数据挖掘课程项目实验报告

中山大学　　数据科学与计算机学院
15336134　莫凡

2018 年 5 月 12 日

目录

Project I

Supervised Discrete Hash

Project II

使用 PageRank 评估 Longman 的词汇重要性

1 背景介绍

市面上有各种各样的英文（英-英）辞典，其中最为主流常用的当属 Oxford、Merriam-Webster、Longman 与 Collins 等。其中 Longman 字典除去释义准确、词汇丰富以外，最主要的优点是号称几乎所有的单词均用 2000 4000 个核心词汇进行解释。这就避免了英语初学者查找单词的时候不停递归的过程。

这个词汇量的要求对于大学生而言几乎是人人必备的。但是对于更加初级的英语学习者，可否从这几千个单词中筛选出更加“重要”，更加精炼的单词表，使他们在进行单词学习的时候有一个顺序。这种学习是一种基础而非进阶式的学习，目的在于学会这些单词之后，基本整本字典都能看懂，为之后的英语学习扫去障碍。

很显然，最简单的方法就是对整本字典进行词频统计，在释义中出现次数越多越重要。但是这种方法显然是不科学的。比如“蛋白质”一词，在许许多多的生物化学医学词语中都会用到，但是在学习英语中它显得不是那么至关重要。

如果，我们给每一个单词重要程度，那么一个被更加重要的词汇（例如常见的食物、运动等）引用的单词，直观上应当比重要度低的单词（例如医学术语）引用的单词更加重要。因此，根据单词的重要性不同，它们对于其引用的单词的重要度的贡献是不同的。这一点和网页的重要度排名十分相似，所以可以应用 PageRank 算法来进行单词重要程度的衡量

2 假设条件

我们只考虑有实际含义的单词，亦即所有名词、动词、形容词和副词。每个单词看做一个图论意义上的节点。如果单词 A 的释义中包含单词 B，则称单词 B 被单词 A 引用，建立一条 A 到 B 的边，构建了初始的图论模型。

当 B 在 A 的释义中出现多次时，我们只记为一次引用。一个单词的不同词性合并成一个节点。同样的字母拼写，若大小写不同，视作不同单词。

由于一个单词的释义长度有限，如此保证了他的出边数量有限，就可以规避大量相互引用、无效引用的问题。也就是说，链接的质量较高

3 数据处理

我们采用了 `http://global.longmandictionaries.com` 上面的第六版 Longman 辞典作为数据源。由于全文检索是通过 JQuery 进行动态加载的，有较难的爬取难度。但是通过对网站进行分析之后，我们找到入口，发现已知一个单词爬取它的释义是相对容易的。

接下来考虑能否使用这种方法获取我们想要的数据呢？如果一个单词比较重要，那它一定会包含在某些单词的释义中。所以我们只需要一个足够长且常用的初始单词列表，然后通过单词释义，递归搜索出我们需要的单词。显然，这种方法会使得许多单词被排除在外。但是我们有充足的理由去断定

这些未能包含在大量常用单词的释义中的单词对理解整本字典没有较大的价值。因此，我们可以放心地使用这种方式来获取数据.

为了选择一个足够长且常用的单词列表, 我们选择 `https://www.examword.com/ielts-list` 上面的 IELTS 单词列表的 4000 个单词作为搜索起点, 然后对字典进行爬虫.

爬虫使用 Python 的

4 算法实现

5 实验结果

6 总结与反思

7 版权声明与致谢

感谢王明哲和潘嘉慧对这项工作作出的鼓励与帮助。

感谢中山大学第三附属医院。

Project III

使用 k-means 算法进行歌曲分类

1 版权声明与致谢