转折，已属乾坤大挪移心法的第七层神功，灭绝师太武功虽高，但于对方刚猛掌力袭体之际，再也难以拆解他转折轻柔的擒拿手法。张无忌虽然得胜，但对灭绝师太这般大敌，实是戒惧极深，丝毫不敢怠忽，以倚天剑指住她咽喉，生怕她又有奇招使出，慢慢的退开两步。”

“灭绝师太横剑一封，正要递剑出招，张无忌早已转得不不知去向。他在未练乾坤大挪移法之时，轻功已比灭绝师太为高，这时越奔越快，如风如火，似雷似电，连韦一笑素以轻功睥睨群雄，也自暗暗骇异。但见他四下转动，迫近身去便是一刀，招术未老，已然避开。这一次攻守异势，灭绝师太竟无反击一剑之机，只是张无忌碍于倚天剑的锋锐，却也不敢过份逼近。”

两段均是对主人公使用乾坤大挪移的描述，其余弦相关度为0.72，较好的说明了段落之间语义关联，从而验证了词向量的有效性。

# **Conclusion**

本次作业是基于中文语料库，利用Word2Vec神经语言模型，将经过jieba分词后的中文词语进行训练得到词向量。后经过四种方法来验证向量的有效性。1.计算词向量之间的语义距离（如：郭靖-黄蓉，杨过-小龙女）；2. 测试相关词对的有效性（如：杨过-小龙女=黄蓉-？）；3.利用K-means来验证某一类词语的聚类效果（如：各部小说中的主人公）；4.两个相关段落之间的直接语义关联。经过上述四种验证方法均较好的说明了词语之间的有效性。

# **Referances**

[1] <https://www.cnblogs.com/lfri/p/15032919.html>

[2] https://blog.csdn.net/weixin\_44586473/article/details/104158302