Word2Vec 模型训练词向量并验证有效性

Abstract：

利用给定语料库（金庸小说语料如下链接），利用 1～2 种神经语言模型（如：基于 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。

模型选择：Word2Vec 语言模型

验证方式：

1、计算词向量之间的语意距离。

2、利用 K-Means 聚类计算轮廓系数验证。

3、构建段落的词向量计算语意距离，判断语义关联。

Introduction：

1、Word2Vec 模型

Word2Vec 模型是一种用于生成词向量的语言模型，由 Google 的 Tomas Mikolov 等人在 2013 年提出。它是一种基于神经网络的方法，旨在将词语映射到一个具有一定维度的向量空间中，其中相似的词在这个向量空间中距离较近。Word2Vec 主要有两种模型结构：连续词袋模型（CBOW）和跳字模型（Skip-Gram）。 两者的不同在于：CBOW 用上下文词来预测中心词，而 Skip-gram 用中心词来预测上下文词

该模型假设文本中离得越近的词语相似度越高，基于此需要设置多个参数，包括：

1、词向量的维度；

2、窗口大小

3、把词频低于 n 的词去掉

4、模型训练的迭代次数

5、负采样，每次采样本个数。

2、词向量

自然语言处理相关任务中要将自然语言交给机器学习中的算法来处理，通常需要将语言数学化，因为机器不是人，机器只认数学符号。向量是人把自然界的东西抽象出来交给机器处理的东西，基本上可以说向量是人对机器输入的主要方式了。

词向量就是用来将语言中的词进行数学化的一种方式，顾名思义，词向量就是把一个词表示成一个向量。 我们都知道词在送到神经网络训练之前需要将其

编码成数值变量，常见的编码方式有两种：One-Hot Representation 和 Distributed Representation。

Methodology

M1 ： 实验步骤

1、对小说进行句子的结巴分词，生成可以被 Word2Vec 模型处理的数据
2、设置参数，利用 gensim 中的 Word2Vec 模型进行模型训练
3、利用训练模型，计算词向量之间的语意距离（余弦相似性），分析模型
4、利用 K-Means 聚类，计算轮廓系数，分析模型的有效性。
5、选取段落，将段落的词向量的平均值作为该段落的“词向量”，判断语义关联

Experimental Studies

M1 ： 计算词向量之间的语意距离

1、分析与“杨过”语义关系最近的词：

('小龙女', 0.8592096567153931),

('黄蓉', 0.8368332982063293),

('李莫愁', 0.7859969735145569),

('郭靖', 0.7774397134780884)

分析：“小龙女”是“杨过”的师傅和伴侣因此相似度最高，“黄蓉”是在小说中与“杨过”存在矛盾，“李莫愁”与“杨过”关系由坏变好是小说的剧情之一，“郭靖”是“杨过”的叔叔。通过以上分析，“杨过”与以上人物的关系较大，因此语义关系很近，验证了模型的有效性

2、对比小说主人公与关系最好的角色和关系没那么好的角色的语义对比：

'郭靖' 和 '黄蓉'的语义距离相似为: 0.8958644866943359

'郭靖' 和 '欧阳锋'的语义距离相似为: 0.7684441804885864

'杨过' 和 '小龙女'的语义距离相似为: 0.8592096567153931

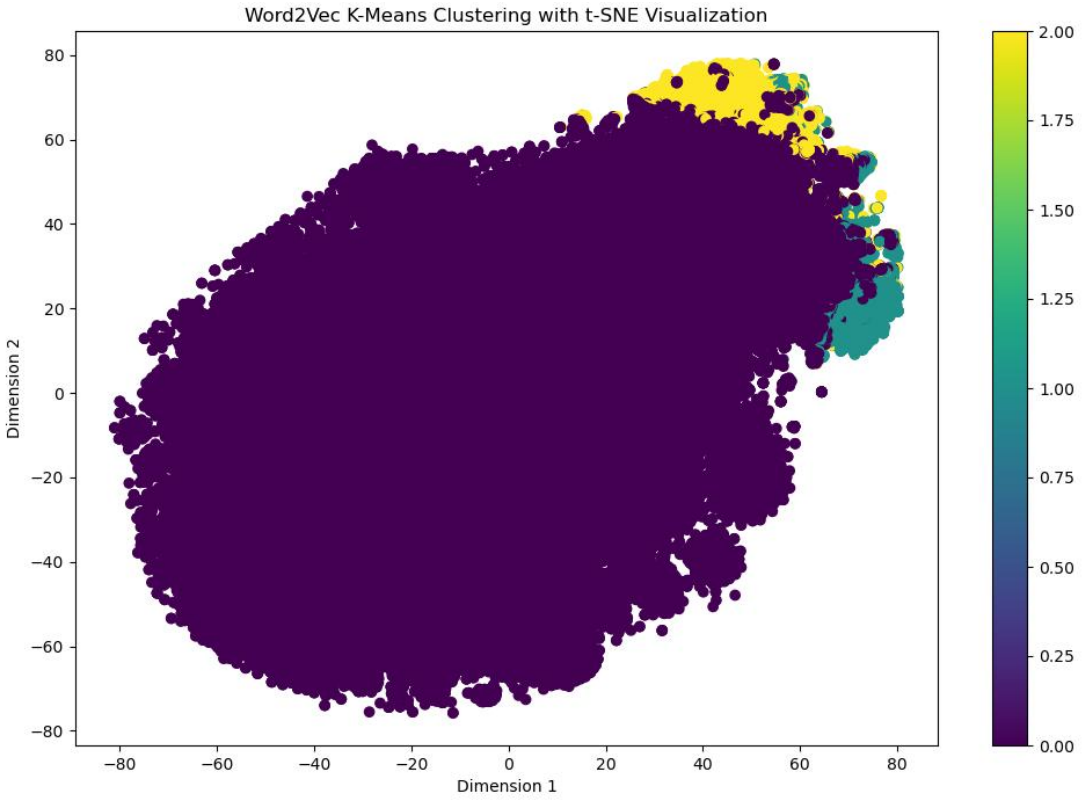
'杨过' 和 '尹志平'的语义距离相似为: 0.5546279549598694

'韦小宝' 和 '康熙'的语义距离相似为: 0.7256892919540405

'韦小宝' 和 '吴三桂'的语义距离相似为: 0.34598487615585327

分析：“郭靖”与“黄蓉”的语义距离比“郭靖”与“欧阳锋”的近；“杨过”与“小龙女”的语义距离比“杨过”与“尹志平”的近；“韦小宝”与“康熙”的语义距离比“韦小宝”与“吴三桂”的近。以上分析与小说剧情相符；因此验证了模型的有效性

M2 ： 利用 K-Means 聚类计算轮廓系数



轮廓系数: 0.8028797507286072

上图为聚类数为 3 的聚类结果，如图所示,基本聚为三类，聚类分明，说明聚类结果较好。

其中，轮廓系数（Silhouette Score）是用于评估聚类效果的一种指标。其取值范围在-1 到 1 之间，值越高表示聚类效果越好。其中 1 到 0.5 表示聚类效果很好；0.5 到 0 表示聚类效果一般，聚类结构不是很明显；0 到-1 表示聚类效果差，很多样本可能被错误聚类。而该模型的轮廓系数为 0.8028797507286072，聚类效果较好；证明了模型的有效性。

M3 ： 构建段落的词向量计算语意距离，判断语义关联。

在语料库中随机选择段落 1：

众人吃了一惊，一齐回过头来。公孙止听了喝声，本已大感惊诧，眼见杨过与女儿安然无恙，站在这蒙面客身侧，更是愕然不安，喝道：「尊驾何人？」

在语料库中随机选择段落 2：

突然之间，乐声一停，随即奏得更紧，正在歌舞的男女纷纷手携手散开，脸上均露诧异之色，向木卓伦等一群人凝望。陈家洛随着他们眼光看去，只见那白衣少女已站起身来，正轻飘飘的走向火堆。众回人大为兴奋，窃窃私议。陈家洛听得身旁的骑兵队长道：“咱们香香公主也有意中人啦，谁能配得上她呢？”木卓伦见爱女忽然也去偎郎，大出意外，很是高兴，眼中含着泪光，全神注视。霍青桐从不知妹子已有情郎，也是又惊又喜。原来她妹子喀丝丽虽只十八岁，但美名播于天山南北，她身有天然幽香，大家叫她香香公主。回族青年男子见到她的绝世容光，一眼也不敢多看，从来没人想到敢去做她的情郎，此时忽见她下座歌舞，那真是天下的大事。

段落相似度：0.869075477123260

在非语料库中随其抽取段落 1：

下面我们通过穿越机的视角，沿长江而下，一起“云游”这跨越 1800 多公里的长江“绿色生态纽带”。

在非语料库中随其抽取段落 2：

欢迎总统先生访问中国并出席中阿合作论坛第十届部长级会议开幕式。

段落相似度：0.4514807462692261

分析：在语料库中随机选择的段落 1 和段落 2 的语义相似度是 0.86907547712326，说明了段落之间的语义相似度较高，验证了模型训练的有效性。同时抽取了非语料库的段落进行语义分析，语义相似度为 0.451480746269，低于语料库中随机选择的段落，证明了模型对语料库内抽取的段落关联性较强。验证了模型的有效性。