第一讲 - 文本分析与数据挖掘概论

张建章

阿里巴巴商学院
杭州师范大学

2023-02-22

目录

目录

课程考核说明

根据教学大纲要求，本课程的考核办法为：

$$\begin{aligned} \text{总成绩} = \text{期末成绩} \times 50\% \ &+ \text{日常作业} \times 30\% \\ + \text{日常考勤} \times 10\% \ &+ \text{课堂表现} \times 10\% \end{aligned}$$

其中，期末考试采用**项目实验作业**形式。

课程简介

课程名称：《文本分析与数据挖掘》

课程目标：

① 理解经典的文本数据分析方法；
② 了解最新的文本数据分析方法；
③ 掌握文本数据分析实验方法；
④ 培养数据驱动的商务计算思维；
⑤ 应用文本挖掘方法高效解决商务分析问题。

授课方式： 课堂讲授 + 实践案例

实验环境

编程语言： Python 3.X

开发环境： Pycharm + Anaconda

交互环境： Jupyter-lab (Anaconda 已内置)

常用软件包： NLTK，scikit-learn，pandas，numpy，matplotlib，上述软件包 Anaconda 均已内置，MLxtend，huggingface，需要通过 pip 命令自行安装。

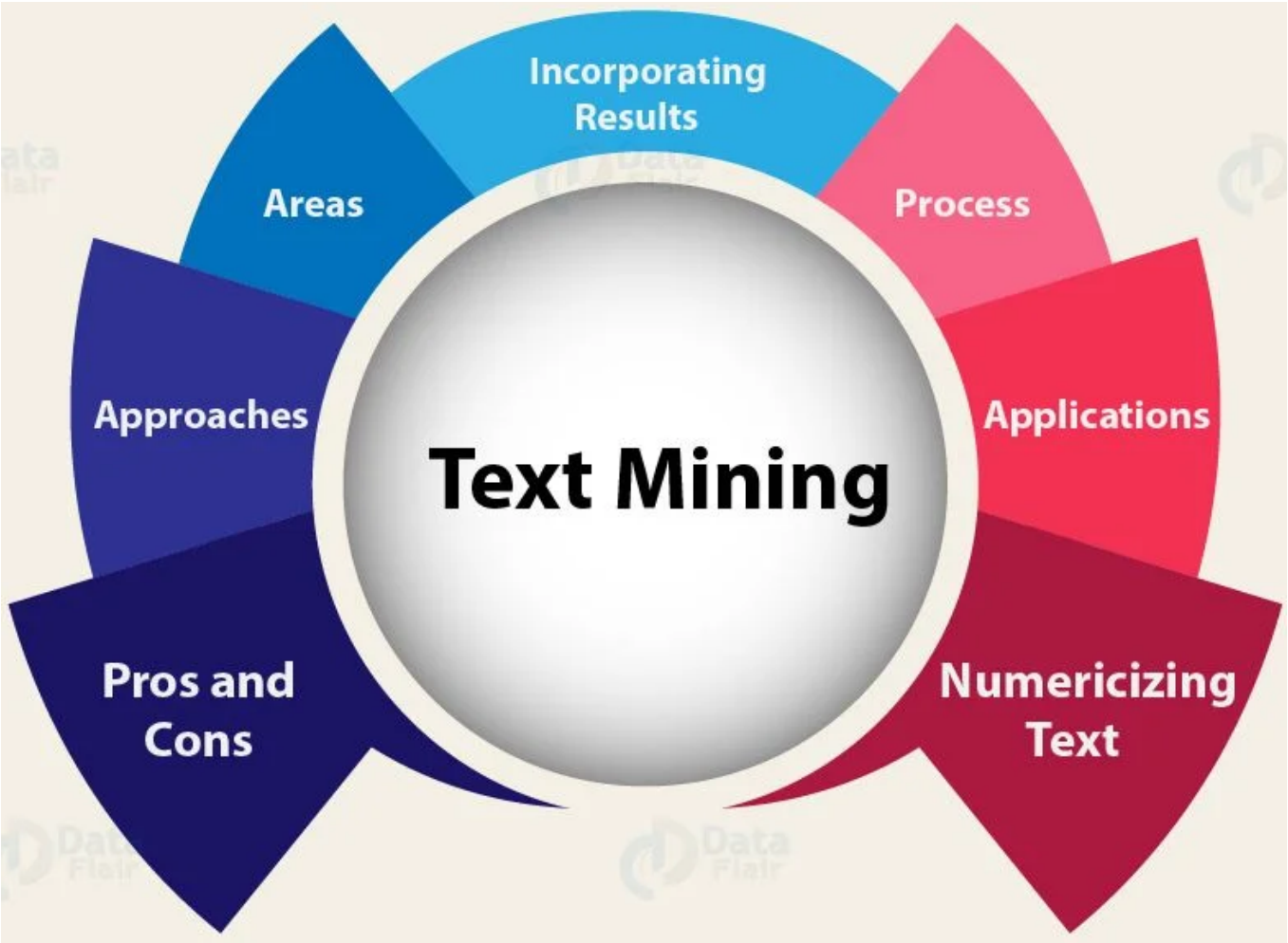
操作系统： Linux 桌面版 (推荐)，Windows，Mac OS (推荐)

在线环境： Kaggle (推荐)，Google colab

学习资源： Kaggle (推荐)，Towards Data Science (推荐)，Stack Overflow (推荐)，Github，CSDN，阿里云天池

目录

文本挖掘 (Text Mining)：是利用**自然语言处理技术**从大量文本数据中提取有价值信息的过程，如提取关键信息、识别主题和趋势、发现文本之间的关系等。广泛用于舆情检测、办公自动化、智能助手等多领域。



文本挖掘最新进展

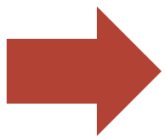
以 BERT、GPT 为代表的超大规模语言模型，训练语料以 TB 计，参数量以十亿计，可支持多种下游任务，在几乎所有自然语言处理任务上取得了突破性进展。

文本挖掘流程 I

1. **文本预处理：**包括文本清洗、分词、词性标注等；

2. **特征提取：**将文本转换为可用于分析的数值特征，如词频、TF-IDF、embedding 向量等；

Vocabulary:
Man, woman, boy, girl, prince, princess, queen, king, monarch



	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets a 1x9 vector representation

文本挖掘流程 II

3. **文本分析：**利用机器学习、统计学等技术对文本数据进行分析，如分类、聚类、情感分析、实体识别等；

> 改了好几次，感觉终于可以确定了。这次的真丝是做了古董感的米金色染色，法蕾也做了同样的颜色。

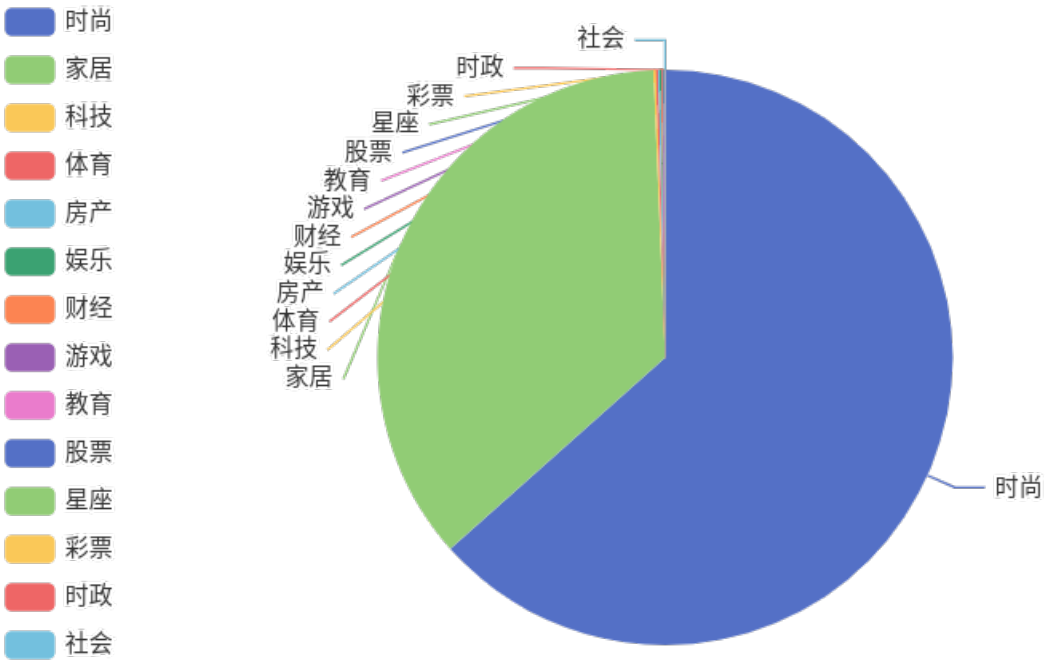


图 1: 文本分类结果

文本挖掘流程 III

4. **可视化展示：**将文本挖掘结果可视化展示，帮助人们更好地理解和使用挖掘结果。

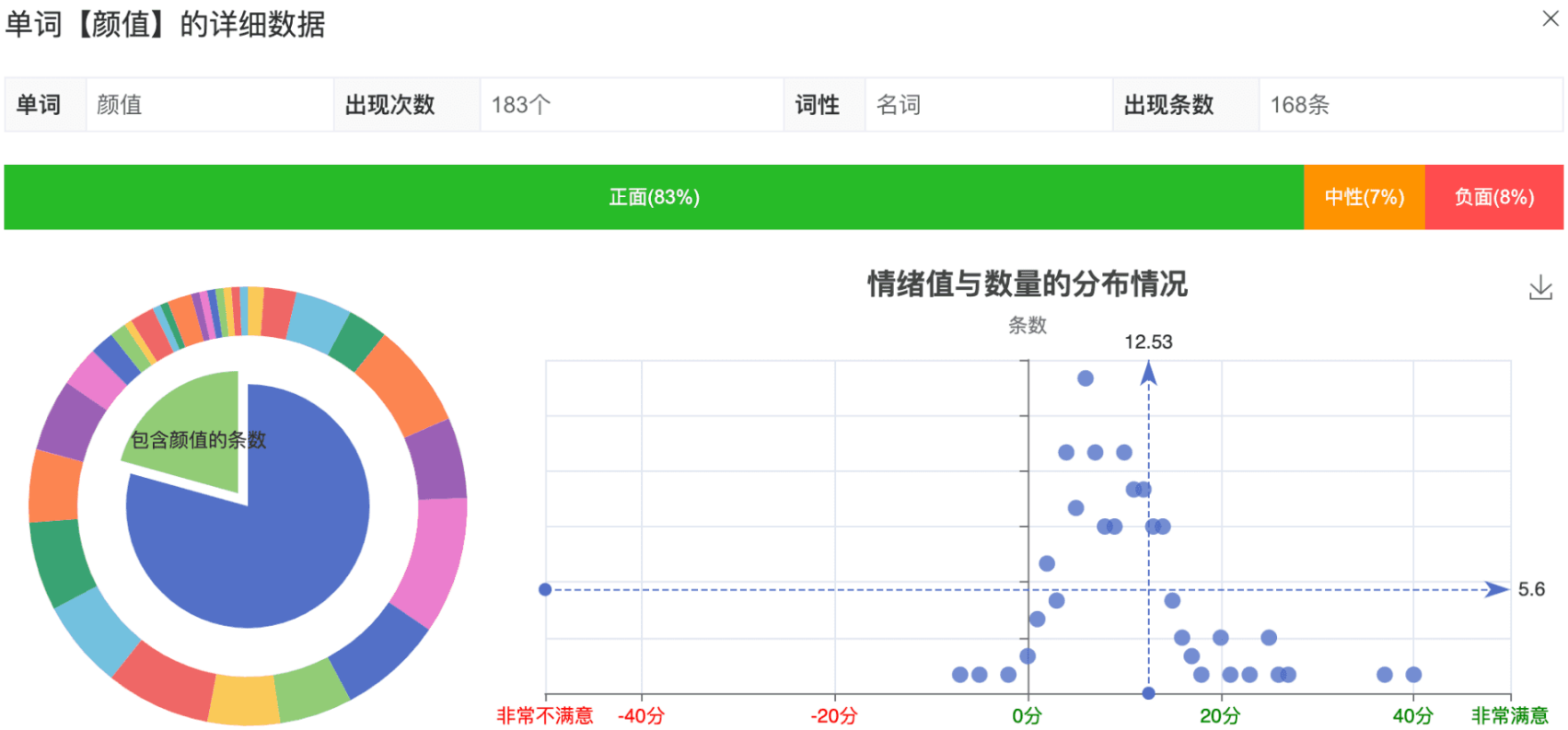
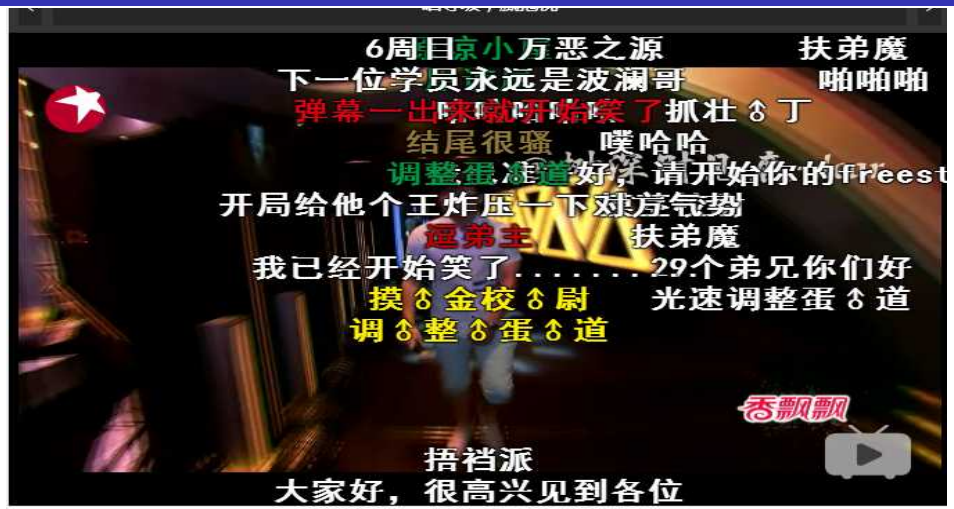


图 2: 文本分析结果可视化示例

目录

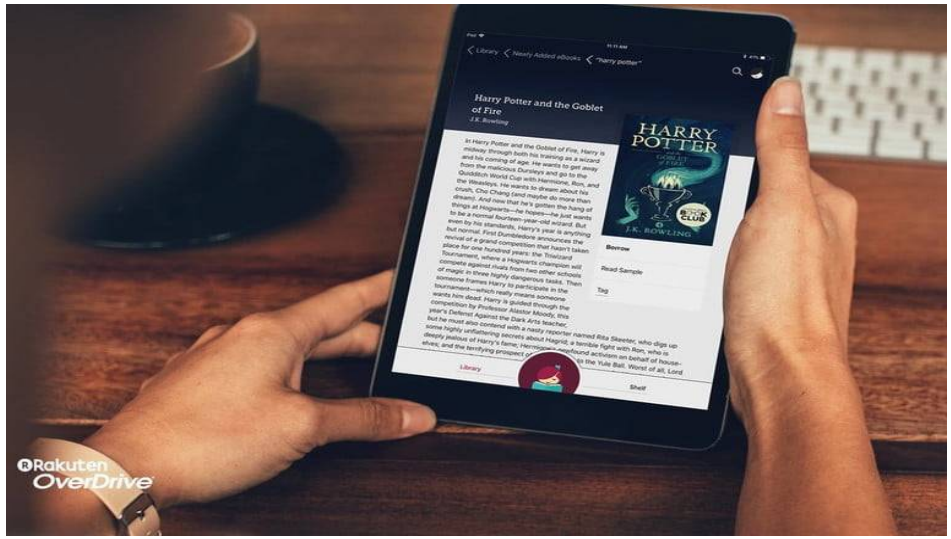
自然语言文本无处不在



视频弹幕

文件报告

新闻媒体



图书资料

自然语言的特点

- **多样性**：多种语言、同一语言中多种方言、专业领域术语；
- **灵活性**：同一含义有不同的表达 (主动句-被动句)，同一表达可表达不同的含义 (如，我不介意)；

示例

你把我灌醉。
我被你灌醉。

- **上下文依赖性**：自然语言的含义往往依赖于上下文，即前后文的语境和背景。

示例

千元智能机就够用了，**小米**不错。
杂粮对身体好，**小米**不错。

- **歧义性：**同一词语或句子可能有多种解释。

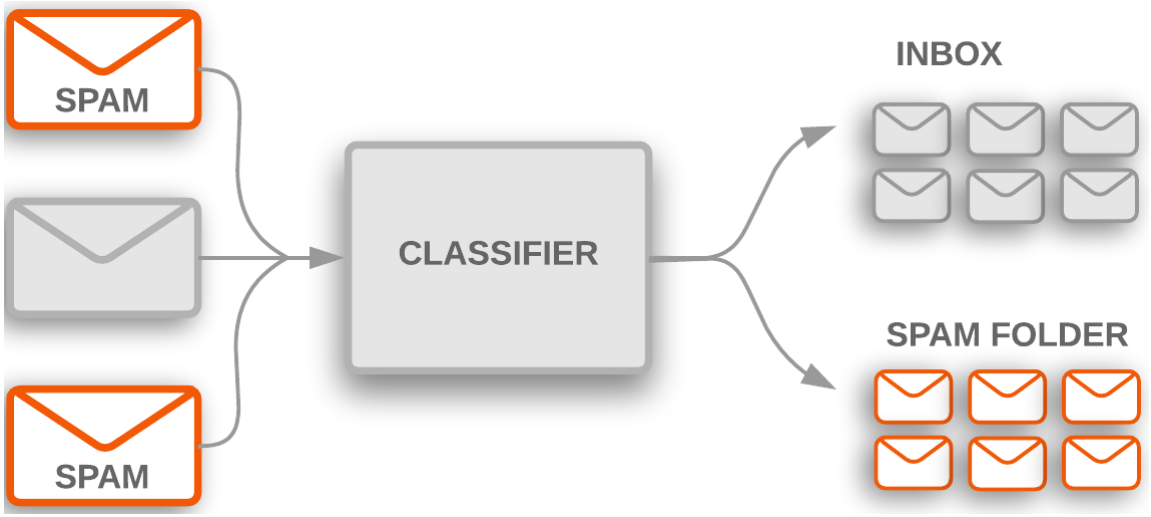
示例

爸爸抱不动儿子了，因为**他**太胖了。
爸爸抱不动儿子了，因为**他**太瘦了。
本届世界杯**中日韩**都进 16 强！他们是亚洲之光。

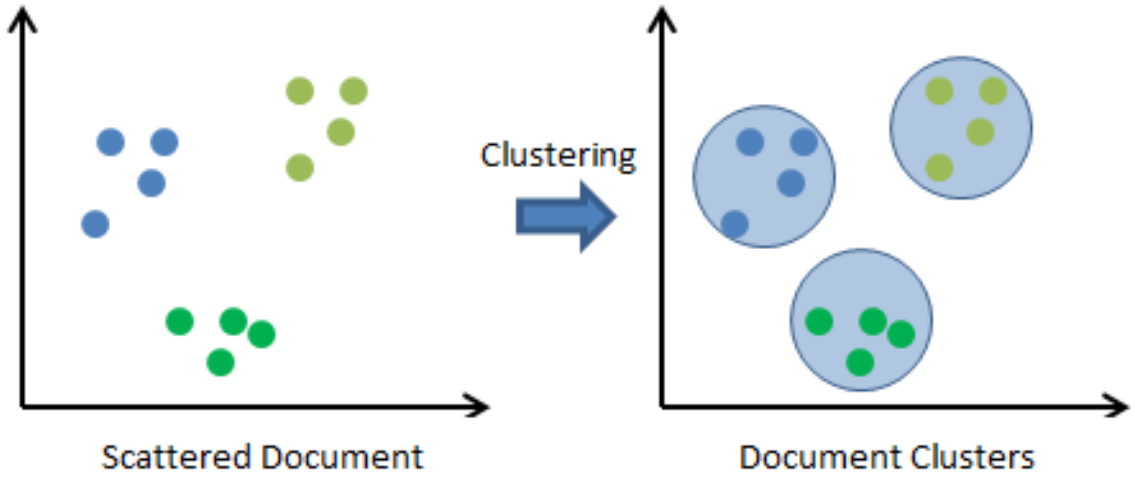
- **错误容忍性：**自然语言容忍语法和拼写错误，人们仍然能够理解其含义。这也是自然语言处理中一个重要的挑战。

目录

- **文本分类**：将给定的文本划分到事先规定的文本类型。



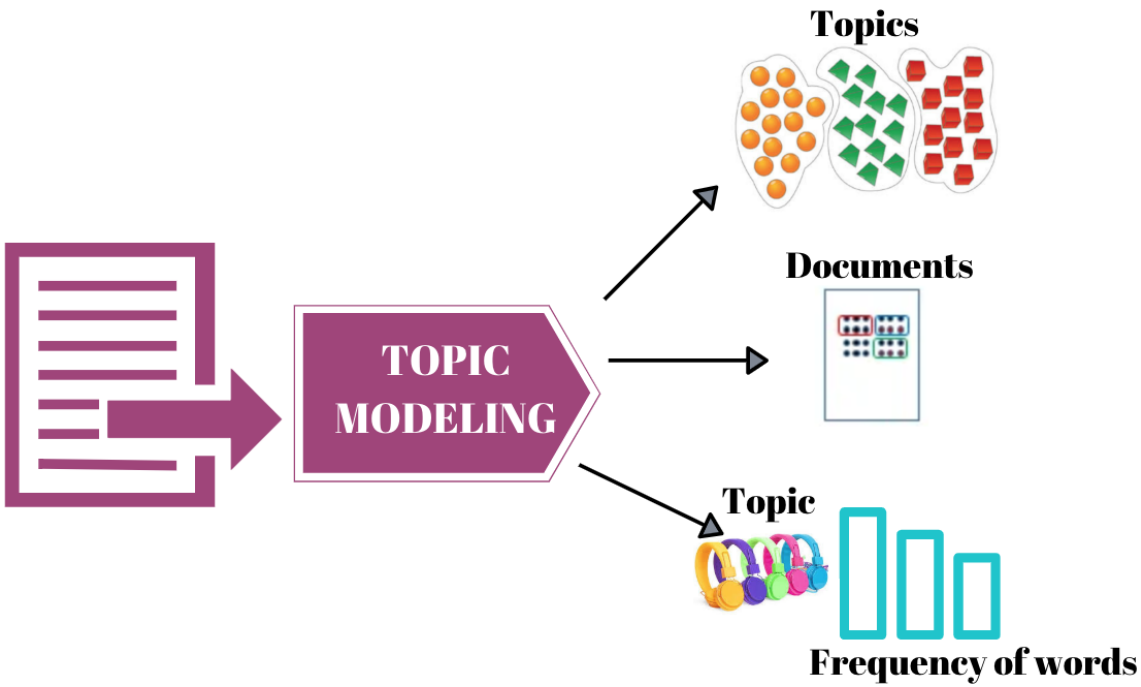
- **文本聚类**：将给定的文本集合划分成不同的类别，通常情况下从不同的角度可以聚类出不同的结果。



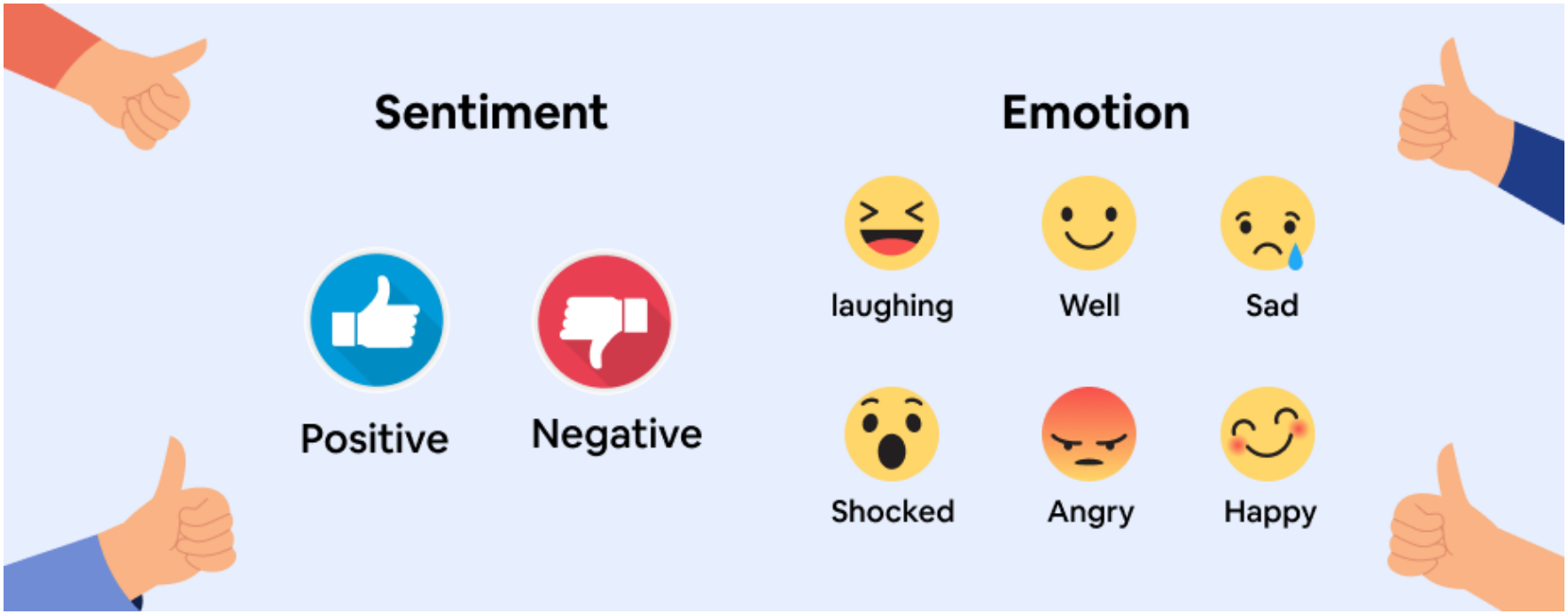
- **主题模型**：通常情况下每一篇文章包含多个主题，而主题可以用一组词汇表示，这些词汇之间有较强的相关性，且其概念和语义基本一致。我们可以认为 (假定) 某个文档以一定概率选择某个主题，某个主题以一定的概率选择某个词汇，如下 (全概率公式)：

$$P(W_j|D_i) = \sum_k P(W_j|T_k)P(T_k|D_i)$$

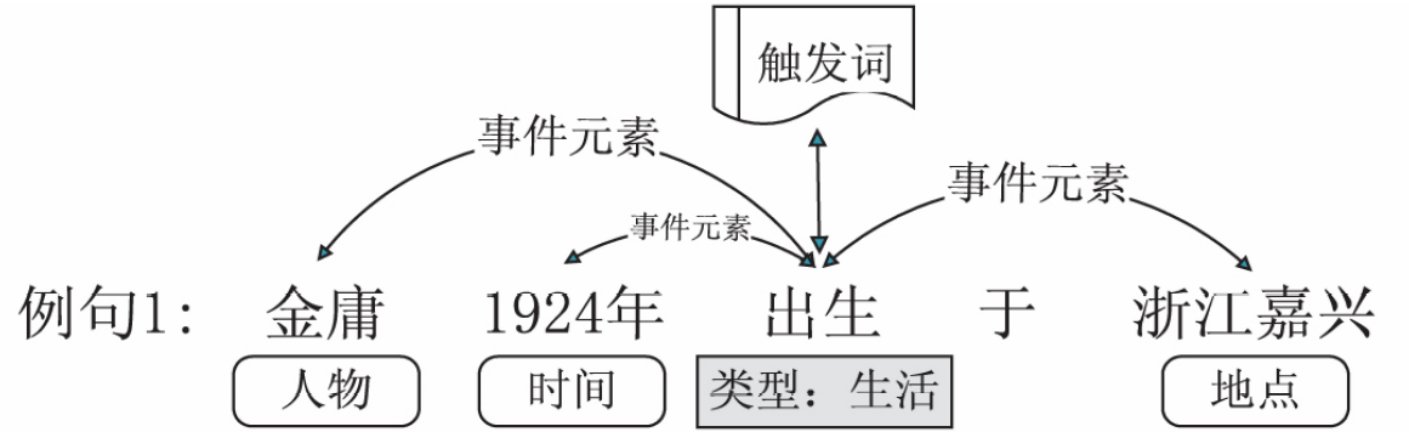
其中 W 表示词语，D 表示文档，T 表示主题。



- **情感分析：**指根据文本所表达的观点和态度等主观信息识别作者对事物 (及其属性) 的情感态度，包括属性识别和情感分类。

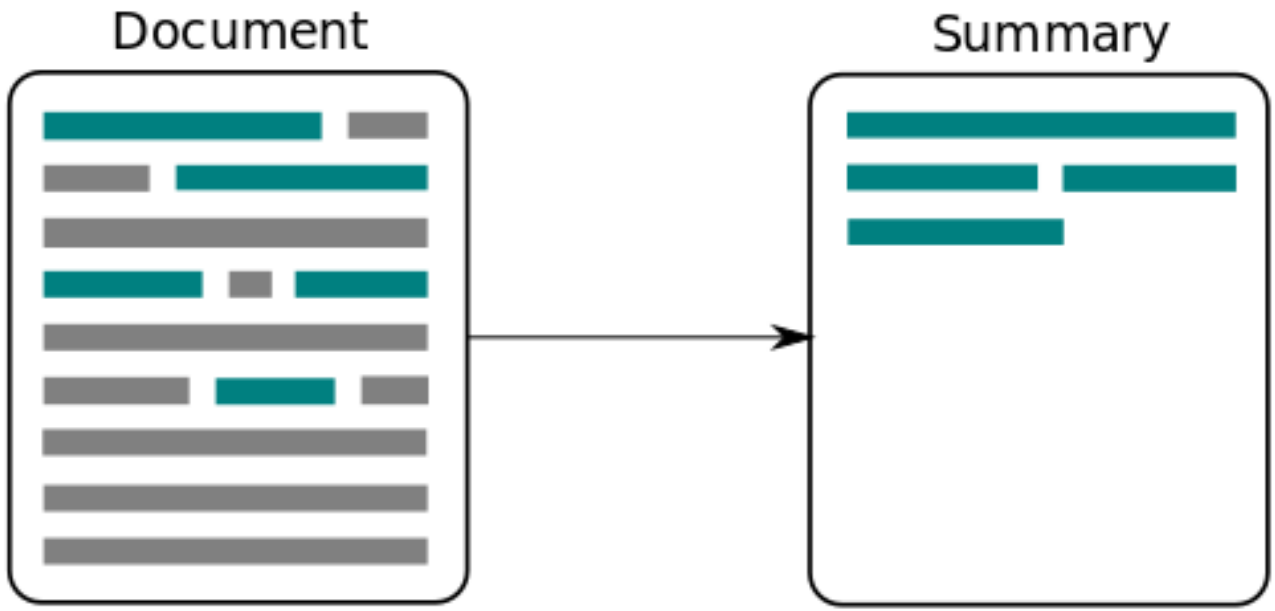


- **信息抽取：**从非结构化、半结构化的自然语言文本 (如网页新闻、学术文献、社交媒体等) 中抽取实体、实体属性、实体间的关系以及事件等事实信息，并形成结构化数据输出的一种文本数据挖掘技术。金融和生物医学文本信息抽取近年来热度上升。



事件类型	生活（出生）	
事件触发词	出生	
事件元素	金庸	角色=人物
	1924 年	角色=时间
	浙江嘉兴	角色=地点

- **文本摘要：**为长文本生成表达其核心意思的短摘要，有效应对信息过载。例如，信息服务部门需要对大量的新闻报道进行自动分类，然后形成某些个事件报道的摘要，推送给可能感兴趣的用户，或者某些公司、政府舆情监控部门想大致了解某些用户群体所发布言论 (短信、微博、微信等) 的主要内容，自动摘要技术就派上了用场。



目录

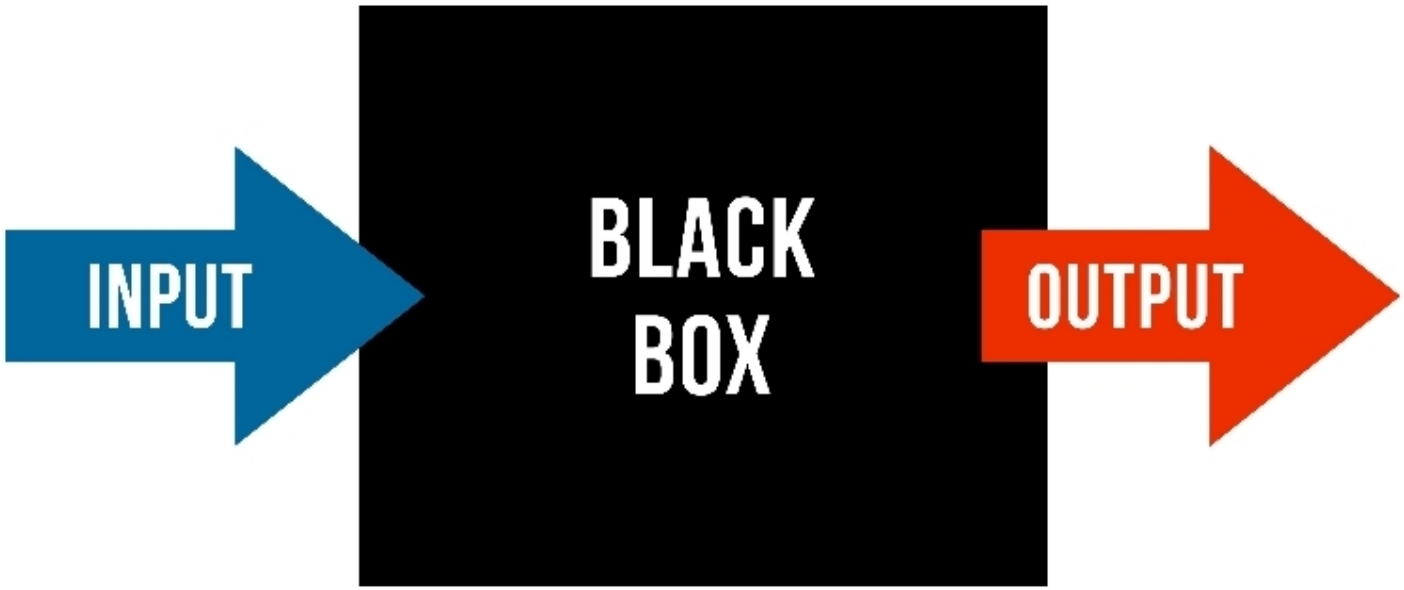
- 文本噪声和非规范性表达使得在规范语料上训练得到的 NLP 模型的准确性降低；
- 歧义表达与文本语义的隐蔽性使得推理能力较差的 NLP 模型的准确性降低；

- 样本收集和标注困难使得主要依赖监督学习算法的 NLP 模型训练成本提高。

- 挖掘目标和结果的要求难以准确表达和理解；

例如，我们可以从某些文本中抽取出频率较高的、可以代表这些文本主题和故事的热点词汇，但如何将其组织成以流畅的自然语言表达的故事梗概 (摘要)，却不是一件容易的事情。

- 语义表示和计算模型不甚奏效；



目录

安装所需软件环境

1. 熟悉Kaggle的用法，学习使用常用的 shell 命令 (文件操作，目录跳转等)。

2. 复习程序设计基础课程中的字符串相关操作。

THE END