数据库系统概论新技术篇

文本大数据分析及应用案例

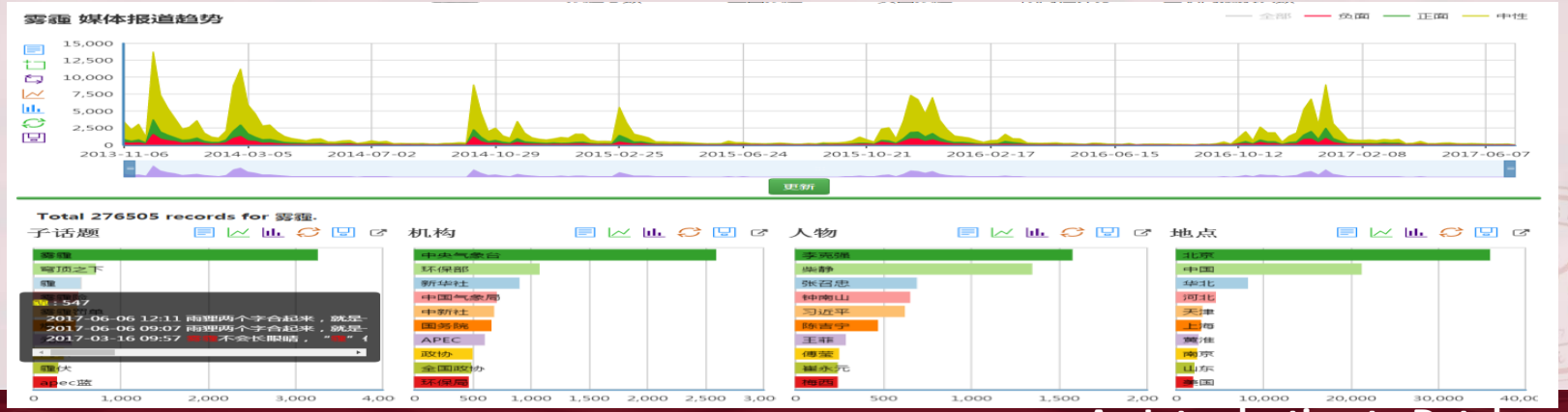
窦志成

中国人民大学信息学院

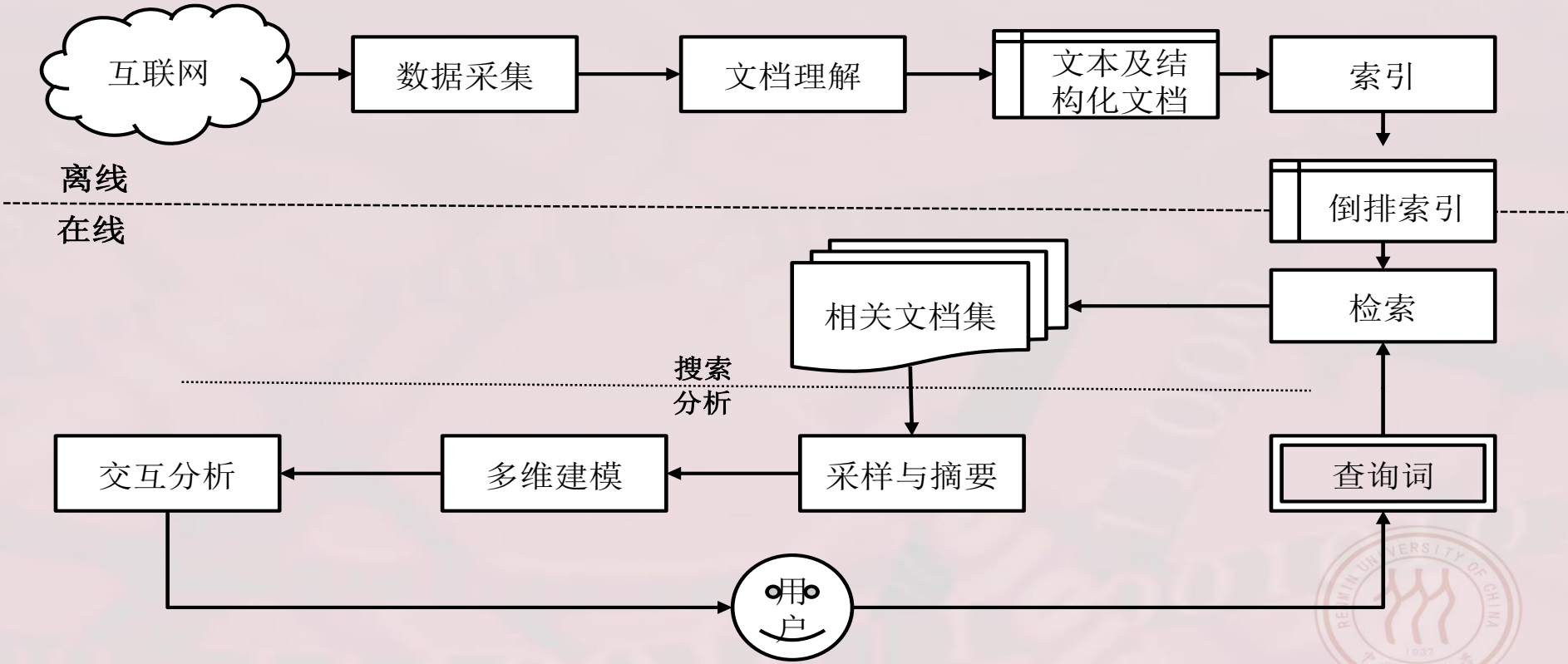
2017年7月

文本大数据分析引擎

- 核心特点
  - 搜索->分析 （OLTP ->OLAP）
  - 简单结果列表->高阶知识统计结果（结果集->立方体）
  - 交互式：在文本数据上进行类似于切片和切块的交互分析功能



文本大数据分析引擎 - 系统构架



文本大数据分析及应用案例

- 课程内容
  - 交互式文本大数据分析系统：时事探针
  - 文本处理：自然语言处理与文本挖掘基础算法
  - 文本搜索、文本分析系统构建

文本处理-正文抽取

- 给一个HTML网页，抽取该网页的标题、正文、图片、来源、时间等信息
- 方法：
  - 基于监督学习的方法：CRF等
  - 基于规则的方法：字体大小、位置、布局、CSS等

http://info.ruc.edu.cn/news_convert_detail.ph

中国人民大学信息学院
SCHOOL OF INFORMATION RENMIN UNIVERSITY OF CHINA

首 页 | 学院概况 | 新闻公告 | 人才培养 | 师资科研 | 学生发展

新闻公告

- 学院新闻
- 学院公告
- 信息相册
- 新闻人物
- 信息影像

您所在的位置：新闻公告 > 学院新闻

“明德图灵”厚重人才成长支持计划启动仪

发布时间：2016-10-25 07:15 浏览量：927 来源：摄影/新闻 分团委

2016年10 月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼4

任、信息学院院长文继荣教授，学生处陈虹百副处长，信息学院党委副书记张国富

授、范举副教授等以及参与项目的全体同学出席此次会议。

项目执行委员会主任文继荣教授致辞。他分析了大数据、计算机技术的广泛应

域人才的重要性，强调了明德图灵厚重人才培养项目的意义。他指出指出，“明德

目，而其人才培养目标、培养方式与培养设想也符合中国人民大学新型人才培养趋

望同学们珍惜机会，努力培养专业能力，能真正成为“厚重”人才。

陈虹百副处长表示，“明德图灵”项目从全校各院系、年级层层选拔出对计算

学，希望学员们能够主动寻求提升能力的途径，积极探索发现，突破自我，向“厚

“明德图灵”厚重人才成长支持计划启动仪式顺利举行

2016年10 月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼417会议室顺利举行。项目执行委员会主任、信息学院院长文继荣教授，学生处陈虹百副处长，信息学院党委副书记张国富，项目导师窦志成副教授、陈跃国副教授、范举副教授等以及参与项目的全体同学出席此次会议。

项目执行委员会主任文继荣教授致辞。他分析了大数据、计算机技术的广泛应用与发展前景，强调培养优秀计算机领域人才的重要性，强调了明德图灵厚重人才培养项目的意义。他指出指出，“明德图灵”项目是我院大胆创新的试点项目，而其人才培养目标、培养方式与培养设想也符合中国人民大学新型人才培养趋势，因此文院长对同学们寄予厚望，希望同学们珍惜机会，努力培养专业能力，能真正成为“厚重”人才。

陈虹百副处长表示，“明德图灵”项目从全校各院系、年级层层选拔出对计算机专业怀有浓厚兴趣且能力出色的同学，希望学员们能够主动寻求提升能力的途径，积极探索发现，突破自我，向“厚重人才”的目标努力。同时，陈老师表示学生处将持续对本项目给予大力

抽取

VOW64; Trident/6.0)

于信息楼417会议室顺利举行。项目执行委员会主任、信息学院院长文继荣教
志成副教授、陈跃国副教授、范举副教授等以及参与项目的全体同学出席此次
机技术的广泛应用与发展前景，强调培养优秀计算机领域人才的重要性，强调
我院大胆创新的试点项目，而其人才培养目标、培养方式与培养设想也符合中
同学们珍惜机会，努力培养专业能力，能真正成为“厚重”人才。 陈虹百副处长
有浓厚兴趣且能力出色的同学，希望学员们能够主动寻求提升能力的途径，积
学生处将持续对本项目给予大力支持，并与文继荣教授一同为本项目启动仪式
选择的特聘导师，是信息学院计算机领域最年轻有为的导师团队，希望导师
责。 导师代表窦志成副教授发言，欢迎各位同学加入“明德图灵”项目，并解释
通过该项目，专项培养一批优秀的计算机人才，开拓项目成员的国际视野，引
张国富副书记为同学们颁发荣誉证书，鼓励学员们挑战自我，在即将到来的技
表吴紫君同学代表全体学员发言。她表示，期待在未来的日子里和同学们共同
员刘玲初介绍了北美数学建模竞赛的情况，她希望项目组同学们再接再厉，跟
交流，并一同观看计算机之父艾伦·图灵的传记片«模仿游戏»。 明德图灵“厚重
是在学校学生处支持下，信息学院结合学科特点和专业特色，优化整合现有的
了积极尝试。该项目以“家国情怀，文理交融”为理念，把国情教育、专业知识
学院学科优势和学校社会科学的特长，全面提升学生的综合素质。 该项目为
同学们培养专业研究能力，接触世界前沿领域，成为“厚重”的计算机、数学人
研究与世界领先水平接轨，强化我院教育与研究实力。

文本处理-中文分词

❖将中文字串切分成有意义的单词

2016年10月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼417会议室顺利举行。项目执行委员会主任、信息学院院长文继荣教授，学生处陈虹百副处长，信息学院党委副书记张国富，项目导师窦志成副教授、陈跃国副教授、范举副教授等以及参与项目的全体同学出席此次会议。

❖方法

- 基于匹配的方法，如正向最大匹配法、逆向最大匹配法
- 基于统计和机器学习的方法：HMM, CRF等

文本处理-命名实体抽取

- 识别文本中有意义的命名实体，如人名、地名等

2016年10月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼417会议室顺利举行。项目执行委员会主任、信息学院院长文继荣教授，学生处陈虹百副处长，信息学院党委副书记张国富，项目导师窦志成副教授、陈跃国副教授、范举副教授等以及参与项目的全体同学出席此次会议。

- 方法
  - CRF
  - 深度学习：BI-LSTM+CRF

文本处理-关键词抽取

❖识别文本中最重要的词

“明德图灵”厚重人才成长支持计划启动仪式顺利举行
2016年10月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼417会议室顺利举行。项目执行委员会主任、信息学院院长文继荣教授，学生处陈虹百副处长，信息学院党委副书记张国富，项目导师窦志成副教授、陈跃国副教授、范举副教授等以及参与项目的全体同学出席此次会议。
项目执行委员会主任文继荣教授致辞。他分析了大数据、计算机技术的广泛应用与发展前景，强调培养优秀计算机领域人才的重要性，强调了明德图灵厚重人才培养项目的意义。他指出指出，“明德图灵”项目是我院大胆创新的试点项目，而其人才培养目标、培养方式与培养设想也符合中国人民大学新型人才培养趋势，因此文院长对同学们寄予厚望，希望同学们珍惜机会，努力培养专业能力，能真正成为“厚重”人才。
陈虹百副处长表示，“明德图灵”项目从全校各院系、年级层层选拔出对计算机专业怀有浓厚兴趣且能力出色的同学，希望学员们能够主动寻求提升能力的途径，积极探索发现，突破自我，向“厚重人才”的目标努力。同时，陈老师表示学生处将持续对本项目给予大力支持，并与文继荣教授一同为本项目启动仪式授旗。
文继荣教授为特聘导师颁发聘书，他表示，本次“明德图灵”计划选择的特聘导师，是信息学院计算机领域最年轻有为的导师团队，希望导师们能够带领学员们在科研创新、学科竞赛等多项活动中取得优异的成绩。
导师代表窦志成副教授发言，欢迎各位同学加入“明德图灵”项目，并解释了本项目的设立初衷及导师团队规划的成员培养方式。他表示，希望通过该项目，专项培养一批优秀的计算机人才，开拓项目成员。。。

关键词

培养/0.4405933969127764　人才/0.38184761065773953　计算机/0.26435603814766584
同学们/0.2349831450201474　图灵/0.2349831450201474　导师/0.2349831450201474
项目/0.22794539914073708　能力/0.205610251892629　明德/0.20450457180343978
学员/0.17623735876511057　厚重/0.14686903784739558　教育/0.14686446563759212
希望/0.14686446563759212　领域/0.1174915725100737　我院/0.1174915725100737
水平/0.1174915725100737　副教授/0.1174915725100737　同学/0.1174915725100737
信息学院/0.1174915725100737　专业/0.09129973647243243

❖方法

- TF-IDF
- TextRank

文本处理-情感分类

- 识别文本中最重要的词

“明德图灵”厚重人才成长支持计划启动仪式顺利举行
2016年10月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼417会议室顺利举行。项目执行委员会主任、信息学院院长文继荣教授，学生处陈虹百副处长，信息学院党委副书记张国富，项目导师窦志成副教授、陈跃国副教授、范举副教授等以及参与项目的全体同学出席此次会议。
项目执行委员会主任文继荣教授致辞。他分析了大数据、计算机技术的广泛应用与发展前景，强调培养优秀计算机领域人才的重要性，强调了明德图灵厚重人才培养项目的意义。他指出指出，“明德图灵”项目是我院大胆创新的试点项目，而其人才培养目标、培养方式与培养设想也符合中国人民大学新型人才培养趋势，因此文院长对同学们寄予厚望，希望同学们珍惜机会，努力培养专业能力，能真正成为“厚重”人才。
陈虹百副处长表示，“明德图灵”项目从全校各院系、年级层层选拔出对计算机专业怀有浓厚兴趣且能力出色的同学，希望学员们能够主动寻求提升能力的途径，积极探索发现，突破自我，向“厚重人才”的目标努力。同时，陈老师表示学生处将持续对本项目给予大力支持，并与文继荣教授一同为本项目启动仪式授旗。
文继荣教授为特聘导师颁发聘书，他表示，本次“明德图灵”计划选择的特聘导师，是信息学院计算机领域最年轻有为的导师团队，希望导师们能够带领学员们在科研创新、学科竞赛等多项活动中取得优异的成绩。
导师代表窦志成副教授发言，欢迎各位同学加入“明德图灵”项目，并解释了本项目的设立初衷及导师团队规划的成员培养方式。他表示，希望通过该项目，专项培养一批优秀的计算机人才，开拓项目成员。。。

情感分析

1. 2016年10 月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼417会议室顺利举行 顺利/1
2. 他分析了大数据、计算机技术的广泛应用与发展前景，强调培养优秀计算机领域人才的重要性，强调了明德图灵厚重人才培养项目的意义 优秀/1
3. 他指出指出，“明德图灵”项目是我院大胆创新的试点项目，而其人才培养目标、培养方式与培养设想也符合中国人民大学新型人才培养趋势，因 大胆/1 创新/1 厚望/1 珍惜/1 真正/1
4. 陈虹百副处长表示，“明德图灵”项目从全校各院系、年级层层选拔出对计算机专业怀有浓厚兴趣且能力出色的同学，希望学员们能够主动寻求提 出色/1
5. 文继荣教授为特聘导师颁发聘书，他表示，本次“明德图灵”计划选择的特聘导师，是信息学院计算机领域最年轻有为的导师团队，希望导师们能 创新/1 多项活动中取得优异/1.5
6. 他表示，希望通过该项目，专项培养一批优秀的计算机人才，开拓项目成员的国际视野，引导同学们接触国际前沿领域，与国际领先水平接轨 优秀/1
7. 她表示，期待在未来的日子里和同学们共同相处进步，并表示自己会不负众望，积极学习，潜心探究 期待/1 积极/1 潜心/1

- 方法
  - 基于词典匹配的方法
  - 基于SVM、CNN、LSTM等监督学习的情感分类

文本处理- 其他

- 词性识别
- 词法分析
- 语法分析
- 观点抽取
- 文本分类
- ...

文本处理

“明德图灵”厚重人才成长支持计划启动仪式顺利举行2016年10月22日，“明德图灵”厚重人才成长支持计划启动仪式于信息楼417会议室顺利举行。项目执行委员会主任、信息学院院长文继荣教授，学生处陈虹百副处长，信息学院党委副书记张国富，项目导师窦志成副教授、陈跃国副教授、范举副教授等以及参与项目的全体同学出席此次会议。

项目执行委员会主任文继荣教授致辞。他分析了大数据、计算机技术的广泛应用与发展前景，强调培养优秀计算机领域人才的重要性，强调了明德图灵厚重人才培养项目的意义。他指出指出，“明德图灵”项目是我院大胆创新的试点项目，而其人才培养目标、培养方式与培养设想也符合中国人民大学新型人才培养趋势，因此文院长对同学们寄予厚望，希望同学们珍惜机会，努力培养专业能力，能真正成为“厚重”人才。

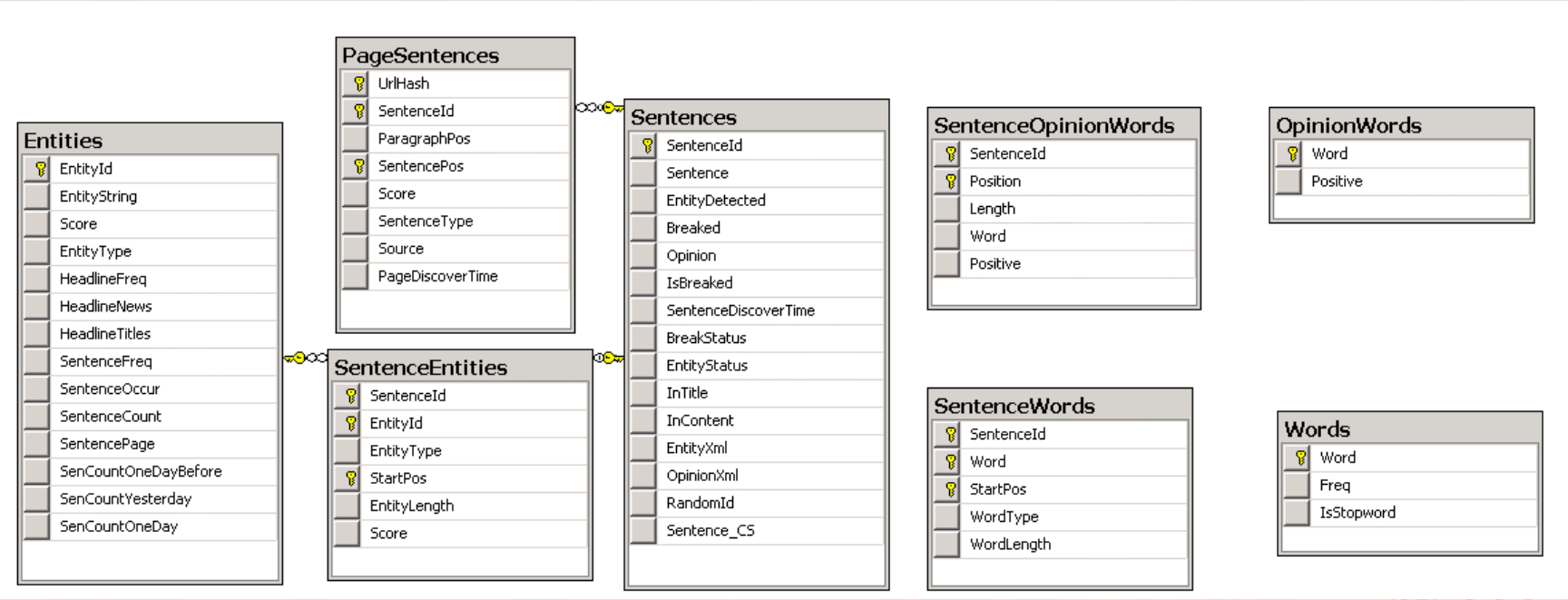
陈虹百副处长表示，“明德图灵”项目从全校各院系、年级层层选拔出对计算机专业怀有浓厚兴趣且能力出色的同学，希望学员们能够主动寻求提升能力的途径，积极探索发现，突破自我，向“厚重人才”的目标努力。同时，陈老师表示学生处将持续对本项目给予大力支持，并与文继荣教授一同为本项目启动仪式授旗。

文继荣教授为特聘导师颁发聘书，他表示，本次“明德图灵”计划选择的特聘导师，是信息学院计算机领域最年轻有为的导师团队，希望导师们能够带领学员们在科研创新、学科竞赛等多项活动中取得优异的成绩。

导师代表窦志成副教授发言，欢迎各位同学加入“明德图灵”项目，并解释了本项目的设立初衷及导师团队规划的成员培养方式。他表示，希望通过该项目，专项培养一批优秀的计算机人才，开拓项目成员。。。

Organization	**Key**	**Value**
	项目执行委员会	2
	信息学院	2
	信息学院党委	1
	中国人民大学	1
Person	**Key**	**Value**
	文继荣	4
	陈虹百	2
	张国富	2
	窦志成	2
	陈跃国	1
	陈百	1
	吴紫君	1
	刘玲初	1
	艾伦·图灵	1
Idiom	**Key**	**Value**
	广泛应用	1
	发展前景	1
	人才培养	5
	寄予厚望	1
	各院系	1
	浓厚兴趣	1
	坚持到底	1
	不负众望	1
	再接再厉	1
	骄人成绩	1

文本数据的结构化存储


Entities
EntityId
EntityString
Score
EntityType
HeadlineFreq
HeadlineNews
HeadlineTitles
SentenceFreq
SentenceOccur
SentenceCount
SentencePage
SenCountOneDayBefore
SenCountYesterday
SenCountOneDay
PageSentences
UrlHash
SentenceId
ParagraphPos
SentencePos
Score
SentenceType
Source
PageDiscoverTime
SentenceEntities
SentenceId
EntityId
EntityType
StartPos
EntityLength
Score
Sentences
SentenceId
Sentence
EntityDetected
Breaked
Opinion
IsBreaked
SentenceDiscoverTime
BreakStatus
EntityStatus
InTitle
InContent
EntityXml
OpinionXml
RandomId
Sentence_CS
SentenceOpinionWords
SentenceId
Position
Length
Word
Positive
OpinionWords
Word
Positive
SentenceWords
SentenceId
Word
StartPos
WordType
WordLength
Words
Word
Freq
IsStopword

```
db.getCollection('Ruc_index').find({})
```

Ruc_index 0.197 sec. 0 50

	title	content	user	userId	source	topics	lan	time	fetchTime	paragraphs	docDatas
1	中国人民...	12月28日...	null	null	文/校报 学...	[1 eleme...	null	2016-12-...	-6213559...	[22 elem...	[0 eleme...
2	中国人民...	中国人民...	null	null	中新网	[1 eleme...	null	2016-12-...	-6213559...	[6 eleme...	[0 eleme...
3	中国人民...	北京科技...	null	null	中国教育...	[1 eleme...	null	1978-12-...	-6213559...	[21 elem...	[0 eleme...
4	深圳高等...	好消息，...	null	null		[1 eleme...	null	2016-12-...	-6213559...	[34 elem...	[0 eleme...
5	推进大学...	《2016中国	null	null	中国网	[1 eleme...	null	2016-12-...	-6213559...	[11 elem...	[0 eleme...
6	中国人民...	人民网北...	null	null	人民网-教...	[1 eleme...	null	2016-12-...	-6213559...	[9 eleme...	[1 eleme...
7	中国人民...	12月27日...	null	null	人大新闻网	[1 eleme...	null	2016-12-...	-6213559...	[5 eleme...	[0 eleme...
8	News	峥嵘八秩...	null	null	离退休工...	[1 eleme...	null	2016-12-...	-6213559...	[6 eleme...	[0 eleme...
9	中国人民...	中国网财...	null	null	中国网	[1 eleme...	null	2016-12-...	-6213559...	[6 eleme...	[1 eleme...
10	中国人民...	28日上午，	null	null	京华时报(...	[1 eleme...	null	2016-12-...	-6213559...	[8 eleme...	[0 eleme...
11	人民大学...	中国 人民...	null	null	大公网	[1 eleme...	null	2016-12-...	-6213559...	[3 eleme...	[1 eleme...
12	人民大学...	法制晚报...	null	null	法制晚报	[1 eleme...	null	2016-12-...	-6213559...	[10 elem...	[0 eleme...
13	中国人民...	新华网...	null	null		[1 eleme...	null	1978-12-...	-6213559...	[8 eleme...	[0 eleme...
14	中国人民...	[摘要]深...	null	null	晶报	[1 eleme...	null	2016-12-...	-6213559...	[5 eleme...	[0 eleme...
15	今后在深...	深圳市政府	null	null	深圳晚报	[1 eleme...	null	2016-12-...	-6213559...	[11 elem...	[1 eleme...
16	深圳与中...	深圳特区...	null	null	深圳特区报	[1 eleme...	null	2016-12-...	-6213559...	[7 eleme...	[1 eleme...
17	深圳又双...	（原标题...	null	null	金羊网	[1 eleme...	null	2016-12-...	-6213559...	[27 elem...	[1 eleme...
18	中国人民...	据报告调查	null	null	新华网	[1 eleme...	null	2016-12-...	-6213559...	[5 eleme...	[0 eleme...
19	中国人民...	本报北京	null	null	人民日报	[1 eleme...	null	2016-12-...	-6213559...	[4 eleme...	[1 eleme...

文本大数据分析引擎 - 系统构架