武汉大学电子信息学院本科生
实习报告

院（系）名 称：　电子信息学院

专 业 名 称 ：　电子信息工程

学 生 姓 名 ：　张凌霄

实 习 公 司 ：　北京明日时尚信息
技术有限公司

二〇一七年十二月

摘　要

在当前的互联网大数据时代，众多互联网公司都转型为数据公司。在机器学习和人工智能大热的环境下，优秀、精确的数据无疑是大数据时代的“土壤”。在企业实习期内，我主要从事数据研发工程师的岗位。接触学习了数据获取、数据存储、数据检查、数据清洗、数据处理、数据挖掘等流程，主要处理的数据是文本数据。利用 NLP 的方法对所互联网所爬取的文本数据进行处理，得到用户需要的指标和产品。

关键词： 文本数据、网络爬虫、NLP

目 录

第 1 章 绪论

实习过程的研发阶段主要包括数据获取、数据存储、数据清洗（检查）、数据挖掘等步骤。我在实习期内，针对不同的业务需求，主要从事了数据获取、数据存储、数据挖掘的工作。图 1.1 展示了研发岗位的数据处理基本流程。

图 1.1 研发岗位的数据处理基本流程

第 2 章 数据获取

在研发过程中，采用 Python 爬虫获取互联网数据。

2.1 运行环境

2.1.1 Python 环境：PyCharm

PyCharm 是一种 Python IDE，带有一整套可以帮助用户在使用 Python 语言开发时提高其效率的工具，比如调试、语法高亮、Project 管理、代码跳转、智能提示、自动完成、单元测试、版本控制。此外，该 IDE 提供了一些高级功能，以用于支持 Django 框架下的专业 Web 开发。[1]

2.1.2 浏览器环境：Chrome

Google Chrome 是一款由 Google 公司开发的网页浏览器，该浏览器基于其他开源软件撰写，包括 WebKit，目标是提升稳定性、速度和安全性，并创造出简单且有效率的使用者界面。[2]

2.2 技术路线

针对不同目标网页的特点，我主要采用了三种技术路线。

2.2.1 Webdriver 模拟浏览器

该方法主要特点是元素定位方便，并且可是操作动态生成的网页内容，不必去

专门解析 HTML。但是该种方法受网速影响较大，元素定位较慢。在爬取小数据量数据或复杂网页时适宜使用。

2.2.2 requests 库+beautifulsoup 库

利用 requests 库获取目标 URL 的 HTML 信息，并利用 beautifulsoup 库解析 HTML 格式。该方法优点是元素定位更快捷，缺点是不如模拟浏览器操作方面，操作动态加载信息相对困难。目标 URL 如需用户登录，需要传入 cookie 才可进行爬取。通常在爬取大数据量数据时采用。

2.2.3 Scrapy 库

Scrapy 是 Python 开发的一个快速、高层次的屏幕抓取和 web 抓取框架，用于抓取 web 站点并从页面中提取结构化的数据。Scrapy 用途广泛，可以用于数据挖掘、监测和自动化测试。在爬取最大量数据时采用。图 2.1 显示了 Scrapy 的主要框架。

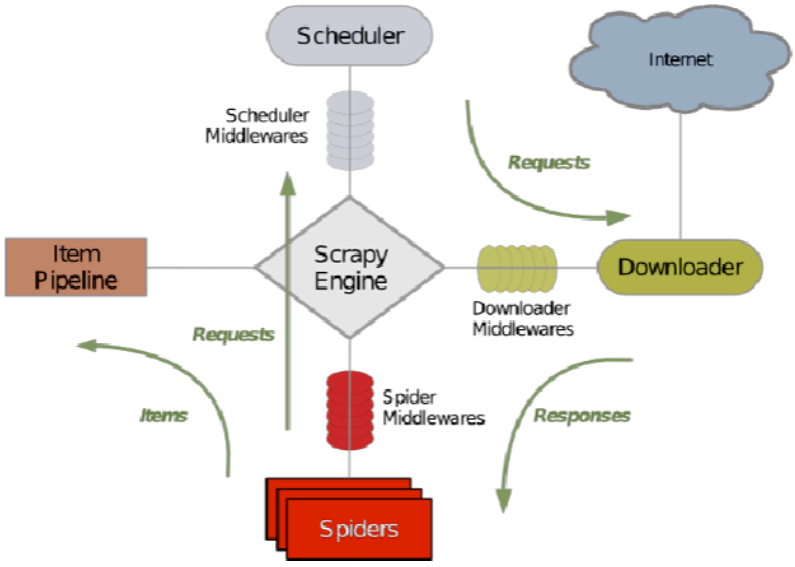

图 2.1 Scrapy 主要框架。

2.2.4 异常处理

网络爬虫受网速的影响十分明显，因此爬虫经常报“链接超时”或“元素定位失败”等错误，此时利用 Python 中 try-catch 方法进行异常处理，记录当前发生异常界面的 URL 并保存在 txt 文件中，延时一段时间后重新链接该 URL，从而获取数据。

第 3 章 数据存储

根据业务的不同采用不同的数据库进行存储

3.1 MySQL

在关系型数据库中，我们采用了 MySQL 进行数据存储。主要为了后续的数据分析方便，采用了第一范式设计数据表格。关系型数据库可以进行用 SQL 语句方便的在一个表以及多个表之间做非常复杂的数据查询，并且使得对于安全性能很高的数据访问要求可得以实现。

MySQL 是一个关系型数据库管理系统，由瑞典 MySQL AB 公司开发，目前属于 Oracle 旗下产品。MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL 是最好的 RDBMS (Relational Database Management System，关系数据库管理系统)应用软件。MySQL 具有体积小、速度快、成本低，开放源码等优点。[3]

3.2 MongoDB

在非关系型数据库中，我们采用了 MongoDB 进行数据存储。非关系型数据库基于键值对进行存储（类型 json 格式），不需要经过 SQL 层的解析，所以性能非常高，并且数据之间没有耦合性，所以容易水平扩展。利于保存网页原始数据，减少了设计数据表格所花费功夫。

MongoDB 是一个基于分布式文件存储的数据库。由 C++语言编写。旨在为 WEB 应用提供可扩展的高性能数据存储解决方案。它支持的数据结构非常松散，是类似 json 的 bson 格式，因此可以存储比较复杂的数据类型。Mongo 最大的特点是他支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，几乎可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立索引。[4]

第 4 章 数据清洗与处理

在经过数据爬取和数据存储之后，需要对存储的文本数据进行清洗和处理。

4.1 数据清洗

数据清洗阶段主要针对不同需求构建完整的语料库，在此期间需要对目标语料进行提取，并且尽可能除去原始语料中噪声。在这里，由于我们的原始语料是 PDF 格式文档，为了方便后续处理，我们首先要将 PDF 格式文档转换为 TXT 格式文档。具体流程图如图 4.1。

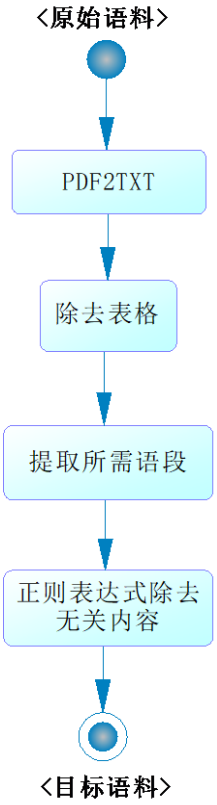

图 4.1 数据清洗流程图

为了使得数据清洗干净，我们在机器处理的基础上，抽样进行人工处理（添加标签）。其中具体清洗的步骤由于涉及到保密协定，在这里不做具体说明。

4.2 数据处理

首先，我们将已经清理好的文本数据进行分句处理；其次，由于我们的目标语料是中文语料，因此我们需要对其进行分词处理；最后，针对用户需求，我们提供了词频统计、情感分析、主题模型等一系列功能。具体流程图如图 4.2。

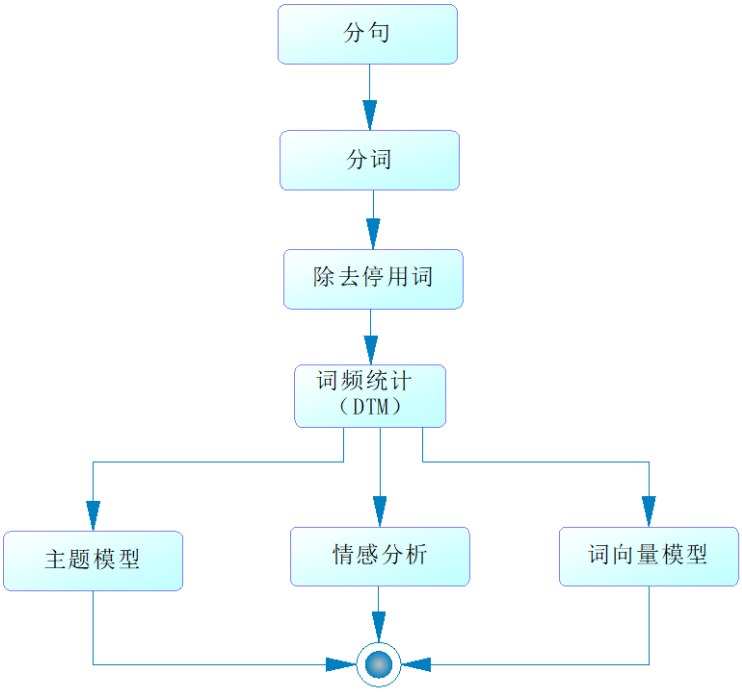

图 4.2 数据处理（NLP）流程图

4.2.1 分词器选择

在自然语言处理过程中，分词效果的好坏直接影响到后续处理，因此我们在分词器的选择与分词效果评估上花费的较多时间。现有的主流分词器如图 4.3。

	分词系统	标识
1	BosonNLP	
2	IKAnalyzer	
3	NLPIR	
4	SCWS	
5	结巴分词	
6	盘古分词	
7	庖丁解牛	
8	搜狗分词	
9	腾讯文智	
10	新浪云	
11	语言云	

图 4.3 主流分词器

针对不同的语料，各分词器的分词效果如图 4.4。所有数据采用北大现代汉语基本加工规范对所有数据进行分词作为标准。

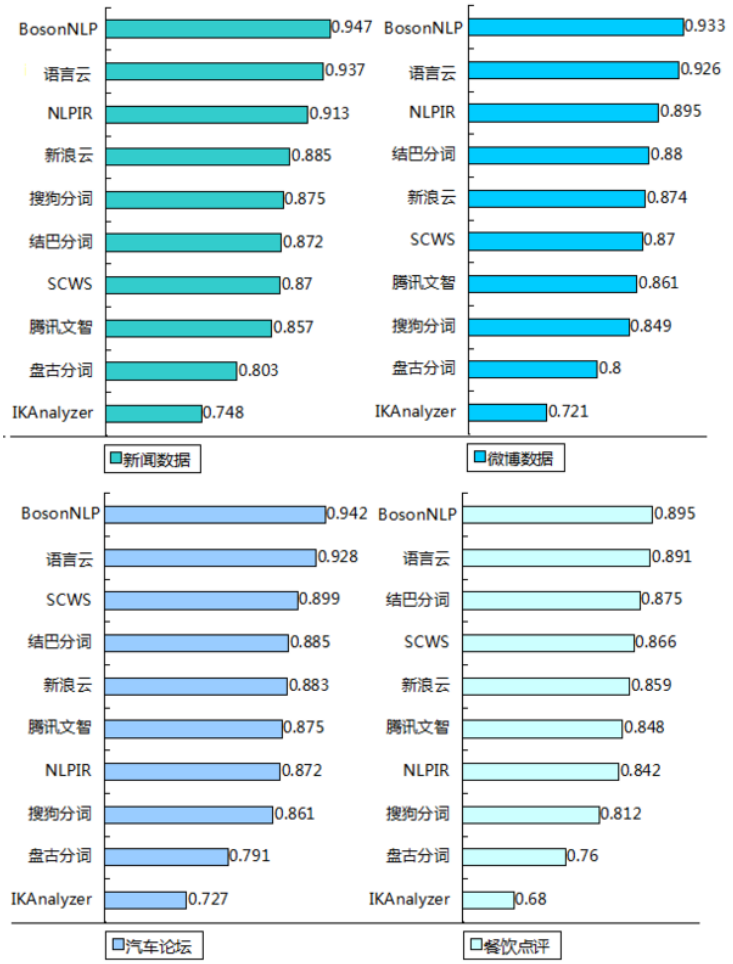

图 4.4 各分词器分词效果

（图中数值为各分词器 F 值，数据采用 BosonNLP 测试数据集[5]）

分词算法主要包括：基于词典的分词方法、基于序列标注的分词方法。而基于词典的分词方法存在着歧义切分、未登录词无法识别等缺点，因此主流分词器主要采用基于序列标注的分词方法。

基于序列标注的分词方法又可分为 HMM (HiddenMarkov Model，隐马尔科夫模型)和 CRF (Conditionalrandom field，条件随机场)：

① HMM

HMM 基本的思想就是根据观测值序列找到真正的隐藏状态值序列。在中文分词中，一段文字的每个字符可以看作是一个观测值，而这个字符的词位置 label(BEMS)可以看作是隐藏的状态。使用 HMM 的分词，通过对切分语料库进行统计，可以得到模型中 5 大要要素：起始概率矩阵，转移概率矩阵，发射概率矩阵，观察值集合，状态值集合。在概率矩阵中，起始概率矩阵表示序列第一个状态值的概率，在中文分词中，理论上 M 和 E 的概率为 0。转移概率表示状态间的概

率，比如 B->M 的概率，E->S 的概率等。而发射概率是一个条件概率，表示当前这个状态下，出现某个字的概率，比如 p(人|B)表示在状态为 B 的情况下人字的概率。 有了三个矩阵和两个集合后，HMM 问题最终转化成求解隐藏状态序列最大值的问题，求解这个问题最长使用的是 Viterbi 算法，这是一种动态规划算法。HMM 算法示意图如图 4.5。

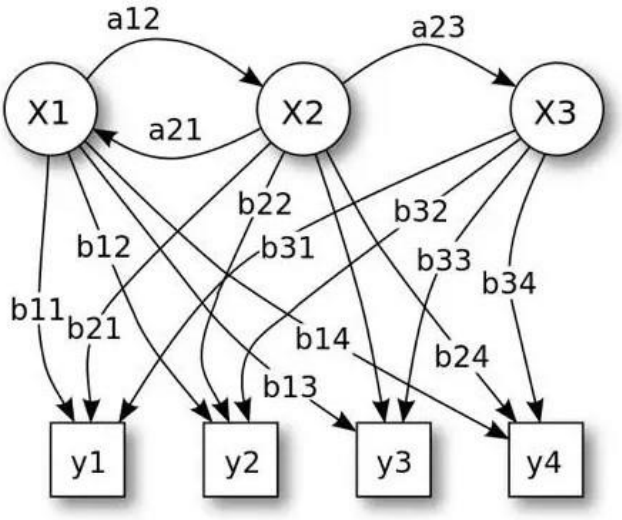

图 4.5 HMM 算法示意图

② CRF

CRF 是用来标注和划分结构数据的概率化结构模型，通常使用在模式识别和机器学习中，在自然语言处理和图像处理等领域中得到广泛应用。和 HMM 类似，当对于给定的输入观测序列 X 和输出序列 Y，CRF 通过定义条件概率 P(Y|X)，而不是联合概率分布 P(X，Y)来描述模型。CRF 算法示意图如图 4.6。

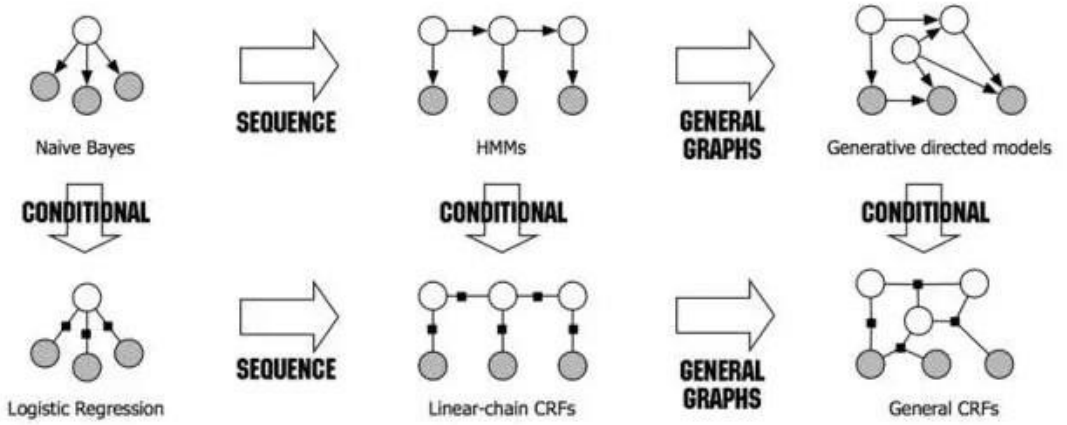

图 4.6 CRF 算法示意图

4.2.2 分词器评价

正确评价一个分词器好坏，首先要有一个“公认正确”的分词结果作为黄金标准分割。

精度（Precision）、召回率（Recall）、F 值（F-mesure）是用于评价一个信息检索系统的质量的 3 个主要指标，以下分别简记为 P，R 和 F。同时，还可以把错误率（Error Rate）作为分词效果的评价标准之一（以下简记为 ER）。

直观地说，精度表明了分词器分词的准确程度；召回率也可认为是“查全率”，表明了分词器切分正确的词有多么全；F 值综合反映整体的指标；错误率表明了分词器分词的错误程度。

P、R、F 越大越好，ER 越小越好。一个完美的分词器的 P、R、F 值均为 1，ER 值为 0。通常，召回率和精度这两个指标会相互制约。P、R、F、ER 计算公式如式 4.1-式 4.4 所示。

$$R = \frac{c}{N} \tag{4.1}$$

$$P = \frac{c}{c+e} \tag{4.2}$$

$$F = \frac{2 \times R \times P}{R+P} \tag{4.3}$$

$$ER = \frac{e}{N} \tag{4.4}$$

其中，N 为黄金标准分割的单词数，e 为分词器错误标注的单词数，c 为分词器正确标注的单词数。

4.2.3 词频统计

单文档词频统计可使用 python 自带 sort()函数。

多文档词频统计可以使用 CountVectorizer 库构建 DTM（Document Term Matrix）矩阵。同时使用 CountVectorizer 库可以生成词频－逆向文件频率（TF-IDF）。

TF-IDF 是一种在文本挖掘中广泛使用的特征向量化方法，它可以体现一个文档中词语在语料库中的重要程度。其计算公式如式 4.5-式 4.6 所示。

$$IDF(t, D) = log \frac{|D|+1}{DF(t,D)+1} \tag{4.5}$$

$$TFIDF(t, d, D) = IDF(t, D) \times TF(t, d) \tag{4.6}$$

词语由 t 表示，文档由 d 表示，语料库由 D 表示。词频 TF(t,d)是词语 t 在文档 d 中出现的次数。文件频率 DF(t,D)是包含词语的文档的个数。如果我们只使用词

频来衡量重要性，很容易过度强调在文档中经常出现而并没有包含太多与文档有关的信息的词语，比如“a”，“the”以及“of”。

如果一个词语经常出现在语料库中，它意味着它并没有携带特定的文档的特殊信息。逆向文档频率数值化衡量词语提供多少信息。

4.2.4 主题模型[6]

LDA 模型是语义挖掘的一种主题模型。这类模型主要为了解决根据 TF-IDF 方法没有考虑到文字背后的语义关联的问题。进而获得多篇文章的主题关联。

LDA 模型的主要思路为：一篇文章的每个词都是通过以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语，即公式 4.7。

$$\mathrm{p}(\text{词语}|\text{文档}) = \sum_{\text{主题}} p(\text{词语}|\text{主题}) \times p(\text{主题}|\text{文档}) \qquad (4.7)$$

4.2.5 情感分析[7]

情感分析是一种常见的自然语言处理（NLP）方法的应用，特别是在以提取文本的情感内容为目标的分类方法中。通过这种方式，情感分析可以被视为利用一些情感得分指标来量化定性数据的方法。在工程中我们利用了 Word2Vec 和 Doc2Vec 方法，通过词向量或段落向量捕捉上下文信息，预测未知数据的情感状况。

总结

在几个月的企业实习中，我着手参与了数据获取、数据存储、数据清洗（检查）、数据挖掘等数据处理步骤，主要完成了网络爬虫的编写、原始数据标记以及自然语言处理等工作。通过企业实习，我进一步了解到了公司运行的机制、数据产品的研发流程，学习并掌握了机器学习的方法，这对我今后的学业和职业发展来说是重要的财富，也使我更明确了今后的发展方向。

参考文献

[1] PyCharm 介绍[EB/OL]. pycharm 开发商官方主页，http://www.jetbrains.com/pycharm/

[2] 谷歌浏览器[EB/OL]. 谷歌浏览器官方正式版，http://www.googlechromer.cn/, 2015-08-28

[3] MySQL 教程[EB/OL]. w3cschool, http://www.runoob.com/, 2014-03-30

[4] 分布式文档存储数据库 MongoDB[EB/OL]. 开源社区网，http://www.oschina.net/p/mongodb, 2012-09-08

[5] 11 款开放中文分词引擎测试数据[EB/OL], Boson 中文语义开放平台，http://bosonnlp.com/dev/resource, 2015-11-06

[6] David M. Blei, AndrewY. Ng, Michael I. Jordan, Latent Dirichlet Allocation[J], Journal of Machine Learning Research 3, 993-1022, 2003

[7] Michael Czerny, Modern Methods for Sentiment Analysis, https://districtdatalabs.silvrback.com/modern-methods-for-sentiment-analysis

致谢

在企业实习结束之际，我向北京明日时尚信息技术有限公司的领导表示感谢，并同时感谢之前在工作中帮助过我的同事。作为大学未毕业的实习生，公司领导给予我充分的信任与学习空间，使得我的编程能力和业务水平都有了很大的提高。大多数公司同事使我在公司感到了归属感，从他们身上学习到的不仅是知识，更是一些为人处世的方式方法。