

地理选择题问答实验报告

姓名: 周小多 学号: MF1733086 邮箱: xiaoduo_zhou@163.com

(计算机软件新技术国家重点实验室(南京大学),江苏 南京 210023)

1 任务描述

本实验要求完成一个地理选择题的问答系统，根据提供的背景知识对测试集中的问题进行选择。

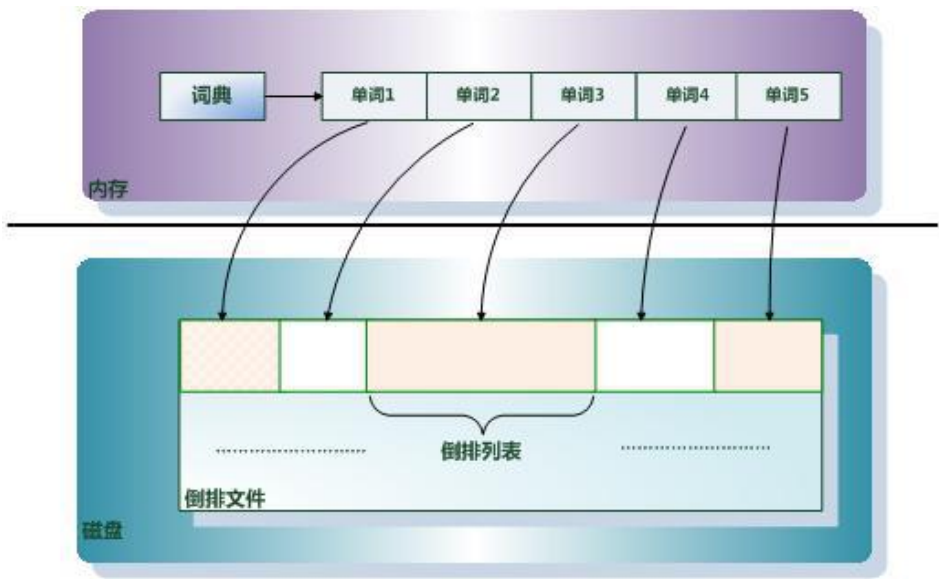
2 相似度计算模块

本模块的任务为根据问题及背景知识在知识库中选择合适的知识来作为该问题的背景知识，相当于做了初筛。极大的减小了计算量，提高了系统效率。

本实验采用了倒排索引来设计这个模块，通过倒排索引能快速的找到和问题相似的知识。

倒排索引原理如下：

- 1. 首先将每个文档的 content 域拆分成单独的 词（我们称它为 词条 或 tokens ），创建一个包含所有不重复词条的排序列表，然后列出每个词条出现在哪个文档
- 2. 如果我们想搜索某个词条，我们只需要查找包含每个词条的文档
- 3. 两个文档都匹配，但是第一个文档比第二个匹配度更高。如果我们使用仅计算匹配词条数量的简单 相似性算法，那么，我们可以说，对于我们查询的相关性来讲，第一个文档比第二个文档更佳。



构造倒排索引的过程还涉及到计算 tf-idf 的过程，TF-IDF（term frequency-inverse document frequency）是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时

会随着它在语料库中出现的频率成反比下降。其步骤如下：

1. 在一份给定的文件里，词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数(term count)的归一化，以防止它偏向长的文件。对于在某一特定文件里的词语 t_i 来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

2. 逆向文件频率 (inverse document frequency, IDF) 是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$





3 答案选择模块

本模块没有采取神经网络的做法，而是将训练集合的题目和正确答案拼接起来丰富知识库的内容，然后将问题先分词，本实验用的是 jieba 分词，然后过滤掉停用词。将问题的背景知识和题目字符串拼接起来，利用前面的相似度计算模块选出 10 个相似的背景知识，然后将 ABCD 四个答案依次与背景知识库做匹配，匹配到关键词最多的就作为最后的答案，如果匹配到的都为 0，则默认选择 A 选项，因为根据统计结果 A 选项的概率较其他选项大。





最后程序运行的准确度为 0.30274

4 运行环境

本程序运行在 ubuntu 平台上，Python 版本为 3.6.2，需要外部的 jieba 的包。文件中包含三个文件和 data 文件夹

 data	2017/12/21 18:59	文件夹	
 Inverted_Index.py	2017/12/12 20:20	PY 文件	4 KB
 Load_data.py	2017/12/12 20:24	PY 文件	3 KB
 Questions.py	2017/12/12 17:20	PY 文件	1 KB

data 文件夹中包含训练文件和验证文件停用词文件，分别为

 knowledge1.txt	2017/12/11 14:29	文本文档	61 KB
 stopword.txt	2017/12/11 21:28	文本文档	12 KB
 test.txt	2017/12/12 19:47	文本文档	624 KB
 train.txt	2017/12/12 17:00	文本文档	3,065 KB

运行方法：python3 Load_data.py