反爬机制的应对

在开发数据采集模块的过程中，由于某一类网站使用了反爬机制。反爬机制是网站在面对爬虫采取的服务器保护机制，通过IP访问频率的限制，账户验证，ajax，内容加密等技术手段防止网站内容被大量爬取。

国内常用的服务器防火墙有“安全狗”，“风云”等系统。经过对防火墙软件的了解，初步判定被爬取的网页采用的是IP流量限制和IP封锁技术。因此必须使用vpn网络代理。

但是网络上的代理质量参差不齐，代理服务器网络速度快慢不一，对代理的判断也非常重要。

因此，针对特定的代理列表网站结构，反爬机制的应对过程如下：

1.爬取最新的代理服务器列表，包括IP地址和端口，用对称密钥AES解密出明文信息。

2.验证代理的可用性：是否能访问，是否做到更换IP，是否快速访问

3.用代理访问网站，轮流使用可用IP进行爬取。

该模块的设计类图如下：

LDA算法的设计与实现

LDA算法输入与输出

算法输入：分词后得到的文档的集合，主题数目K，超参数alpha和beta

算法输出：

1.每篇文章中的每个词的所对应的主题编号

2.每篇文章的主题概率分布theta

3.每个主题下词的概率分布phi

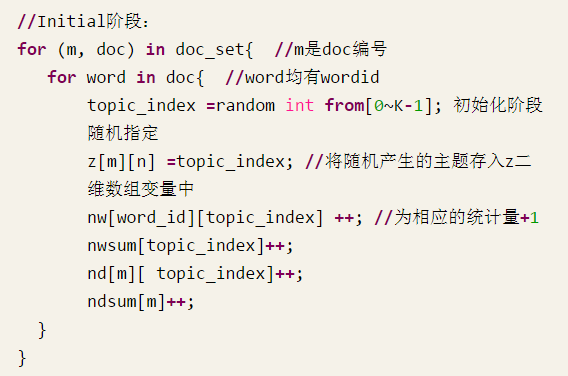
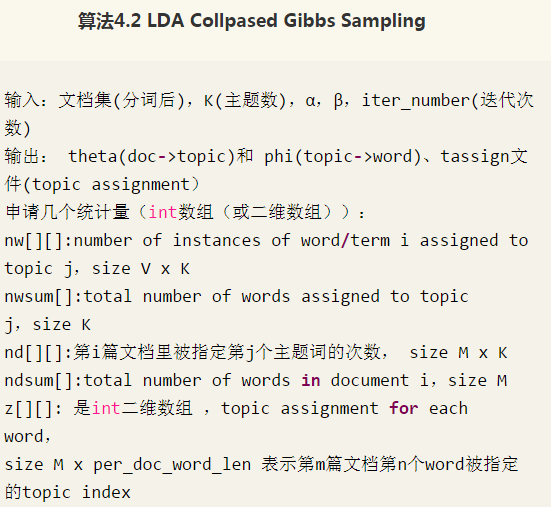
4.程序中词的id映射表

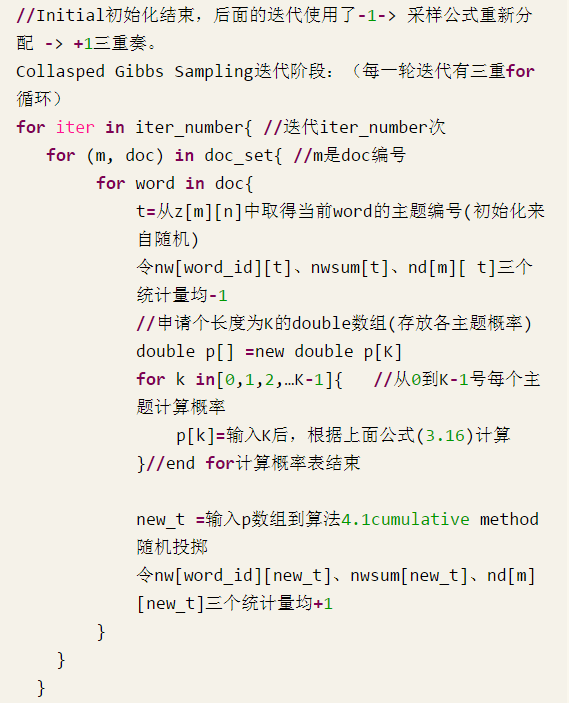
5.在每个主题下phi从高到低的排名

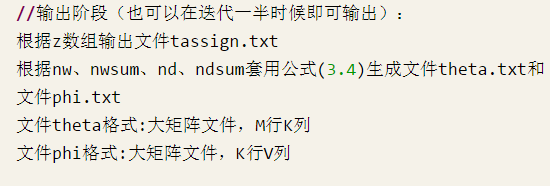
伽马函数

狄利克雷分布

其中，







情感词典的获取：

由于之前项目积累的词库并不是针对于财经股吧，博客或者媒体新闻的，为了保证情感词典的有效性，开发过程中将情感词典的构建作为一个重要任务，同时也为今后能够对词库进行更新打下良好的基础。使用的PMI-IR(Point-wise Mutual Information，点互式信息；IR，Information Retrieval, 信息检索)【引用http://www.doc88.com/p-0794369385446.html】，