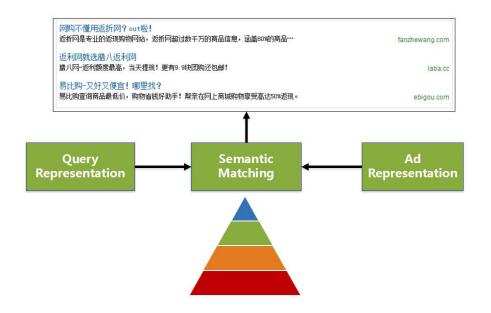
火光摇曳

夜幕降临之际,火光摇曳妩媚、灿烂多姿,是最美最美的.....

[我们是这样理解语言的-1]文本分析平台 **TextMiner**

②2014/06/19 ▶搜索技术、自然语言处理、计算广告学 ✔LDA、NLP、TextMiner、关键词抽取、文本分类、文本聚类、文本语义分析、自然语言处理 ♣ fandywang 互联网上充斥着大规模、多样化、非结构化的自然语言描述的文本,如何较好的理解这些文本,服务于实际业务系统,如搜索引擎、在线广告、推荐系统、问答系统等,给我们提出了挑战。例如在效果广告系统中,需要将 Query(User or Page) 和广告 Ad 投影到相同的特征语义空间做精准匹配,如果 Query 是用户,需要基于用户历史数据离线做用户行为分析,如果 Query 是网页,则需要离线或实时做网页语义分析。



文本语义分析(又称文本理解、文本挖掘)技术研究基于词法、语法、语义等信息分析文本,挖掘有价值的信息,帮助人们更好的理解文本的意思,是典型的自然语言处理工作,关键子任务主要有分词、词性标注、命名实体识别、Collection 挖掘、Chunking、句法分析、语义角色标注、文本分类、文本聚类、自动文摘、情感分析、信息抽取等。

making good progress

mostly solved Sentiment analysis still really hard Best roast chicken in San Francisco! Tokenizer Question answering (QA) The waiter ignored us for 20 minutes. 严守一/ 把/ 手机/ 关/了 Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness? Coreference resolution 严守一把手机关了 Carter told Mubarak he shouldn't run again. Paraphrase Part-of-speech (POS) tagging Word sense disambiguation (WSD) XYZ acquired ABC yesterday ADJ ADJ NOUN VERB ADV ABC has been taken over by XYZ Colorless green ideas sleep furiously. I need new batteries for my mouse. Summarization Parsing Named entity recognition (NER) The Dow Jones is up The S&P500 jumped PERSON ORG I can see Alcatraz from the window! good Einstein met with UN officials in Princeton Housing prices rose Machine translation (MT) Dialog 第13届上海国际电影节开幕... Where is Citizen Kane playing in SF? Text classification The 13th Shanghai International Film Festival... Castro Theatre at 7:30, Do Let's go to Agra! not spam you want a ticket? Information extraction (IE) Buv V1AGRA .. spam You're invited to our dinner May 27 party, Friday May 27 at 8:30

(摘自https://class.coursera.org/nlp/, 稍作修改)

在解决文本处理需求过程中,我们发现保证文本分析相关的概念、数据和代码的一致性,避免重复开发是非常关键的,所以设计并搭建一套灵活、可扩展、通用的文本分析底层处理平台,供上层应用模块使用,是非常必要的。

既然是文本分析,我们很自然的想到是否可以使用已有的自然语言处理开源代码呢?为此,我们不妨一起了解下常见的相关开源项目:

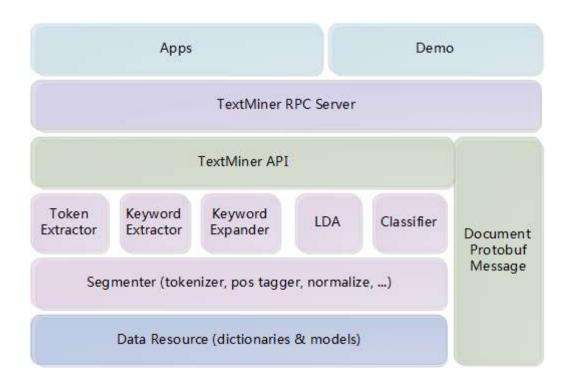
- 1. Natural Language Toolkit (NLTK), http://www.nltk.org/, In Python, 主要支持英文
- 2. Stanford CoreNLP,http://www-nlp.stanford.edu/software/index.shtml,In Java,主要支持英文,阿拉伯语,中文,法语,德语
- 3. 哈工大-语言技术平台(Language Technolgy Platform,LTP),http://www.ltp-cloud.com/,In C/C++,支持中文
- 4. ICTLAS 汉语分词系统,http://ictclas.org/, In C/C++, 支持中文

遗憾的是,我们发现尽管这些项目都极具学习和参考价值,和学术界研究结合紧密,但并不容易直接用于实际系统。也许这正源于学术界和工业界面临的问题不同,定位不同。对比如下:

	学术界	工业界
处理单元	句子级	篇章级
处理维度	词法、句法、语义	词法、语义
处理数据	小规模新闻语料	Web 大数据
处理方法	统计+规则	统计+规则
典型任务	分词、词性标注、命名实体识别、句法分析、语义角色标注、 信息抽取、问答系统等	分词、词性标注、命名实体识别、 关键词抽取、文本分类、文本聚 类等

5. 提供统一的数据资源管理功能,尤其,要支持同时加载多份不同版本的数据资源,便于进行更新及效果对比。

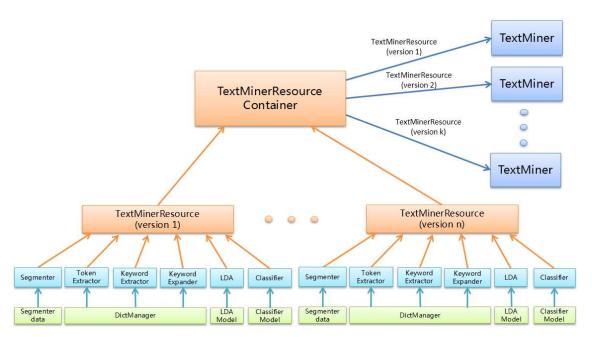
参考斯坦福大学自然语言处理组开源项目: Stanford CoreNLP 和 哈尔滨工业大学社会计算与信息检索研究中心开源项目: 语言技术平台 (Language Technology Platform, LTP) 设计思想,结合实际业务系统常见需求, TextMiner 系统架构如下图所示:



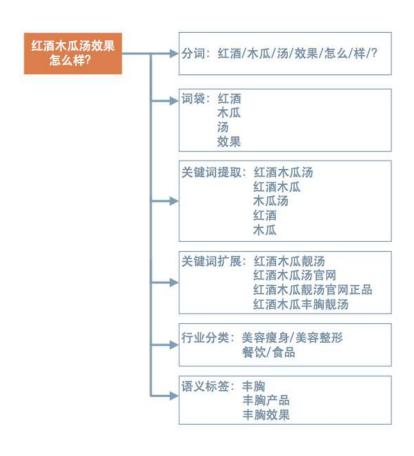
TextMiner 制定了基于 Google Protocal Buffer (简称 Protobuf, Thrift 也是不错的选择) 的文本分析处理结果表示方法,集成了一整套自底向上的文本分析基础模块,主要包括:

- 1. 分词器 (Segmenter): 对纯文本进行切词、词性标注和通用命名实体识别 (如人名、地名、机构名);
- 2. Token 抽取 (Token Extractor): 构建 Bag Of Words (BOW) 模型,并支持标点符号、停用 词、功能词(如连词、代词、助词)等过滤,Token 不考虑词序(词独立)、句法等信息:
- 3. Keyword 抽取 (Keyword Extractor): 匹配抽取与文本语义相关的词或短语,并识别出专有名词,如书名、产品名、品牌名、游戏名、疾病名等,一定程度上考虑了词序和句法,语义更明确;
- 4. Keyword 扩展 (Keyword Expander): 对匹配抽取的少量高质量 Keywords 进行语义扩展,获取更多与文本语义相关且未在文本中提及的词条,解决 Vocabulary Gap 问题;
- 5. Topic 识别 (LDA): 采用 Unsupervised Learning 的方法,对文本进行聚类分析,识别能够表达文本语义的 Topics;
- 6. 层次化文本分类 (Classifier): 采用 Supervised Learning 方法,在人工构建的大规模层次类别体系基础上,对文本进行类别判断,标识出文本所属的行业语义。

中,往往需要通过小流量 A/B Test 实验验证效果正向后,才能完成全量平滑升级,这一点非常重要。所以,这里特别介绍一种可行的设计方案。如下图所示:



TextMinerResource 负责数据资源统一管理,调用者需要基于 TextMinerResource 初始化 TextMiner 对象,然后,各功能模块均围绕 Document Message 进行文本分析处理,上层应用模块也只需要从 Document 中获取所需要的字段即可。 TextMiner 和 TextMinerResource 是一对一的关系。但是,TextMiner 平台设计本身支持多份 TextMinerResource 的存在,即多份数据资源的存在,并使用 version_id/resource_name 进行版本标识,这些数据资源由 TextMinerResourceContainer 维护,初始化时解析配置文件加载数据,使用者需要指定使用哪个 version 的数据资源(即算法策略)做文本处理。 至此,我们就可以使用 TextMiner 处理文本了。比如,给定文本"红酒木瓜汤效果怎么样",可以得到类似下面的文本分析结果:



[我们是这样理解语言的-2]统计语言模型

[我们是这样理解语言的-3]神经网络语言模型

[我们是这样理解语言的-4]说说中文分词

[我们是这样理解语言的-5]命名实体识别

[我们是这样理解语言的-6]关键词抽取

[我们是这样理解语言的-7]关键词扩展

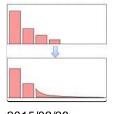
[我们是这样理解语言的-8]主题模型

[我们是这样理解语言的-9]层次文本分类器

[我们是这样理解语言的-10]在线广告系统中的语义分析

本文链接: [我们是这样理解语言的-1]文本分析平台TextMiner 本站文章若无特别说明,皆为原创,转载请注明来源:火光摇曳,谢谢! ^^

相关文章



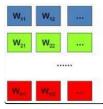
2015/02/28 [我们是这样理解语 言的-2]统计语言模 型

nc. is a well know

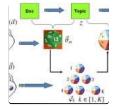
2015/03/02 Peacock: 大规模 主题模型及其在腾 讯业务中的应用



2014/07/21 [LDA工程实践之算 法篇-1]算法实现正 确性验证



2014/06/17 [LDA数学八卦-4] 文本建模



2014/06/17 [LDA数学八 卦-5]LDA 文本建 模

关于fandywang



14条评论

最新 最早 最热



ruoshui1126

感觉和Itp的架构比较类似,都是一整套自然语言处理任务的集合, Document Message 具体实现是XML吗?

2014年6月30日 回复 项 转 发



王利锋Fandy

不是,推荐使用 protobuf or thrift

 2014年6月30日
 回复
 项
 转

 发



hanshumin

代码公开吗?

2014年7月**21**日 回复 顶 转 发



王利锋Fandy

抱歉,代码不公开。

2014年**7**月**23**日 回复 顶 转发



清心竹影

感觉没有干货啊?? 🤮

2014年8月3日 回复 顶 转发



王利锋Fandy

2014年8月18日 回复

顶 转发



暴君祥子

RPC Server 也是自己开发的吧,还是有开源的替代?

2014年10月14日 回复 顶 转 发



王利锋Fandy

自己开发的,当然,也可以使用开源 的,如Thrift

2014年10月27日 回复 顶 转发



机器学习爱好者

有从网页文本到行业类目的训练集么

2015年2月26日 回复 项 转发



test

后面的还没写吗?

2015年2月27日 回复 顶 转发



王利锋Fandy

马上会有后续文章发出,谢谢关注

2015年2月28日 回复 项 转 发



aeolus

怎么没有3的?

2015年**5**月**14**日 回复 顶 转 发



宅男潇涧

博主继续写啊,等着看呢,好期待啊



 2015年6月20日
 回复
 项
 转

 发

分享到::

火光摇曳正在使用多说