中文智能问答系统—项目报告

一、组员

成员	学号
陈小康	1500012741
金典	1500012815
赵浩然	1500012797

二、分工情况

1、陈小康

- 架构设计、文本预处理、问题分析、开放测试、文档撰写

2、金典

- 段落检索、答案抽取

3、赵浩然

- 问题调研、开放测试、文档撰写

三、实验环境

●运行环境：Mac OS sierra 10.12.6

●编程语言：python 3.6.2

●第三方库：jieba、gensim、opencc、pandas、sklearn、tensorflow、numpy、urllib、requests、bs4、哈工大分词工具—pyltp

四、封闭测试系统架构关键技术

系统主要架构分为四个部分：**语料预处理**、**问题分析**、**段落检索**、**答案抽取** 。

1、预处理

●对原始 Wiki 中文预料，使用网上代码WikiExtractor.py进行提取，并用opencc进行繁简转换。对提取结果再用正则表达式抽取有效信息

●对抽取结果进行分词操作，获得分词后的语料，对其中每个词建立倒排索引，方便检索

2、问题分析

●用pyltp对问题进行分词，进行词性标注

●利用制定好的规则，提取问题中的疑问词、中心词、关键词

●利用规则+深度神经网络（DNN）来判断问题类型（设置五种）

- person、time、place、organization、number、other

●DNN实现多分类

- 对训练集进行分词，对每个词用word2vec训练4维词向量
- 让疑问词和中心词的词向量组成8维向量，作为问题的特征值
- 将特征值和标签放入网络中训练，使用TensorFlow，输入为8维向量，隐藏层为三层（10 * 20 * 10），输出6维向量（问题的标签）

3、段落检索

●对每个问题，返回相关的篇章若干

●建立倒排索引：根据维基语料建立每个词的倒排索引

●筛选出包含之前提取出关键词前3个的段落

●计算这些关键词的 *tf-idf* 值

●对词性为名词、专有名词的关键词分别赋予不同的权重，计算它们的和作为文章的权重，普通名词权重为1，专有名词权重为10

●按照权值排序，返回权值最高的若干篇（不多于5篇）文章

4、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5个句子，具体计算公式见参考文献

●从上一阶段的句子中抽取出符合要求的答案，计算这些答案与问题的相似度（权值），计算方法为距离问题关键词的加权距离计算

- 根据答案的权值进行排序，从高到低判断答案的词性是否符合问题分类结果
- 第一个符合词性要求的即为答案

五、开放测试

- 使用爬虫技术，将测试集问题调用百度搜索API，取前面几个网页的内容简介，输出到新文件passage.txt。
- 对于每个问题，在passage.txt中进行检索，根据一些关键词，比如***最佳答案***，***[专业]答案***以及利用和问题的模糊匹配进行查找，并结合封闭测试结果，选取出可能的答案。保留不多于20个字符。

六、理论部分—计算公式

1、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）。

2、答案提取阶段

相似度的计算见参考文献

备选答案权值计算（距离问题关键词的加权距离的计算）：

对于某个备选句子S，计算其每个词与问题中词的最大的相似度，当S中某个词的最大相似度大于0.5，则对其左右各5个位置增加权重，遍历完该备选句子后，备选答案所在位置权值即为该备选答案的权值。

最终某个答案的权值为它在每个句子中权值之和。

七、结果分析

1. 封闭测试

   a) 方法：随机抽取给定测试集的一百个问题进行分析，统计其中的事实性问题个数、事实性问题中回答大致正确的个数（包含正确答案的个数）

   b) 统计结果

事实性问题个数	74
回答大致正确问题个数	10
正确率	13.5%

2. 开放测试

a) 方法：随机抽取给定测试集的一百个问题进行分析，统计其中的事实性问题个数、事实性问题中回答大致正确的个数（包含正确答案的个数）

(b) 统计结果

事实性问题个数	70
回答大致正确问题个数	58
正确率	83%

八、源码使用说明

- 请参见各个文件夹里的 README.txt