基于词向量的文本搜索推荐系统

1. 开发概述
2. 开发目的：开发一个基于词向量的文本搜索推荐系统，在用户输入关键字后，可以根据语料库自动检索出类似的文本，并呈现给用户，和搜索引擎上的搜索提示比较相近。
3. 开发环境：

操作系统：windows10

开发语言：Python3.6

编译工具：PyCharm 2018.2版本

1. 参考资料：

《Python核心编程》

word2vec相关博客

1. 项目介绍
2. 工程目录，如图2.1所示

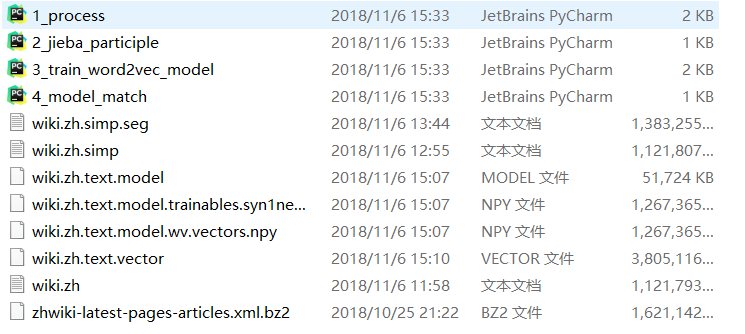


图2.1 工程目录

1. 项目介绍

在图2.1中，zhwiki-latest-pages-articles.xml.bz2是下载的维基百科的语料库，是压缩之后的；然后在1\_process.py脚本中，我按照每行一片文章的规则将压缩的语料库转换为txt格式，程序运行之后的产物是wiki.zh.txt文件。然后运行OpenCC程序将wiki.zh.txt通过繁体转简体转换为wiki.zh.simp.txt文件；在此基础上，脚本2\_jieba\_participle.py文件借助结巴分词工具将wiki.zh.simp.txt文件进行分词，分词之后的文件是wiki.zh.simp.seg.txt。进行分词之后，就可以训练词向量模型了，程序的第三步，即脚本3\_train\_word2vec\_model.py，该脚本完成了词向量模型的训练工作，生成的wiki.zh.text.model就是最终需要的模型文件。最后，使用训练得到的模型进行测试，脚本4\_model\_match.py文件就是一个验证模型的脚本，在该文件中，用户输入一个想要搜索的关键词，程序运行之后，输出在语料库中与之相关性比较大的词汇，并按照相关性的大小，即词与词之间的联系密切程度降序排列，在测试文件中，只列出了前十个相关词汇，从而可以达到智能搜索推荐的效果。

1. 实例展示

在输入“吕布“之后，得到的推荐结果如图2.2所示：

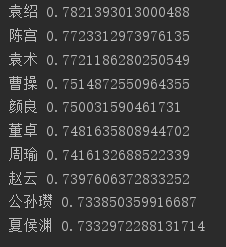


图2.2 推荐结果

图2.2中，左侧一栏是推荐词汇，右侧一栏是该词与搜索关键词之间的相关性。

1. 实验体会和总结

在完成该小项目时，主要做了以下几方面的工作：语料库的下载，语料库的解析，分词，模型训练，模型验证。主要使用到了包含word2vec的gensim模块库，还有结巴分词工具，还有一些python的基本库，总的来说，通过这次实验，对于文本检索有了个了解，然后也学习到了词向量在语料处理方面的便捷和相关的应用，最终做出这么个小的程序，实现关键词推荐功能，功能不多，算是一个基本的使用模块，如果需要的话，可以在此基础上进行相关延伸，满足自己的相关需要。