Word2vec复现说明

一、实验步骤

1. 数据处理

由于原训练语料数据量太大，本次试验中我只使用了前1w5行数据，我预先读出写入了一个新的文件“word2vec_train.txt”中。

读入语料后获得了词表，使用set方法去除重复单词，获得最终词表。

2. 构建权重矩阵、正负例

权重矩阵的维度是393612*50，每读入一个单词时，随机选取出五个高频词作为负例来计算，然后更新权重矩阵中的值。

3. 输出词向量文件

词向量文件第一行为单词个数和词向量维度，中间用空格隔开。

从第二行开始，第一列是单词，后面跟着50个特征值，都用空格隔开。

4. 测试

使用资料中给出的评测文件。可以看到词相似度正确率可以达到0.53左右。

```
wangjiaruideMacBook-Air:201811680771_王嘉瑞_word2vec实现 wjr$ python word_similarity.py
.txt
Word Similarity Evaluation
Handling with the 393613 lines, all 393613 lines.
embedding words 393612, embedding dim 50.
+---------------------+-------+-----------+--------------------+
| Dataset             | Found | Not Found |    Score (rho)     |
+---------------------+-------+-----------+--------------------+
| 词相似度_wordsim-353 |  335  |     18    | 0.5319422733270335 |
+---------------------+-------+-----------+--------------------+
All Finished.
wangjiaruideMacBook-Air:201811680771_王嘉瑞_word2vec实现 wjr$
```