四大板块梳理

一、爬虫：爬取特定的网站

1、专利网站

1）爬取企业的专利信息，佰腾网：http://so.baiten.cn/

2）爬取【知识产权局】每周的专利更新

http://epub.sipo.gov.cn/

点击最新的时间的发明公布数据已更新，进入发明专利页面。并提取信息：

- 发明人
- 专利权人
- 地址　排除地址为非国内地址的专利，不爬取。

2、企业基本信息爬取

网站：

- 企查查　http://www.qichacha.com/　　登录才能查看所有信息
- 天眼查　http://www.tianyancha.com/　这个平台最好，所有信息都在同一个页面
- 企多维　http://www.qiduowei.com/　　类似于天眼查

- 悉知 http://www.xizhi.com/ 信息分散在不同栏目下，处基本信息外需登录查看
- 启信宝 http://www.qixin.com/ 必须登录才能查看信息

2.1需要爬取的信息如下：

1）企业基本信息：

- 工商注册号
- 企业类型
- 行业
- 注册资本
- 法定代表人
- 注册地址
- 经营范围

2）股东信息

- 股东名称及出资金额和出资比例（年报中）

3）企业联系信息

- 联系电话
- 联系邮箱
- 网站链接
- 地址

4）对外投资信息

3、股东信息查询

针对企业信息中股东信息中的企业。

1）查询链接 证券投资基金业协会

http://gs.amac.org.cn/amac-infodisc/res/pof/fund/index.html

2）将股东中企业名称输入到页面关键词中，查询后，点击【私募基金管理人名称】，进入查询页面。需要爬取的信息如下：

- 机构诚信信息
- 基金管理人全称
- 注册地址

- 办公地址
- 实收资本
- 机构网址
- 管理基金主要类别
- 员工人数
- 法定代表人
- 产品信息
- 特别提示信息

机构诚信信息:

基金管理人全称(中文):	北京智诚汇志投资有限公司 投诉		
基金管理人全称(英文):	witruth capital		
登记编号:	P1003064		
组织机构代码:	59235474-1		
登记时间:	2014-06-04	成立时间:	2012-03-12
注册地址:	北京市海淀区香山路88号二层A113室		
办公地址:	北京市朝阳区第一商城B座2006		
注册资本(万元):	1,000	实缴资本(万元):	1,000
企业性质:	内资企业	注册资本实缴比例:	100%
管理基金主要类别:	股权投资基金	申请的其他业务类型:	
员工人数:	4	机构网址:	http://www.witruthcapital.com
法定代表人/执行事务合伙人(委派代表)姓名:	鹿峰		
是否有从业资格:	是	资格取得方式:	通过考试

法定代表人/执行事务合伙人(委派代表)工作履历:

时间	任职单位	职务
2011.12 - 2014.07	北京智诚汇志投资有限公司	董事长
2009.11 - 2011.11	清华大学公共管理学院	博士后
2006.03 - 2009.10	民生银行总行	总监,年薪处处长
2001.01 - 2006.03	民生银行西安分行公司业务部	总经理、高新支行行长
2000.01 - 2001.01	华融资产管理公司西安办事处	经理
1998.01 - 2000.01	工商银行西安市城南支行国际业务部	副经理，经理
1995.01 - 1998.01	工商银行陕西省分行国际业务部	科员
1994.07 - 1995.01	商银行陕西省咸阳分行乾县支行信贷科	信贷员

高管情况:

高管姓名	职务	是否具有基金从业资格
鹿峰	董事长	是(通过考试)
岳红云	合规风控总监	是(资格认定)

4、资讯网站信息爬取

1）投资类网站

- 投资界 http://www.pedaily.cn/all/
- 投资中国

http://www.chinaventure.com.cn/cmsmodel/news/capital/list/11.shtml

- 36 氪 http://36kr.com/news

2、生物医药类资讯网站

- 生物谷 http://www.bioon.com/
- 丁香园 http://www.dxy.cn/

跟踪最近一周的新闻资讯，并对文章进行分词，计算词的频数，提取出出现率最高的 30 个词输出，人工判断是否可作为研究线索。

3、涉及的算法：

- 分词
- TF-IDF 计算（直接使用词频的话，会提取出大量的无用词）。

二、搜索

1、目的：通过百度搜索（或 google），提取需要的信息。

2、步骤：

1）给出百度搜索的关键词，并搜索；

2）文章相关性计算，计算文章内容与关键词之间的相关性【？】；

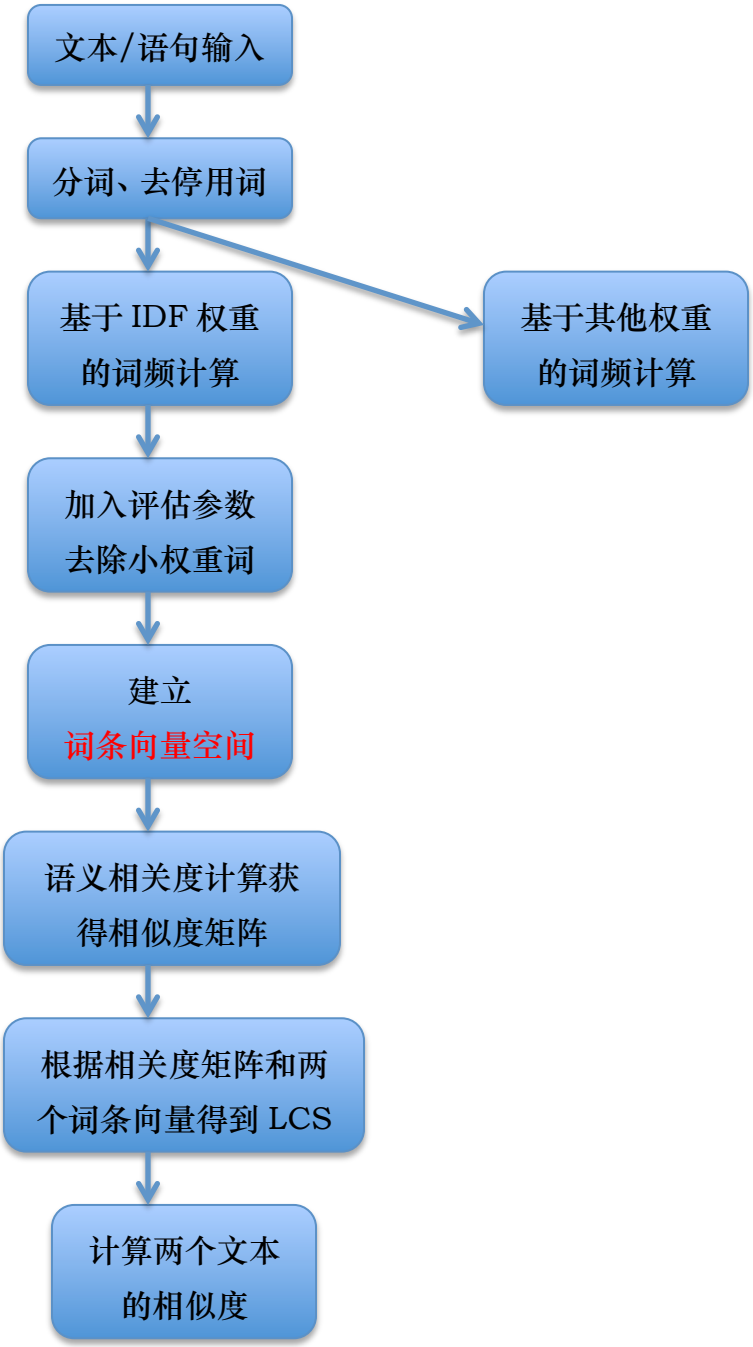
3）筛选相关性超过某一阈值的文章；
4）提取文章摘要。

3、可能涉及的算法
1）VSM、余弦定理-相似性计算
2）聚类算法 K-means
2）摘要提取（见第四部分）

三、排序

四、摘要

五、数据库建立
爬虫部分的资讯网站的爬取时，将分词后的词，进行无用词的剔除后，保存为自己的词库。


文本/语句输入
分词、去停用词
基于 IDF 权重的词频计算
基于其他权重的词频计算
加入评估参数去除小权重词
建立词条向量空间
语义相关度计算获得相似度矩阵
根据相关度矩阵和两个词条向量得到 LCS
计算两个文本的相似度