自然语言处理第一次实验

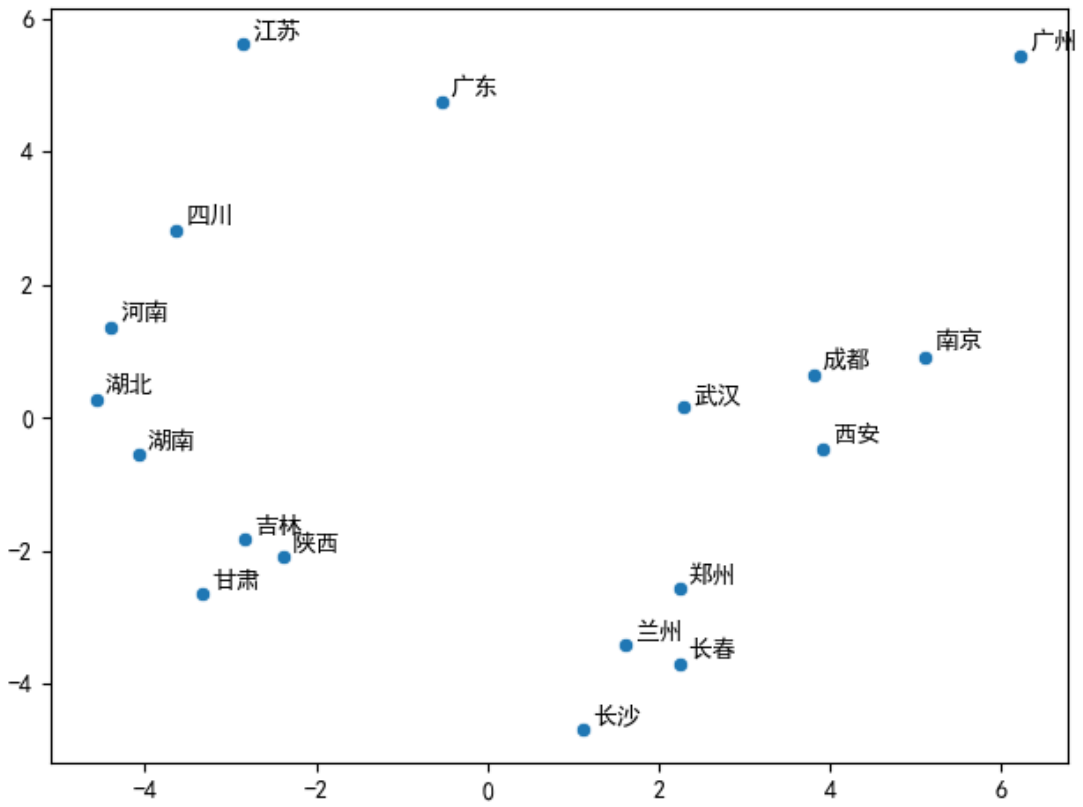
实验环境

python 3 + jieba + gensim + sklearn + matplotlib + numpy + seaborn

实验内容

1. 使用`jieba`分词工具进行分词，使用方法：`jieba.cut(text)`；
2. 使用`gensim`中的`Word2Vec`模型训练词向量：`model = Word2Vec(common_texts, size=100, window=5, min_count=1, workers=4)`；
3. 使用训练好的词向量对指定的词（2个例子）进行相关性比较：`model.similarity('中国','中华')`；
4. 使用训练好的词向量选出与指定词（2个例子）最相似的5个词：`model.wv.most_similar(positive=['武汉'], topn=5)`；
5. 使用训练好的词向量选出与指定词类比最相似的5个词（2个例子），如**湖北 - 武汉 + 成都 = 四川**：`model.wv.most_similar(positive=['湖北', '成都'], negative=['武汉'], topn=5)`；
6. 使用`sklearn`中的`PCA`方法对列表`['江苏', '南京', '成都', '四川', '湖北', '武汉', '河南', '郑州', '甘肃', '兰州', '湖南', '长沙', '陕西', '西安', '吉林', '长春', '广东', '广州', '浙江', '杭州']`（可换成其他）中的所有词的词向量进行降维并使用`seaborn`和`matplotlib`将其可视化：

```
pca = PCA(n_components=2)
results = pca.fit_transform(embeddings)
sns.scatterplot(x=results[:, 0], y=results[:, 1])
```



附加实验（可选）

使用python代替`gensim`实现Word2Vec算法并完成以上 3-6 的实验内容

提交时间

- 基础实验：**10月12号截止**
- 附加实验：**第18周截止**

各班班长或学习委员收集班内所有同学的实验报告和实验代码后发送到`1327793532@qq.com`邮箱中。

提交文件命名方式：`姓名-学号-第X次实验`

实验要求

- 完成所有实验内容
- 良好的代码风格
- 完整的实验报告

参考资料

1. jieba文档
2. Word2Vec Model in gensim
3. PCA in sklearn
4. Word2Vec中的数学原理详解
5. Distributed Representations of Words and Phrases and their Compositionality
6. Efficient Estimation of Word Representations in Vector Space