**Medical Question-Answer System Report**

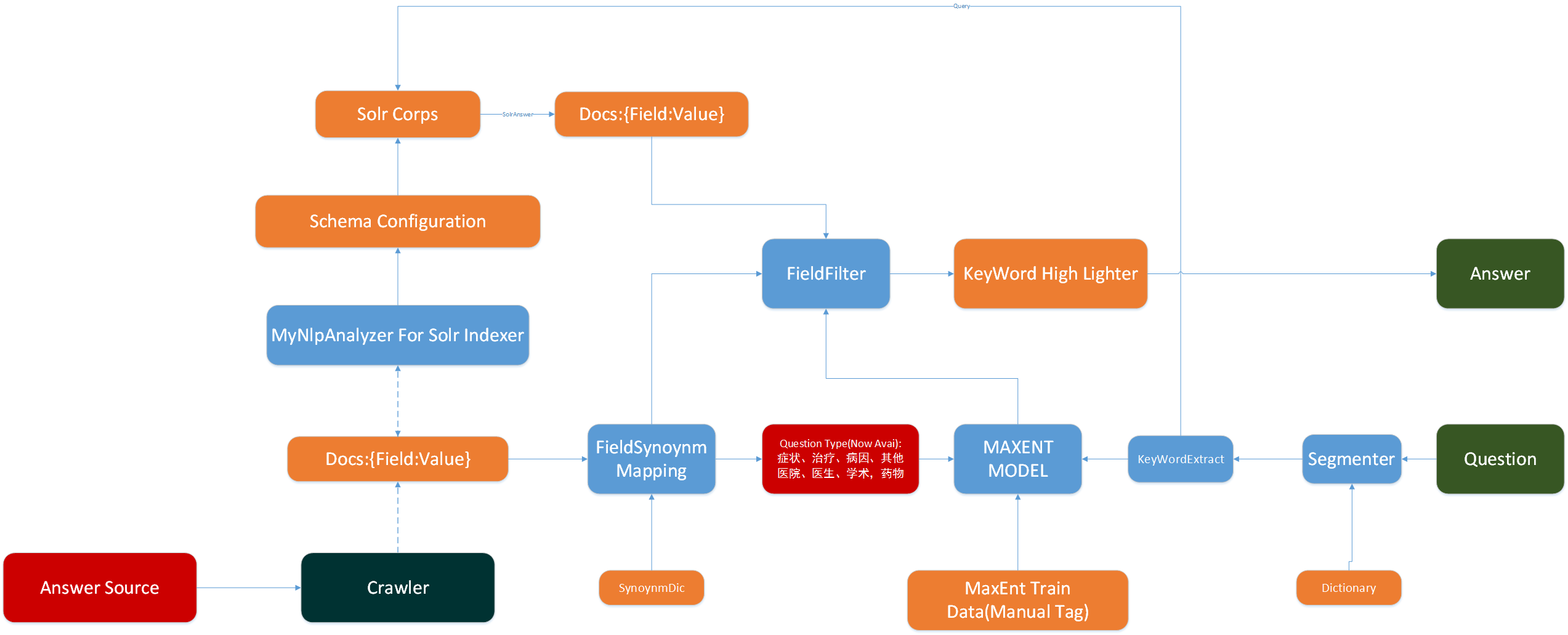
**2015/10/23**

1. **Corpus Introduce**
   1. Question Corpus

324 questions are hand labelled, Include 9 question types: 医生(30)、医院(21)、学术(32)、治疗(41)、病因(30)、病症(33)、症状(29)、药物(33)、其他(75). The questions is random selected from 10000 questions which are crawled from <http://www.120.net/list/>

* 1. Answer Corpus  
     1402 formatted diseases materials are collected and they are crawled from <http://www.baikemy.com/disease/list/0/0>. All diseases are labelled by doctors from the website, but there are 314 sub-items (we can regard them as our answer types), as below:  
     概述,病理分型,临床表现,神经影像学检查,肿瘤标志物检查,诊断,治疗,现代治疗趋势,病因,检查,鉴别诊断,预后,血肿特点,分类,分期,预防,流行病学,病因及分类,病因病机,辨证施治,耐药相关定义,临床分型,并发症,异物的种类,预防与调摄,病因与发病机制,病理,发病机制,预防与治疗,诊断及鉴别诊断,预后与随访,辅助检查,适应证和禁忌证,覆盖义齿类型选择,种植覆盖义齿的维护,生物学行为,即刻种植的优缺点,手术步骤,注意事项,患者术前准备,种植治疗设计,种植患者的口腔维护,相关检查,子词条,肿瘤类型,作用机制,饮食疗法,护理,临床分型及病理,病理生理,中毒机制,治疗及预后,分类与诊断,病因及发病机制,病因及病理特点,常见的癌前病变,诊断标准,失败原因,造血干细胞的分类与功能,造血干细胞异常,鉴别要点,饮食调理,预防护理,分型,腹痛的感觉,腹痛常见疾病,分型标准,趋势,常见非有机磷农药,临床表现及诊断,处理方法,副银屑病分型,常用疗法介绍,锻炼步骤及方法,Richard（1998）前庭操方法,离断伤分类,现场急救,急救,固定方法,杀鼠药分类,适应证,禁忌证,自体造血干细胞移植的步骤,预处理的分类及优劣,现场处理,脐血移植的优点,脐血移植的缺点,脐血移植适应证,疗效,临床常见病,急救处理,急救及搬运,分期标准,分期、分子遗传学发现和治疗,大发作救治,急救要点,常用止血方法,骨髓穿刺部位,认识误区,骨髓穿刺失败原因,建立骨髓库,缺点,救治原则,常备药中毒及救治,白血病十大征象,治疗原则,银屑病的分型,常用药物,光敏剂,光疗方法,副作用,一级预防措施,病因和发病机制,误诊的防控,临床应用,筛查项目,步骤与方法,治疗方法,优点,操作方法,与传统放疗照射的区别,新辅助化疗后接受手术的时机,风险,目的,效果,术前同步放化疗的手术时机,术后护理,作用,优势,原理,前检查及注意事项,效果及安全性,不良反应,方法,检查前准备工作,检查后注意事项,种类,步骤,费用,适用部位,常见疾病,先进技术应用,原则,苯二氮䓬类镇静催眠药,临床常见的误诊,误诊漏诊防范,防治要点,结核菌的传播途径,肺结核,结核病疫情回升的主要原因,我国结核病防治,必要性,预防方法,耳鸣病因,复发,抗体与Ig,Ig的功能,早餐营养学,咳嗽的机制,疾病状态下的咳嗽反射,睡眠解析,养护方法,降压食品,保健抗病功效,具体方法,常见病,家庭小药箱,判断方法,肺癌分类,临床病理类型,脑损伤过程,心肺复苏开始宜早,肺癌传统疗法,体温测量与正常波动,发热,TNM分期系统,打鼾原因,打鼾的影响,治疗模式,肿瘤个性化治疗的核心,综合治疗,致病性,抵抗力,传染性,传播方式,痰成分,痰液处理,“十字”原则的内容,套筒冠义齿,不良反应及对策,误区原因与纠正,肝功能评价指标,缺牙的弊端,缺牙修复,饮食防癌方法,梦游的特点,膳食纤维与健康,二者比较,评判标准,测定体重时的注意事项,营养支持的作用,营养支持的方式,营养素与癌症,抗癌食品,原则内容,缺水可引致疾病,水的功能与喝水,分类及分期,原因,常规临床应用项目,核医学的优势,应用,放射性核素治疗与放疗的区别,安全性,诊断与鉴别诊断,放疗历史,类型,目前较完善和实用的放疗技术,具体内容,意义,病因及常见疾病,睡眠异常与疾病,检测指标,肥胖判定,减肥要科学,预防糖尿病的五个要点,基本原则,五项达标,重建方法,导致发音失败的原因及处理方法,训练方法,部位及范围,术前准备,术后注意事项,常选用的血管,手术方法,术后注意,与疾病的关系,诱发因素,护理调适,特点,来源,临床意义,所致疾病,危害,防治,身高的影响因素,矮小症的治疗,误区,重要性,婴幼儿现场心肺复苏,改善方法,紧急处理,操作步骤,按压技术要求,家庭处理,原因及处理,特色,检查须知,操作要点,常见疾病及症状,抢救,医院救治,方法步骤与要求,心脑血管疾病的预防,人体系统,人体的组成成分,饮食及注意事项,结构与功能,脂肪的功效,脂肪与疾病,能量的用途与需要,中西医结合治疗,饮食调护,施治要点,辩证施治,导致腹痛的疾病,致病原理,可致疾病,预处理及回输,展望,常见过敏性疾病,效果评价,常见并发症,形态,生活史,异常变化相关因素,窦性心动过缓的原因,日常保健,诊断依据,预后转归,预防调摄,并发症和后遗症,疗效判定,辨证要点,实验室检查,并发症的治疗,适应症,脊柱骨折分类,脊柱脊髓损伤评分系统,脊髓损伤的分级,证候及辨证施治,手术指征,手术程序,术后处理,诊断和鉴别诊断,锌的作用,膳食指导,并发证,证候分析,预防调护,审证要点,预防保健,治疗要点.
  2. Mapping between Question type and Answer type  
     Because answer types are more than question types, so the project provide a Suggestor based of the synonym to mapping them. 314 answer types are mapping to 9 question types , for example, answer type “临床表现” to question type “症状”.

1. **System Flow Charty**



1. **Installing and Deployments**
   1. Requirements
      1. JDK1.7
      2. Question type training dataset(83.txt)
      3. Jetty
      4. Lucene
      5. Networking&Access
   2. Solr
      1. Unzip solr-5.3.0
      2. Check core solr-qa’s conf,such as schema.xml make sure it MATCH the answer data source
      3. Start solr service and check $IP:8983/solr/
      4. Copy the field INDEXER Analyzer.jar to $solr/webapps/WEB-INF/lib/
      5. Select the solr-qa core and checkout the data whether integrity



* 1. Dicitionary
     1. All dic should be found at $USER\_HOME/data
     2. Organization机构名词典
     3. Places全国地名大全
     4. Chinese\_Name人名词典
     5. Shanghai\_Places上海地名
     6. Modern Chinese Dictionary现代汉语补充词库
     7. Pinyin拼音转换表
     8. Synonym同义词林中的同义词表(formed as：synonym ID# word1 word2 word3…),all these could accomplish a word net。
  2. Servlets
     1. All rest api are now sealed in mqa.jar
     2. Cp mqa.jar to a java-based OS(Now tested at 121.192.179.234)

All setting up and depolyments step could follow the next command and then ENJOY the mqa:

cd ~

mkdir lib

mkdir solr

mkdir $MQA\_HOME

cd $MQA\_HOME

git clone [git@github.com:changwu/mqa.git](mailto:git@github.com:changwu/mqa.git)

mv solr-5.3.0 ~/solr

mv lib ~/lib

solr/bin/solr start

(wait and check 8983)

export MQASERVICE=com.ibaguo.mqa.services.WebSocketServer

nohup java -Djava.ext.dirs=~/lib/ -classpath mqa.jar $ MQASERVICE >>mqa.log &

iptables -I INPUT -p tcp --dport 8080 -j ACCEPT

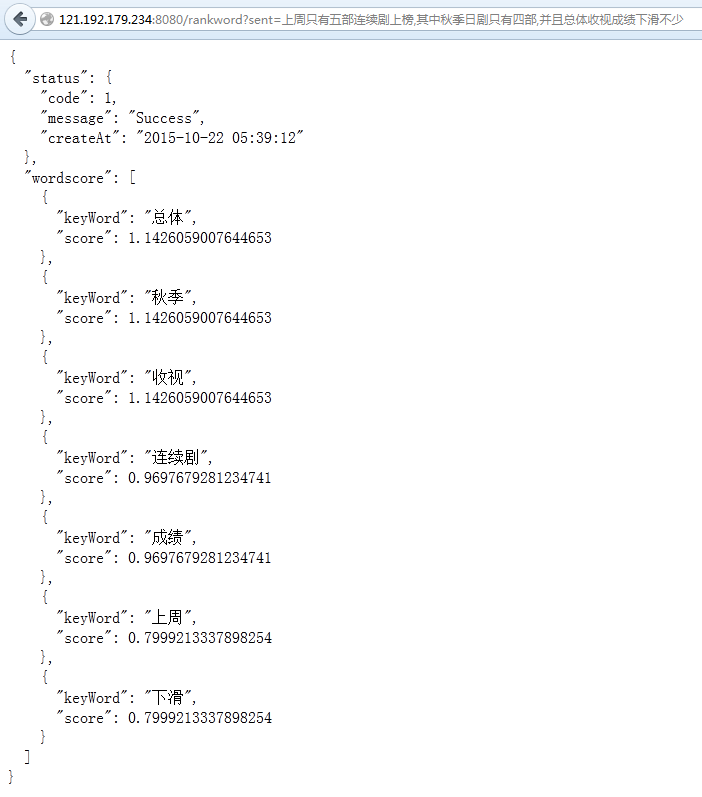
iptables -I INPUT -p tcp --dport 8983 -j ACCEPT

1. **Call Rest Api**
   1. KeyWord(Ranker,Extract)

A sentence could be seg by ibaguo.segment,ibaguo.nlp will give a edge between two word in a sentence determined by both words’ connection in nature lang. And a word rank says the word’s importance in the sentence.

Java src ref TextRankKeyword. rankWordScore.

Rest api named rankword, the param is “sent”，the type of the param value should be a text;eg:



* 1. Question Type Classification

User can obtain the question type by using this api.8 type available by far;ibaguo use a MaxEnt model to train the dataset,the return value contains each type’s score.

Java src ref MaxEnt. Train

Rest api named type,the param is “kwList”,the type of the param value should be a cooma-seperated word in ONE text;eg:

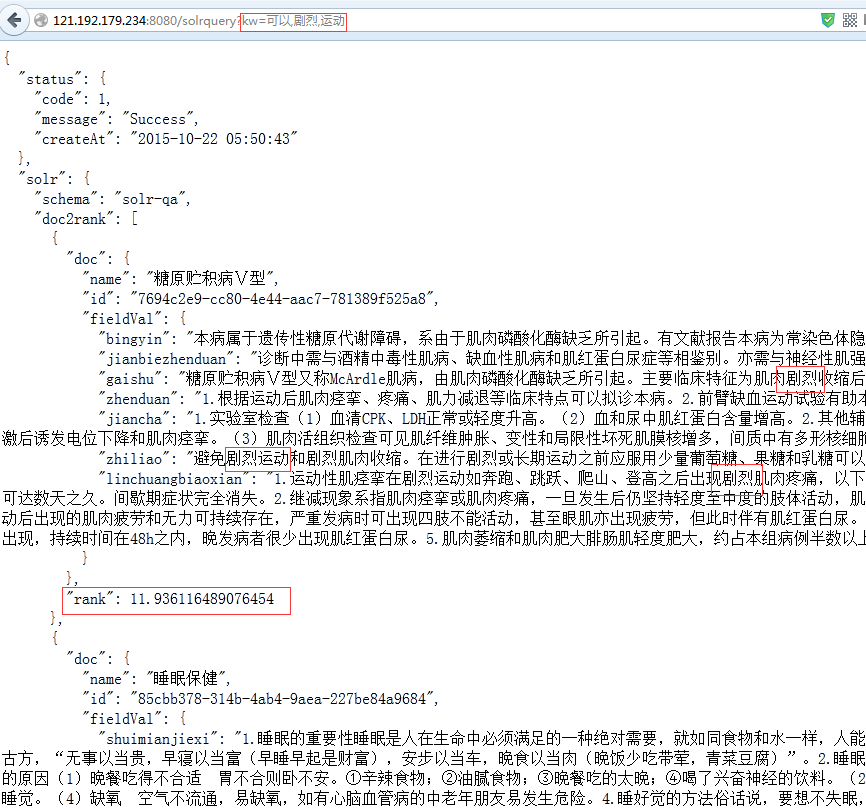


* 1. Search(Query)

Implemented base solr query,search key word list in solr-qa,those setting like highlight, field, sort are hard coded by now .

Java src ref SolrConnector. Search。

Rest api named solrquery, the param is “kwList”,the type of the param value should be a cooma-seperated word in ONE text;eg:



The rank value means each keyword in every doc fields related value and summed, normalized.

* 1. Ask

According to the flow chart, all these function above are bind here. After solrquery, ibaguo provide a Suggestor, base the synonym。There are 300+ fields in the answer dataset, need to mapping to 8 types ,it is a key function to make ”临床表现” to ”症状”。

Rest api named ask, the param is “sent”，the type of the param value should be a text;eg:

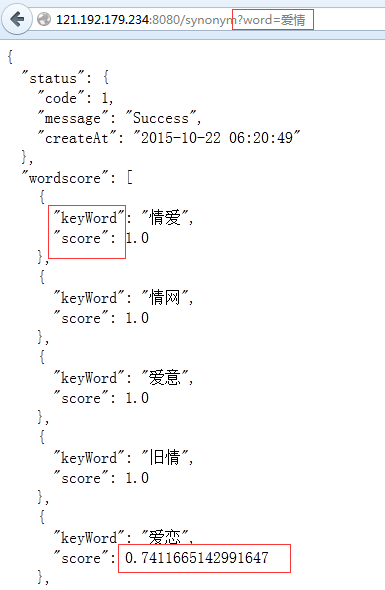


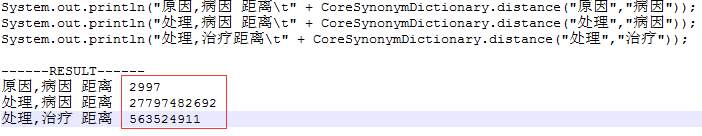
PS: Since a filter,other field value won’t be shown(last example out of “治疗”),there are some difference between search and ask.So ibaguo will test and decide which way to go in classification-after-search or search-after-classification.

* 1. Synonym

Java src ref CoreSynonymDictionary. Distance。

Rest api named synonym the param is “word”,the type of the param value should be a single word;eg:



Load a WordScore as the output,real mean synonymword and distance,return all word under the synonymID,but the distance what mqa used is like：

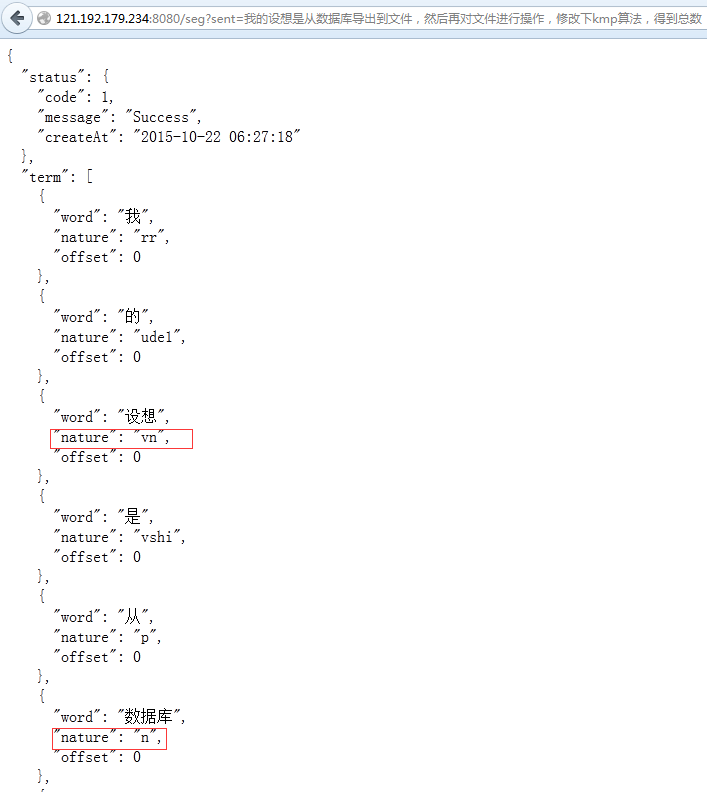
* 1. Segment

Extract the Terms in a sentence.

Java src ref Segment.seg

Segment can set out profile Chinese-Name,Places and stopwords。

Rest api named seg ,the param is “sent”，the type of the param value should be a text;eg:



1. **Secondary Dev by Implements MQA Interfaces**

In order to facilitate the Secondary Dev, all servlets are interfaced as follow:

AnswerSearcher

List<DocRank> search(String[] kw);

KeyWordExtract

List<String> extractKeyword(String document, int size);

List<KeywordScore> rankWordScore(String document); WordSynonym

QuestionClassifier

save(String path);//save a classification model

QuestionClassifier initFromFile(String path);//load a mode from file

rain(**int** maxIt);

String eval(List<String> fieldList);//eval and predict

Map<String, Double> predict(List<String> fieldList);//each lables’ predict score

QuestionToAnswer

List<Doc> makeQa(String q);//return the answer