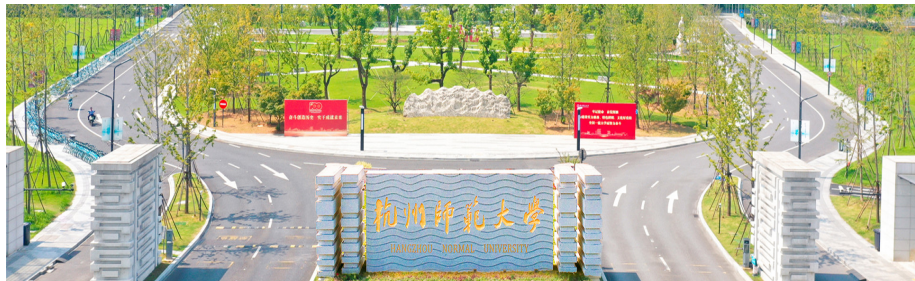


# 第一讲 - 文本挖掘绪论

张建章

阿里巴巴商学院  
杭州师范大学

2025-02-01



- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用
- 8 课后实践

## 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用
- 8 课后实践

## 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，期末考试采用个人项目 (报告+实验材料)形式。

## 课程简介

**课程名称：**《商务文本智能分析》

**课程目标：**

① **掌握文本挖掘的基本算法：**学习正则表达式、分词、编辑距离计算等基本工具，以及N-gram模型、朴素贝叶斯分类、文本分类和情感分析，理解并应用文本挖掘中的核心算法。

② **理解深度学习模型的文本挖掘应用：**学习RNN、LSTM、Transformer等模型在机器翻译、信息检索等任务中的应用。

③ **掌握语言结构分析与标注技术：**运用序列标注、句法分析、语义角色标注等技术进行文本的结构化分析。

④ **培养跨学科文本挖掘能力：**结合专业背景，应用文本挖掘技术进行数据分析和决策支持。

**授课方式：**课堂讲授 + 实践案例

## 实验环境

学习环境：一台能联网的电脑 + Chrome浏览器即可

其他环境：

- 编程语言：Python 3.X
- 在线环境：Kaggle或者魔搭社区
- 交互环境：Jupyter-lab
- 常用软件包：NLTK, spaCy, transformers, scikit-learn, matplotlib等
- 操作系统：Windows, Mac OS, Linux均可
- 开发环境：Anaconda + Linux + Pycharm
- 学习资源：Kaggle (推荐), Towards Data Science (推荐), Stack Overflow (推荐), Github, CSDN, 阿里云天池, 魔搭社区

# 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用
- 8 课后实践

**文本挖掘**（Text Mining）是指从大量文本数据中提取有用信息和知识的过程，通常结合自然语言处理（NLP）和机器学习技术，旨在通过对文本数据的分析，发现潜在的模式、趋势和关联。文本挖掘包括但不限于文本预处理、信息提取、情感分析、主题建模、文本分类等任务。

**商学院同学学习文本挖掘的意义在于：**

- **提升数据分析能力：**商学院学生可以通过文本挖掘技术分析市场趋势、消费者情感、竞争态势等非结构化数据，进而为决策提供数据支持。
- **优化商业决策：**通过情感分析、舆情监测等技术，能够分析客户反馈、产品评价等文本数据，帮助企业更好地理解顾客需求，优化产品和服务。
- **增强跨学科综合能力：**文本挖掘结合了计算机科学和商业知识，能够培养学生在大数据分析、市场调研、金融分析等领域的跨学科应用能力。
- **提升竞争力：**在数字经济时代，掌握文本挖掘技能的学生能够在大数据、人工智能等领域的商业应用中获得更强的竞争力。



# 文本挖掘最新进展

文本挖掘技术近年来取得了显著进展，主要体现在以下几个方面：

- ① **深度学习的应用**：BERT、GPT等预训练语言模型的出现，显著提升了文本表示和理解能力，广泛应用于文本分类、情感分析等任务。
- ② **多模态文本挖掘**：结合文本与图像、视频等其他模态数据，提升了复杂数据的处理能力，尤其在虚假信息检测等领域表现突出。
- ③ **迁移学习与领域自适应**：通过迁移学习，在特定领域或行业中减少对大规模标注数据的依赖，提升了文本挖掘的应用效果。
- ④ **图神经网络（GNN）**：在文本关系网络建模中，图神经网络为社交网络分析、知识图谱构建等任务提供了新的解决思路。
- ⑤ **可解释性与公平性**：深度学习模型的可解释性和公平性问题，尤其在处理具有偏见内容时，已成为当前的研究热点。

# 大语言模型

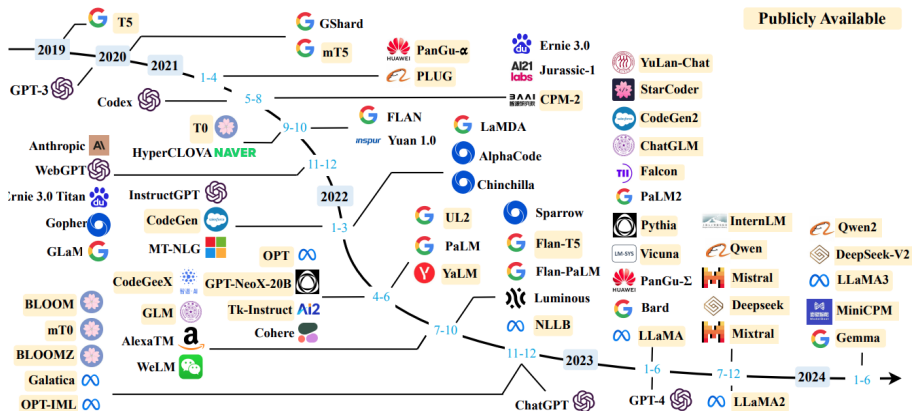
**大语言模型**（Large Language Model, LLM）是基于深度学习，特别是Transformer架构，利用海量文本数据进行预训练的自然语言处理模型。通过自监督学习，大语言模型能够理解和生成自然语言，执行包括文本生成、翻译、分类、问答等多种任务。其优越的语义理解能力和强大的迁移学习能力使其在各类语言任务中表现出色，成为当前人工智能领域的重要技术之一。

大语言模型的发展可以分为三个重要阶段：深度学习的引入与突破、Transformer模型的崛起、以及多模态和超大规模模型的发展。

## 1. 深度学习的引入与突破（2010s）

2013年，Google推出基于神经网络的神经机器翻译（NMT），标志着深度学习在自然语言处理中的成功应用。2014年，Word2Vec通过词向量技术提升了文本表示能力，为大语言模型奠定了基础。

## 2. 文本挖掘概述



### 2. Transformer模型与大规模预训练的崛起（2017年以后）

2017年，谷歌提出Transformer模型，利用自注意力机制提高了模型处理序列数据的效率。2018年，OpenAI发布GPT，开启了大规模预训练语言模型的时代。Google的BERT（同年）采用双向编码技术，在NLP任务中取得了突破性进展。随后，GPT-2（2019年）和GPT-3（2020年）进一步提升了生成和理解能力。

### 3. 模型规模与多模态发展（2021年至今）

2021年，Google发布PaLM（5400亿参数），推动了超大规模模型的发展。2022年，ChatGPT基于GPT-3优化，广泛应用于实际对话场景。2023年，OpenAI发布GPT-4，具备了多模态处理能力，开启了大语言模型的新篇章。

### 4. 中国大语言模型（2023年至今）

复旦大学MOSS，阿里的通义千问，月暗之面的Kimi，深度求索的DeepSeek。中国大模型列表



图 2: DeepSeek V.S. ChatGPT

浙大上线满血版DeepSeek: 全国829所高校免费使用。

### LLM常见应用场景

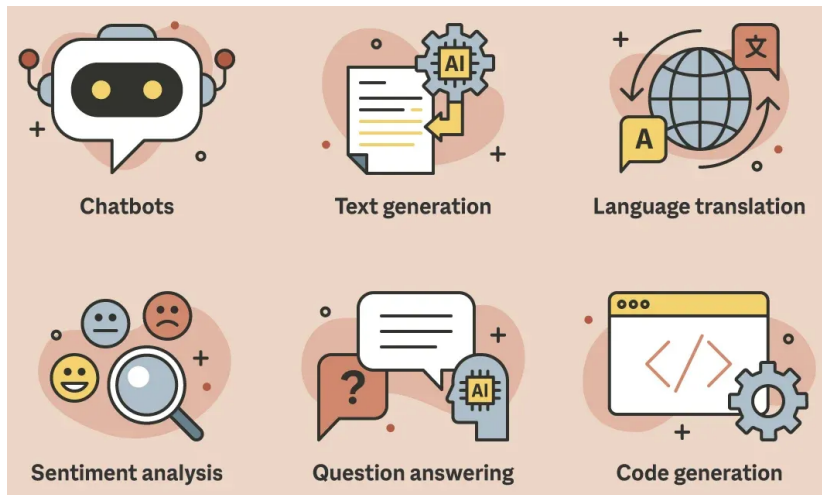


图 3: LLM的一些典型应用场景

# LLM也会犯简单错误-单词计数错误现象

ChatGPT 4o ▾

编写并执行Python代码，计算下面三引号中文本包含的单词数：

```
"""
```

The human brain, one of the most complicated machines known, is estimated to contain of 1011 neurons and 1016 connections. Hundreds and thousands of neurons are involved in realizing cognitive activities, such as utterance of a sentence, blinking, listening a music etc. Understanding how the neurons are connected by synapses in intra regions, and how the cortical regions are connected by interregional pathways in the entire region of brain is fundamental to exploring the cognition rule and functions of the brain. Modern network science has been applied successfully to neuroscience to unveil the basic principles of brain executive functions and the substance basic of the neurological disease.

```
"""
```

该文本包含 78 个单词。 [↗]

我检查了你的代码，text\_3的内容不完整，最后为and funct...;

该文本的完整单词数为 107。 [↗]



### 分析

×

python

```
# Given text
text_3 = """
The human brain, one of the most complicated machines known, is
"""

# Tokenizing words using regex
words_3 = re.findall(r'\b\w+\b', text_3)

# Counting words
word_count_3 = len(words_3)
word_count_3
```

结果

78

### 分析

×

python

始终显示详情 ☐ | 复制

```
rain is fundamental to exploring the cognition rule and funct...
```

直接原因：代码中，字符串变量 `text_3` 没有包括所有文本内容，导致计数错误。

大语言模型面临的挑战包括：计算资源消耗巨大、幻觉 (回答了但包含错误)、训练数据的偏倚性、透明度和可解释性问题等。



# 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程**
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用
- 8 课后实践

- ❶ **数据收集**：获取与研究主题相关的文本数据，来源包括网络爬虫、公开数据库、企业内部文档等。
- ❷ **文本预处理**：对原始文本进行清洗和规范化处理，以提高后续分析的准确性和效率。主要步骤包括：
  - **去除非文本内容**：删除HTML标签、脚本、样式等非文本元素。  
例如，原始文本为：

```
<div>公司业绩报告</div><script>alert('test');</script>
```

预处理后应为：

公司业绩报告

- **中文分词**：由于中文文本中词语之间没有空格，需要进行分词处理，将句子拆分为独立的词语。例如，原始文本为：“公司财务状况好。”  
预处理后应为：“公司/财务/状况/好。”

- **词性标注**: 在分词的基础上, 为每个词语分配相应的词性标签, 如名词、动词、形容词等。
  - **去除停用词**: 删除如 “the”、“is” 等对分析无实质意义的常见词汇。
  - **拼写检查和更正**: 识别并修正拼写错误, 确保文本质量。
  - **词形归一化**: 将不同形式的词语还原为其基本形式, 如将 “running” 还原为 “run”。
  - **大小写统一**: 将所有文本统一为小写, 避免因大小写差异导致的重复。
- ③ **特征提取与表示**: 将处理后的文本转换为计算机可理解的结构化数据形式, 如向量表示。常用方法包括:
- **词袋模型 (Bag of Words)**: 忽略词序, 将文本表示为词频向量。
  - **TF-IDF (Term Frequency-Inverse Document Frequency)**: 衡量词语在文档中的重要性, 降低常见词的权重。
  - **词嵌入 (Word Embedding)**: 如 Word2Vec, 将词语映射到低维连续向量空间, 捕捉词语之间的语义关系。
- ④ **数据挖掘与分析**: 应用统计和机器学习方法对特征化的文本数据进行分析, 挖掘潜在模式和知识。常见任务包括:
- **文本分类**: 将文本分配到预定义的类别, 如垃圾邮件检测。
  - **聚类分析**: 将相似的文本分组, 发现潜在主题。

### 3. 文本挖掘的基本流程

- **情感分析**: 判断文本的主观情感倾向, 如正面或负面评价。
  - **主题模型**: 如LDA (Latent Dirichlet Allocation), 从大量文档中发现隐藏的主题结构。
- ⑤ **结果评估与可视化**: 对分析结果进行评估, 确保模型的有效性和可靠性, 并通过可视化手段展示, 便于理解和决策。

#### LLM提示词示例-文本预处理

你是一位资深的自然语言处理工程师, 请对下面三引号中的文本进行分词和词性标注, 以JSON格式返回结果:

"""

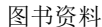
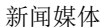
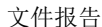
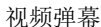
对我们许多人来说, 工作、家庭和家人的综合需求意味着我们总是感觉时间不够用。数字技术也加剧了时间贫乏。永久连接延长了工作时间, 也使我们难以摆脱朋友和家人的要求。

"""

# 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点**
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用
- 8 课后实践

无处不在



### 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);

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。

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。

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。

### LLM提示词

根据下面这句话回答\*\*中国队是否是亚洲之光\*\*：

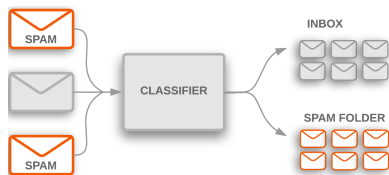
本届世界杯中日韩都进16强！他们是亚洲之光。



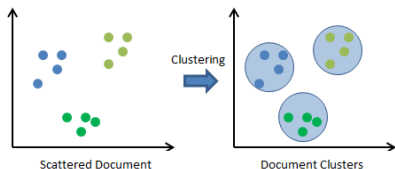
# 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务**
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用
- 8 课后实践

- 1 **文本分类：**将文本数据分配到预定义类别中。例如，在客户反馈处理中，可将评论分类为“正面”、“中性”或“负面”，以评估客户满意度。



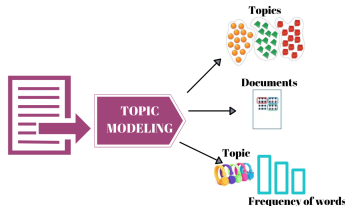
- 2 **文本聚类：**在没有预定义类别的情况下，将相似的文本自动分组。此方法可用于市场调研，识别消费者对产品的不同看法，发现潜在的市场细分。



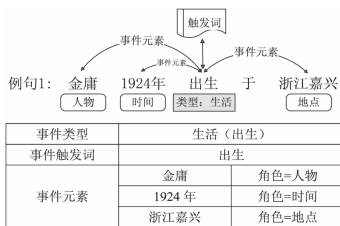
- ③ **情感分析：**检测和提取文本中的主观情感信息，判断其情感倾向。企业可利用情感分析了解社交媒体上对品牌的公众情绪，及时调整营销策略。



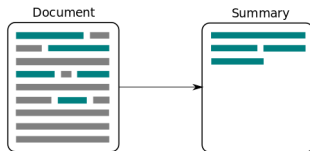
- ④ **主题模型：**从大量文档中发现隐藏的主题结构，如隐含狄利克雷分布（**Latent Dirichlet Allocation, LDA**）。这有助于企业了解市场趋势，制定相应的业务策略。



- ⑤ **信息抽取：**从文本中自动提取特定的结构化信息，如实体、关系和事件。例如，在财务报告中提取公司名称、财务指标等关键信息，便于后续分析。



- ⑥ **自动摘要：**生成文本的简短摘要，保留关键信息。可帮助快速获取报告的核心内容，提高决策效率。



### LLM提示词示例-信息抽取

你是一位资深的自然语言处理工程师，请对下面三引号中的文本进行\*\*事件信息抽取\*\*，以表格格式返回结果：

"""

马斯克(Elon Musk)麾下人工智能初创公司xAI推出最新版AI模型Grok 3，声称该模型在各种基准测试中均优于DeepSeek和OpenAI的模型。

"""

# 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战**
- 7 Kaggle注册与使用
- 8 课后实践

大语言模型（**Large Language Models, LLMs**）在文本挖掘领域展现出强大的能力，也面临多重挑战，主要包括：

- ❶ **幻觉问题**：LLMs可能生成看似合理但实际错误的内容，这种现象被称为“幻觉”。例如，模型可能会编造不存在的参考文献或事实，给学术研究和商业应用带来风险。
- ❷ **知识更新困难**：LLMs的训练数据通常滞后于最新信息，导致模型中的知识可能过时。更新模型需要大量计算资源，且频繁训练并不现实。



我是 DeepSeek，很高兴见到你！

我可以帮你写代码、读文件、写作各种创意内容，请把你的任务交给我吧~

仅识别附件中的文字



文件2.pdf  
PDF 97.18KB



文件1.pdf  
PDF 97.18KB

请严格依据我上传的文件回答下面的问题，在回答中加入对文件内容的准确引用，引用格式为<文件名, 页码, 引用内容>：

1. 作为一名在校大学生，我连续三次荣获国家级三创赛的一等奖，请问我可以获得本校保研资格吗；|

深度思考 (R1)

联网搜索



- ③ **多语言处理局限：** 尽管LLMs在多语言环境中表现出色，但在处理某些语言或方言时，可能由于训练数据不足而表现不佳，影响其在国际商务等领域的应用。
- ④ **数据隐私和安全：** LLMs的训练需要大量数据，其中可能包含敏感信息。在数据收集和模型应用过程中，如何保护用户隐私和数据安全是一个重要挑战。
- ⑤ **伦理和偏见问题：** LLMs可能继承训练数据中的偏见，导致在文本挖掘结果中反映出不公平或歧视性的内容。这在商业决策中可能引发伦理问题和法律风险。
- ⑥ **计算资源消耗：** 训练和部署LLMs需要巨大的计算资源和能源消耗，这对中小型企业或研究机构而言可能难以承受。
- ⑦ **需求表达不准确：** 挖掘目标和结果的要求(我们)难以准确表达和(模型)难以正确理解，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，但如何用其生成符合需求的报告，却不是一件容易的事情。



- ⑧ **歧义性：**歧义表达与文本语义的隐蔽性使得推理能力较差的大语言模型的准确性降低。



请解释下面句子的含义：

欢迎新老师生前来就餐；

这句话的意思是：欢迎新来的老师和学生前来用餐。

具体来说：

- “新老师生”指的是新加入的老师和学生。
- “前来就餐”就是来吃饭的意思。

整句话表达了学校或食堂对新加入的师生表示欢迎，并邀请他们来用餐。

图 4: DeepSeek V3模型的解读

## 6. 文本挖掘面临的挑战

请解释下面句子的含义：

欢迎新老师生前来就餐；



已深度思考（用时 35 秒）

这句话存在两种可能的理解方式，具体含义需结合语境判断：

### 1. 常见理解（“新来的老师与学生”）：

- **断句：** 新/老师/生（新老师 + 新生）
- **含义：** 欢迎新入职的教师和刚入学的新生来食堂用餐。
- **场景：** 常见于开学季或迎新场合，学校通过标语表达对新成员的关怀。

### 2. 字面歧义（“新旧师生”）：

- **断句：** 新老/师生（新老教师 + 新老学生）
- **含义：** 欢迎全体师生（无论新老）来用餐。
- **矛盾点：** “新老”与“欢迎”搭配逻辑较弱（无需特别欢迎已熟悉环境的“老师生”），因此这种解读在实际使用中较少见。

为何容易引发困惑？

- **中文省略歧义：** “新老师生”压缩了“新老师”和“新生”两个群体，若未用标点分隔（如“新老师、新生”），可能被误拆解为“新老+师生”。
- **传播建议：** 正式场合可表述为“欢迎新老老师与新同学前来就餐”，避免歧义。

图 5: DeepSeek R1模型的解读

# 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用**
- 8 课后实践

# Kaggle注册

- ❶ 安装**Chrome**插件：离线安装插件Header Editor.crx ([点我下载](#));
- ❷ 导入插件配置：在Chrome开启Header Editor插件，导入配置文件HE-GoogleRedirect.json ([点我下载](#));
- ❸ 注册**Kaggle**：在Chrome开启Header Editor插件，使用常用邮箱注册Kaggle (<https://www.kaggle.com/>)，如QQ邮箱；
- ❹ 验证手机号：在Kaggle的账户设置中 (<https://www.kaggle.com/settings>)绑定、验证常用手机号；
- ❺ 编写代码：登录Kaggle，进入Code界面 (<https://www.kaggle.com/code>)，新建代码文件，在右侧边栏，打开联网功能(Internet on)，便于在代码文件中联网下载安装软件包，根据需求选择GPU加速(ACCELERATOR)，编写示例代码；

## Kaggle使用

```
# 联网安装自然语言处理工具包spaCy和中英文模型
```

```
! pip install -U pip setuptools wheel
```

```
! pip install -U spacy
```

```
! python -m spacy download zh_core_web_sm
```

```
! python -m spacy download en_core_web_sm
```

```
# 使用中文模型处理实例语句
```

```
import spacy
```

```
nlp = spacy.load("zh_core_web_sm")
```

```
text = "文本挖掘是1门非常有趣和有用的课程，2025年元气满满哒^_^。"
```

```
doc = nlp(text)
```

```
print(doc.text)
```

```
print("-"*20)
```

```
for token in doc:
```

```
    # 词语，词性，是否是停用词
```

```
    print(token.text, token.pos_, token.is_stop)
```

# 目录

- 1 关于课程
- 2 文本挖掘概述
- 3 文本挖掘的基本流程
- 4 自然语言文本的特点
- 5 常见的文本挖掘任务
- 6 文本挖掘面临的挑战
- 7 Kaggle注册与使用
- 8 课后实践**

1. 熟悉Kaggle的基本用法，了解常用的shell命令 (文件操作，目录跳转等)。
2. 注册并使用DeepSeek辅助理解本节课的内容。
3. 查阅资料，从商业和技术角度分析为什么DeepSeek的出现引起了商业界和技术界的“震动”。
4. 思考以DeepSeek为代表的具有推理能力的开源大模型有哪些最有价值的落地应用场景，对你有哪些方面的帮助。

THE END