实习答辩

实习公司：北京明日时尚信息技术有限公司

汇报人：张凌霄

目录
contents

01

实习概要

实习公司

实习公司

北京明日时尚信息技术有限公司

实习内容

Fintech产品开发
基于NLP的金融数据处理

实习岗位

数据获取

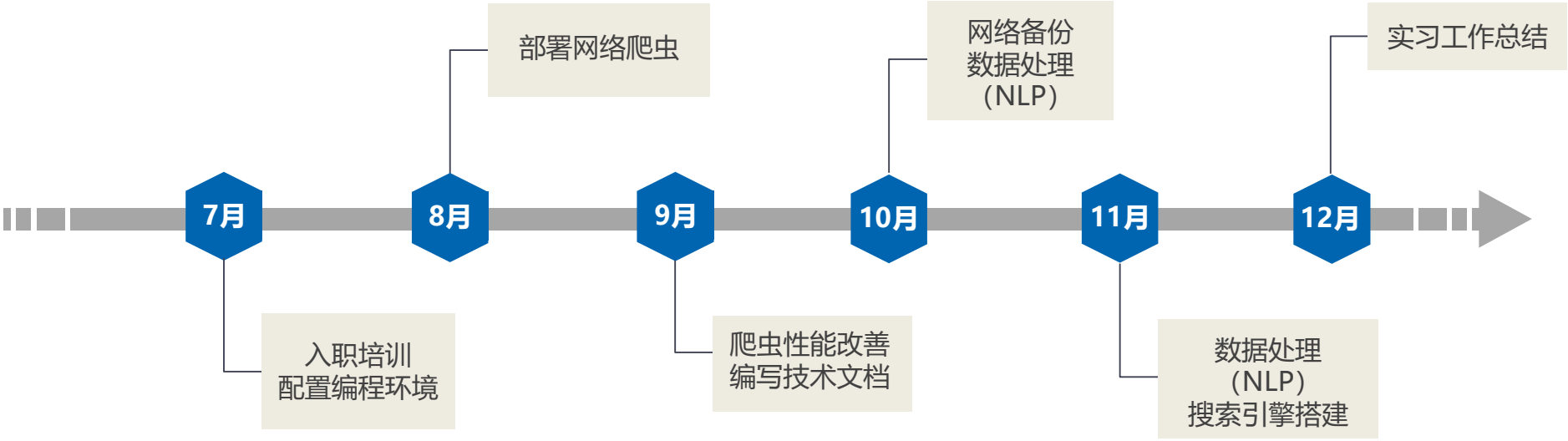
数据存储

数据工程师

数据清洗

数据处理

实习进度


7月
入职培训
配置编程环境
8月
部署网络爬虫
9月
爬虫性能改善
编写技术文档
10月
网络备份
数据处理
(NLP)
11月
数据处理
(NLP)
搜索引擎搭建
12月
实习工作总结

02

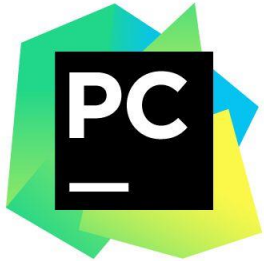
实习内容

运行环境

编程语言

Python

常用作数据爬取和数据处理，
被称为胶水语言。



编程软件

PyCharm

一款Python IDE

浏览器环境

Chrome

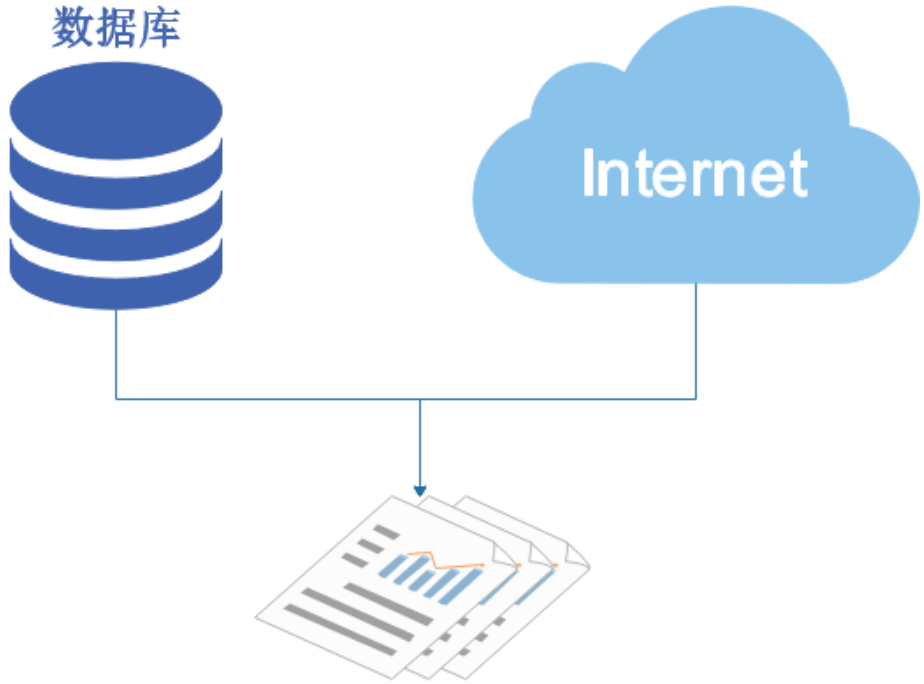
由Google公司开发的网页浏览器

操作系统

CentOS、Windows

CentOS是Linux发行版之一
Windows是基于窗口的操作系统

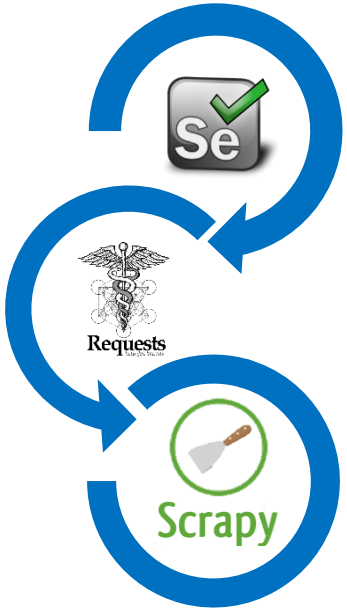
数据获取



数据获取

requests+beautifulsoup

requests库获取目标URL的HTML信息
beautifulsoup库解析HTML格式
元素定位快捷
操作动态加载信息相对困难
爬取大数据量数据时采用



Selenium WebDriver

元素定位方便
可是操作动态生成的网页内容
受网速影响较大，元素定位较慢
爬取小数据量数据或复杂网页时使用

Scrapy

快速、高层次的屏幕抓取和web抓取框架
爬取最大量数据时使用

数据存储

MySQL

1

关系型数据库

SQL语句查询
安全性能高的数据访问
体积小、速度快
成本低，开放源码

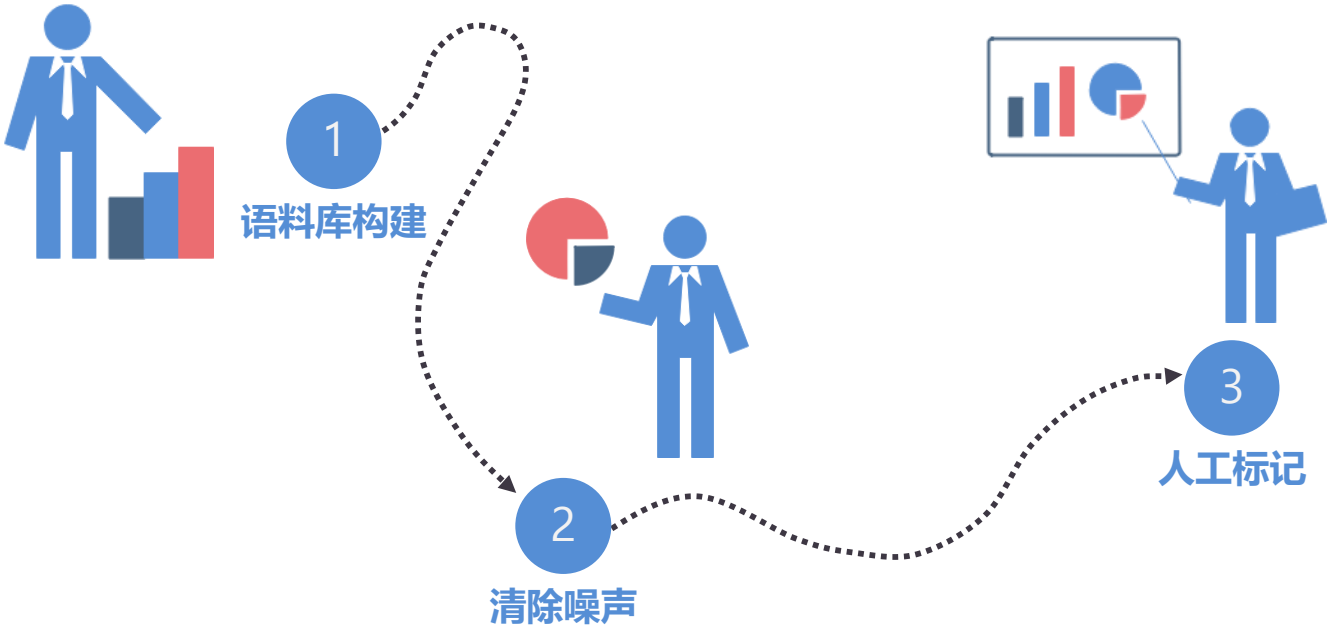
MongoDB

2

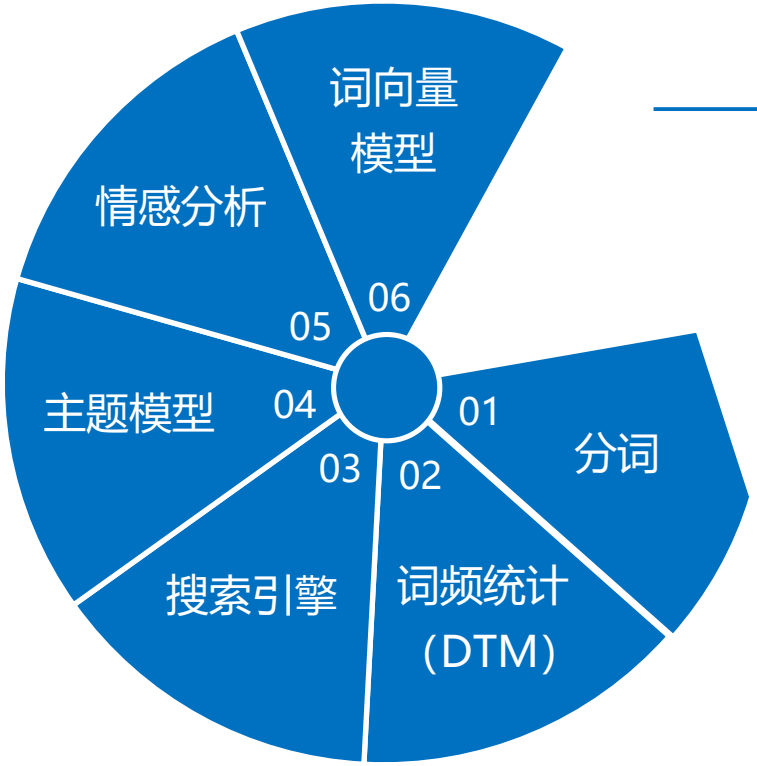
非关系型数据库

基于键值存储
数据之间无耦合性
易水平扩展
无需进行表格设计

数据清理

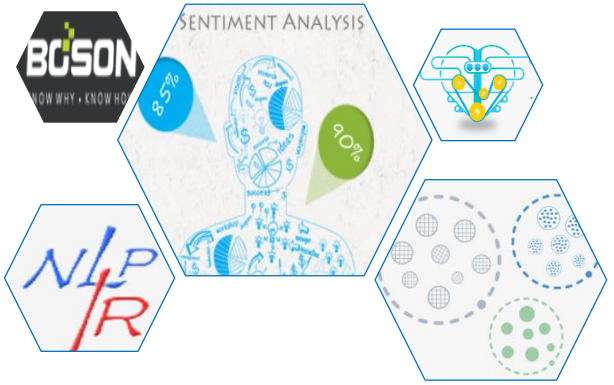


数据处理



主要步骤

自然语言处理（NLP）+搜索



数据处理

分词

运用语言云等分词器对语料库进行分词，
主要分为基于HMM和基于CRF的分词方法
并用人工标注好的样本进行评估

词频统计

主要运用scikit-learn构造词频矩阵（DTM），
并计算词频 - 逆向文件频率（TF-IDF）

语义挖掘

运用主题模型、情感分析等方法对语料进行
语义分析

搜索引擎

使用Elasticsearch或Apache Solr进行关键
词搜索

03

实习成果（心得）

实习过程遇到问题

不同开发人员代码可读性不高

优化模块
降低模块之间的耦合性
提高注释的详细度
规范代码书写格式

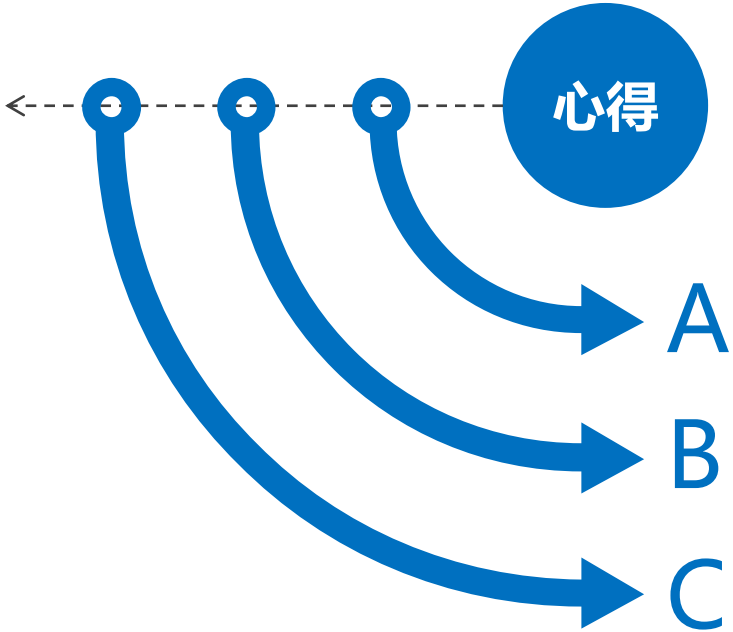
产品经理所提出需求工程不可实现

进行技术例会
与产品经理商议，调整需求

公司人事变更造成工作热情低落

公司开展素拓或聚餐活动缓解工作压力，增强团队凝聚力

实习成果



- 了解了公司运行机制
- 数据产品的研发流程
- 提高团队协作能力

- 提高了编程能力
- 掌握了机器学习、自然语言处理的方法

- 以产品为导向，把控开发时间与质量之间的平衡
- 作为研发人员，与产品经理的及时沟通交流至关重要

谢谢观看