**基于Doc2vec的图书推荐系统**

一、系统说明

功能介绍:

本系统利用Doc2vec来计算文本相似度，从而实现了根据岗位说明来向该岗位用户推荐图书的功能。

首先将所有图书的书名进行分词，并作为语料训练得到模型，然后再将岗位说明进行分词，放入训练好的模型中，便能够得到该岗位说明文本与所有图书书名文本的相似度，最后取相似度最大的前20本书作为推荐结果。

模块：

get\_datasest 读取所有图书书名，加载停用词表，进行结巴中文分词，将结果保存到列表中。

train 根据分词后的数据，训练生成Doc2Vec模型，并保存训练好的模型。

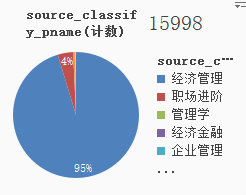
test 加载训练好的模型，获取岗位说明，再次加载停用词表，进行结巴中文分词，载入模型，返回推荐图书以及相似度。

二、数据说明

|  |  |  |  |
| --- | --- | --- | --- |
| 数据名称 | 说明 | 规模 | 格式 |
| newbook | 所有书名，即语料 | 15998个 | txt |
| stopwordss | 停用词库，中文分词的常见停用词，即分词中过滤掉的词汇 | 4000多个 | txt |
| post | post文件夹下有293个txt格式的文件，每个文件对应一个的岗位说明 | 293个txt文件 | 文件夹 |
| result | 推荐结果，每条数据由岗位编号和推荐书目编号组成 | 293条 | txt |

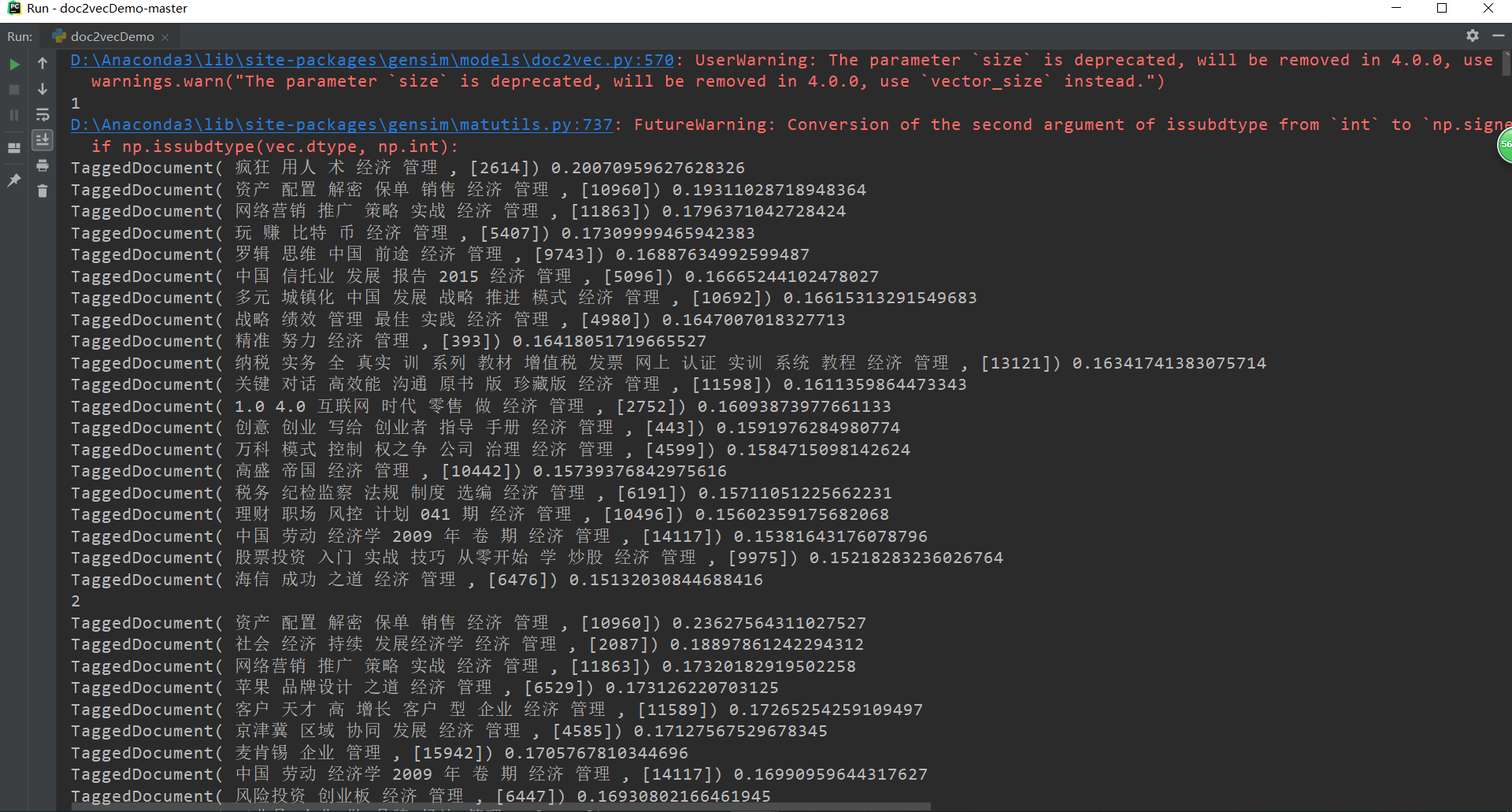
图书选择说明：从公司给的图书源数据中选出15998本图书，用于跟岗位进行文本相似度匹配，其中各类别书籍如下：

|  |  |
| --- | --- |
| 书籍类别 | 数量 |
| 财经人物 | 4 |
| 产品运营 | 1 |
| 成功哲学 | 1 |
| 大学频道 | 1 |
| 管理学 | 51 |
| 互联网思维 | 1 |
| 金融投资 | 1 |
| 经济管理 | 15250 |
| 经济金融 | 15 |
| 领导学 | 9 |
| 名人励志 | 1 |
| 企业管理 | 15 |
| 市场营销 | 14 |
| 演讲口才 | 1 |
| 职场进阶 | 633 |



三、结果说明

系统运行结果如下：



第一行的数字是岗位编号，接下来的二十行是推荐图书的书名分词、图书编号以及相似度，总共293条这样的数据。