# 智能问答系统项目报告

2015年12月 《互联网数据挖掘》大作业

## 目录


- 作者信息
- 分工情况
- 编译 / 运行环境
- 系统架构和关键技术
    - 问题分析
    - 段落检索
    - 答案抽取
    - 分词与预处理
    - 其它技术
- 开放测试
- 是非题
- 外部资源
- 计算公式
    - 段落检索阶段
- 参考文献


## 作者信息

本组成员名单和学号信息如下表：

| 成员 | 学号 |
| --- | --- |
| 史舒扬 | 1300012959 |
| 张闻涛 | 1300012758 |
| 张天宇 | 1300012796 |

## 分工情况

本组成员分工情况信息如下表：

| 成员 | 工作 |
| --- | --- |
| 史舒扬 | 分词、问题分析 |
| 张天宇 | 段落检索 |
| 张闻涛 | 答案抽取 |

# 编译 / 运行环境

- 运行环境: Linux / Unix
- 需要：
  - Python 解释器
  - 哈工大自然语言处理平台ltp(分词需要)

# 系统架构关键技术

系统主要架构分为三个部分： **问题分析** , **段落检索** , **答案抽取** 。下面是具体的解释。

## 问题分析

- 对问题进行分词
- 问题建构：提取关键词，作为段落检索和答案抽取的标准
  - 利用规则寻找中心词
    - 如果动词是“是”，则是动词前的最后一个名词
    - 否则，使用疑问词后的第一个名词
  - 将其它名词和非停用动词放在后面作为关键字
- 问题分类：得到答案类型，作为答案抽取的标准
  - 通过一定的步骤将问题分为若干类
    - `Q_person`
    - `Q_place`
    - `Q_time`
    - `Q_number`
    - `Q_other`
  - 利用三类规则判断出一些问题的类别
    - 疑问词判断，例如“谁”，“哪里”
    - 中心词判断，例如“国家”，“城市”
    - 中心词的POS (Part of Speech) 判断，例如“nh”（表示人名），“ns”（表示地名）

    - 利用学习的方法判断出一些问题的类别
        - 将标注过的sample，以及规则判断出来的问题作为训练集，交给若干个SVM进行训练，对剩余的问题进行类别判断

### 段落检索

- 建立倒排索引：根据维基语料建立每个词的倒排索引
- 对每个问题，返回相关的篇章若干
    - 初步筛选：对出现过所有问题中专有名词的文章进行以下考虑
    - 计算问题里每个词的 *tf-idf* 值
    - 对名词、专有名词分别赋予不同的权重，计算它们的和作为文章的权重
    - 按照权值排序，返回权值最高的若干篇（不多于5篇）文章

### 答案抽取

- 从返回的相关文章中得到最可能包含答案的句子
    - 方法：基于同义词词林的词语相似度，计算句子相似度并取相似度最大的句子

- 从上一阶段的句子中抽取出符合要求的答案
    - 根据问题分类和句子中的词性标注、命名实体识别，找出所有的候选词
    - 按照距离问题关键词的加权距离对候选词排序得到答案

### 分词与预处理

- 预处理是：
    - 对维基语料提取与分词，
    - 对问题进行分词与命名实体识别。

- 利用哈工大的语言平台实现。

### 其它技术

- 使用POSIX线程、多进程等方式进行并行计算；
- 中间结果通过磁盘文件传递。

## 开放测试

将问题复制到百度进行搜索，将搜索到结果的前几条的内容提取出来作为答案。

（其中，百度知道、百度文库、百度百科的结果效果相对比较好）

## 是非题

同样对问题进行分词，进行段落检索。

- 如果检索出没有符合条件的段落，则返回“No”；
- 如果答案抽取过程中得到的关键的句子同时包含那些名词 / 专有名词，则返回”Yes”
- 随即返回"Yes" 或者 "No"

# 外部资源

- 分词平台: 使用哈工大语言云离线版
    - 链接：https://github.com/HIT-SCIR/ltp
- 哈工大ltp的python版本
    - 链接：https://github.com/HIT-SCIR/pyltp
- Wikipedia XML文章提取成文本文件，使用Github开源项目WikiExtractor
    - 链接：https://github.com/attardi/wikiextractor

# 计算公式

## 段落抽取阶段

对一篇长度为 $n$ 的文章 $S$，针对特定的问题，它的权重是

$$S = \sum_{w \in N(S)} tfidf(w, S) + 10 \sum_{w \in Ni(S)} tfidf(w, S)$$

其中 $N(S)$ 是文章 $S$ 的名词集合（可重复），$Ni(S)$ 是文章 $S$ 的专有名词集合（可重复）。

# 参考文献


- Shih, Cheng-Wei, et al. "ASQA: Academia sinica question answering system for NTCIR-5 CLQA." *NTCIR-5 Workshop, Tokyo, Japan. 2005.*
- Lee, Cheng-Wei, et al. "Chinese-Chinese and English-Chinese question answering with ASQA at NTCIR-6 CLQA." *Proceedings of NII-NACSIS Test Collection for Information Retrieval Systems (NTCIR'07) (2007): 175-181.*
- Lee, Yi-Hsun, et al. "Complex question answering with ASQA at NTCIR 7 ACLIA." *Entropy 1 (2008): 10.*
- Hsu, Wen-Lian, Shih-Hung Wu, and Yi-Shiou Chen. "Event identification based on the information map-INFOMAP." *Systems, Man, and Cybernetics, 2001 IEEE International Conference on. Vol. 3.*

*IEEE, 2001.*

- Day, Min-Yuh, et al. "An integrated knowledge-based and machine learning approach for Chinese question classification." *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on. IEEE, 2005.*