作业七. 嵌入计算（Word2Vec与BERT）

生信1802-余思克-2018317220208

实验目的

Word2Vec是一种用于自然语言处理的技术。Word2Vec算法使用神经网络模型从大量文本语料库中学习单词联想。一旦训练完成，这种模型就可以检测同义词或为部分句子建议其他词。BERT是Google开发的一种基于transformer的机器学习技术，用于NLP预训练，是一种较新的、效果较好的模型。本实验将分别利用Word2Vec和BERT进行Covid-19文献的词汇嵌入和展示。

实验步骤

1. 实验数据的获取

本实验将针对Covid-19文献词汇数据开展，使用`git clone`命令从Github中获取相关数据，并将获取的数据压缩包解压。

```
git clone https://github.com/bionlp-hzau/Embedding-experiment-in-BioNLP-
course.git # 获取实验数据
cd Embedding-experiment-in-BioNLP-course/data
unzip data/litcovid-trainingdata.zip # 解压数据压缩包
```

2. 安装实验所需的python依赖包

本实验需要`torch`、`matplotlib`、`nltk`、`scikit_learn`、`transformers`包用于分析，这些包的名称以及版本罗列在了`requirements.txt`文件中，使用`pip3 install`命令即可从文件中安装对应的python依赖包。

```
pip3 install -r requirements.txt # 安装python依赖包
```

3. 修改Word2Vec嵌入计算的相关参数

为了取得不错的嵌入计算效果同时考虑计算计算力，需要对相关参数进行设置调整。

```
self.train_size = 200000 # 45行
self.batch_size = 128 # 47行
self.embedding_size = 100 # 51行
self.epoch = 5 # 57行
self.learning_rate = 0.03 # 59行
```

4. 基于Word2Vec进行嵌入计算与可视化

使用python运行脚本开始进行基于Word2Vec的Covid-19文献词汇嵌入计算。

```
python3 Skip_Gram_basic.py # 开始基于Word2Vec的嵌入计算
```

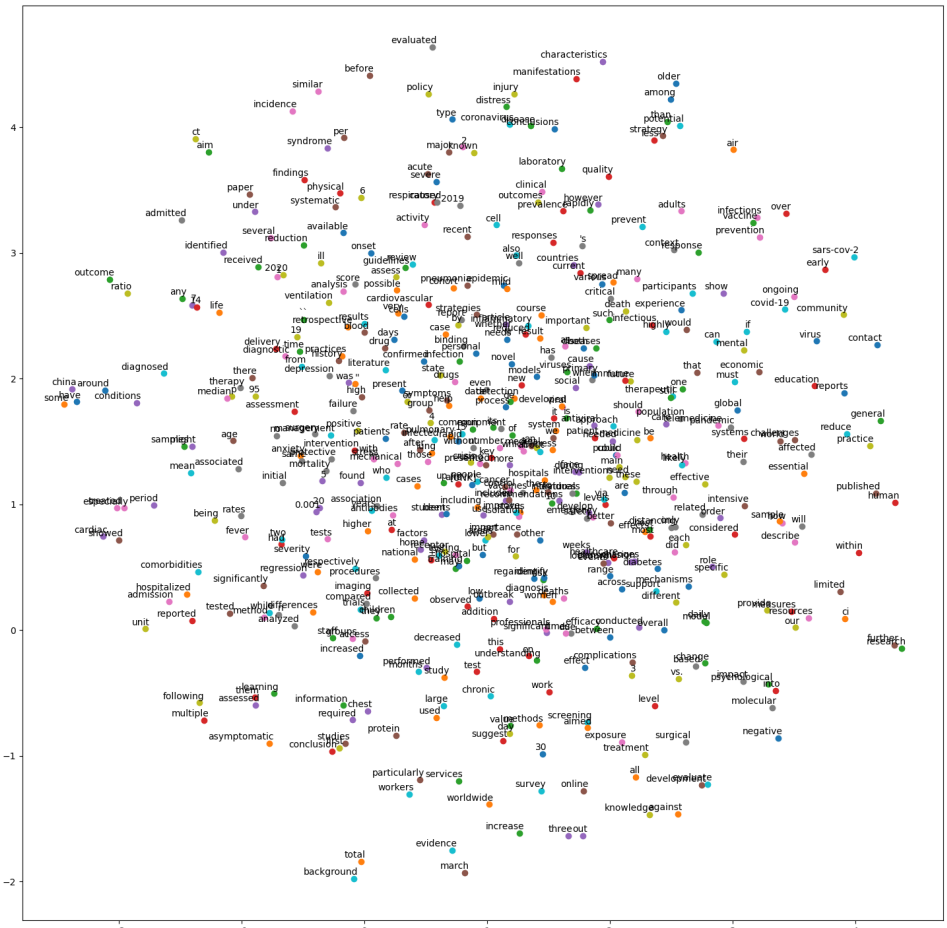
5. 基于BERT进行嵌入计算可视化

使用python运行脚本开始进行基于BERT的Covid-19文献词汇嵌入计算

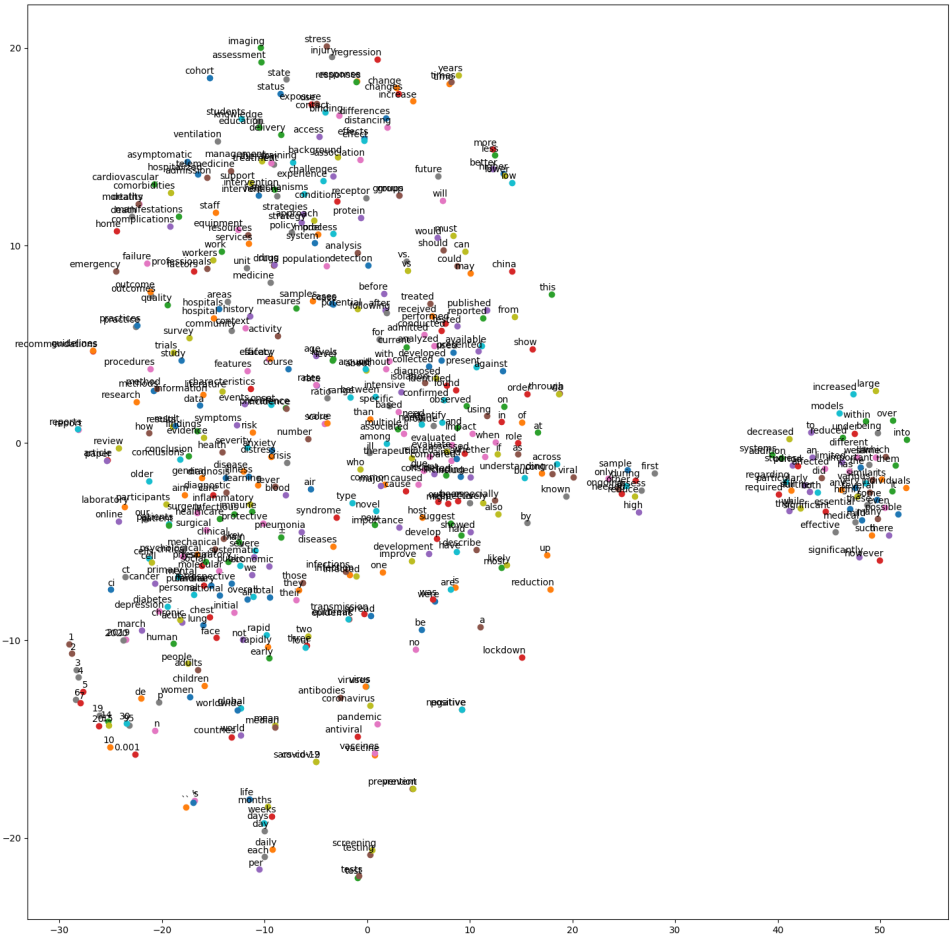
```
python3 Bert_4_Litcovid_WordEmbedding.py # 开始基于BERT的嵌入计算
```

实验结果

1. Word2Vec嵌入计算结果可视化



2. BERT嵌入计算结果可视化



讨论

通过比较两种模型的嵌入计算结果图可以发现，基于BERT嵌入计算的结果图中出现了许多明显的词汇聚集现象，且不同的聚类之间有着较为明显的距离，例如单词"more", "less", "better", "lower", "low"聚集在一起表示“比较”，单词"human", "children", "people", "adults"聚集在一起表示“人类”， 单词"life", "months", "weeks", "days",聚集在一起表示“时间”，阿拉伯数字也有明显聚集在一起。然而，相较于BERT，基于Word2Vec嵌入计算的效果就没有这么明显了，其可视化图片中，大多数词汇的分布较为离散，聚集的现象并没有BERT的结果明显，可以认为，BERT在嵌入计算中的表现效果相对于Word2Vec更为出色。

随着NLP技术的发展，BERT作为一款优质的模型，其训练效果已经得到了广泛的验证，因此，它可以广泛地应用于NLP领域中各种学习问题，例如情感分析、问答系统、阅读理解、信息检索等各种场景。总而言之，作为一款较新的模型，BERT拥有着广阔的应用前景。