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

- 文本是信息表达的主要方式之一
- 互联网上存在着大量的文本数据
  - 网页
  - 新闻
  - 论坛
  - 社交媒体（微博、微信）
  - 评论（新闻评论、购物评论等）
- 海量的文本数据中蕴含着丰富的价值，对文本大数据的分析和挖掘具有重要意义
  - 舆情监控、商业智能、趋势预测、精准营销等

文本数据的特点

- 信息蕴含在自由文本中
- 没有结构化字段可供查询
- 无法直接进行统计分析

10月22日，“明德图灵”厚重人才成长支持计划启动仪式于中国人民大学信息楼举行。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参加了此次会议。

项目执行委员会主任文继荣致辞。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和明德图灵厚重人才培养项目的意义。“明德图灵”项目是信息学院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，他对同学们寄予厚望，希望同学们珍惜机会，努力培养专业能力，能真正成为的“厚重”人才。

文本大数据分析

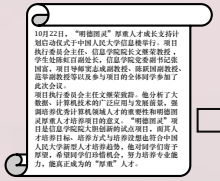
❖ 分析前，需要先进行文本的处理和挖掘

10月22日，“明德图灵”厚重人才成长支持计划启动仪式于中国人民大学信息楼举行。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参加了此次会议。
项目执行委员会主任文继荣致辞。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和明德图灵厚重人才培养项目的意义。“明德图灵”项目是信息学院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，他对同学们寄予厚望，希望同学们珍惜机会，努力培养专业能力，能真正成为的“厚重”人才。

非结构化
文本数据

✓ 自然语言处理
✓ 文本挖掘

从非结构化的文本数据中挖掘出结构化信息以供分析



文本数据

搜索

结构化信息

分析

字段	值
人名	文继荣，陈虹百，张国富，窦志成，陈跃国，范举
地名	中国人民大学信息楼
机构名	中国人民大学，信息学院
关键词	明德图灵，人才，启动仪式，会议
主题	…

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
  - 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