# Data Mining and Optimization

## Lecture 2: Basic Concepts in Text Data Mining

Liu Yang

Nanjing University

Spring, 2025

# Table of Contents

# 什么是文本数据挖掘?

文本数据挖掘（text data mining）是指从自然语言文本中挖掘研究者感兴趣的模式与知识的方法和技术。

- 文本：txt、doc/docx、pdf 和 HTML 等以语言文字为主要内容的数据文件；
- 与通常的 excel 数据不同，文本是非结构化数据（unstructured data）；
- 文本数据挖掘的最大挑战在于对非结构化自然语言文本的分析和理解。

# 文本数据挖掘的应用场景

- 社交网络分析：通过分析微博、微信、短信等社交网络信息，及时准确了解民意，把稳舆情；
- 经济预测：通过深入挖掘大量新闻报道、财务报告和网络评论等文字材料，预测经济走势和股市行情；
- 质量管理：分析用户对产品的评价及市场反应，改进产品质量，为用户提供个性化服务；
- 风险识别：通过分析大量的化验报告和病例记录，发现某种传染性疾病的发展规律。

# Table of Contents

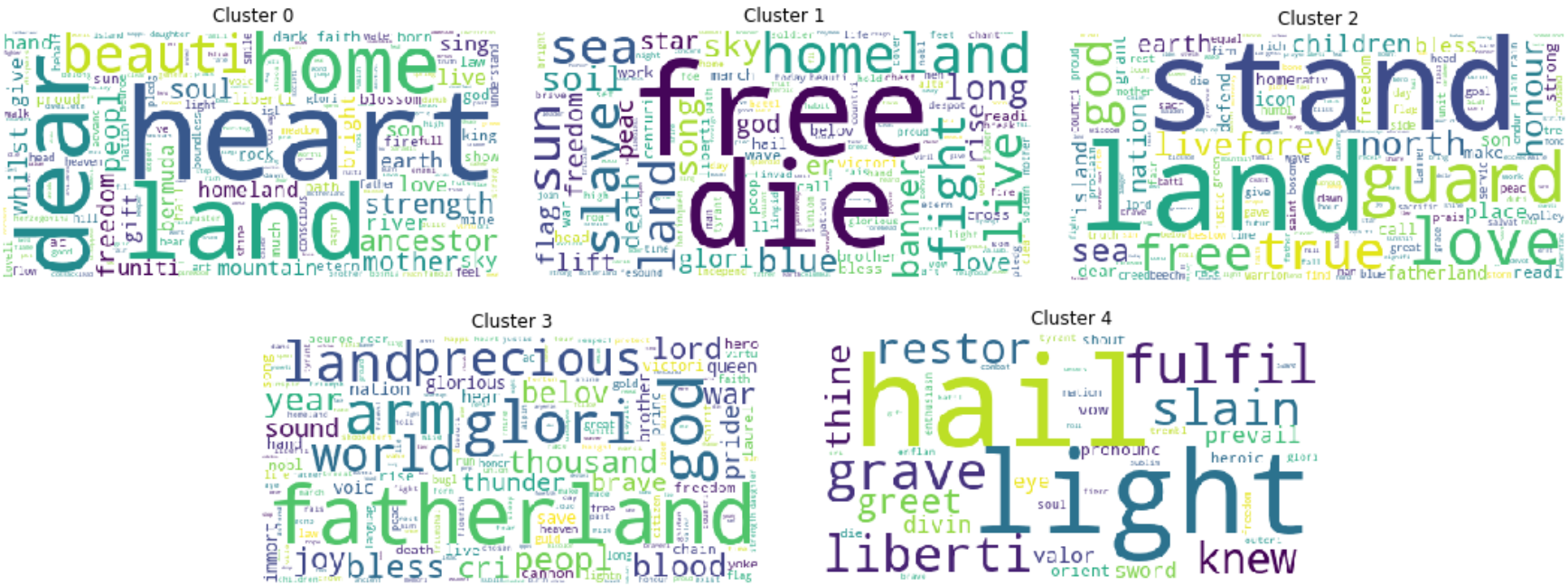
# 文本挖掘任务——文本分类

文本分类（text classification）是根据一部图书或者一篇文章的内容将其划分到事先指定的文本类型。

# 文本挖掘任务——文本聚类

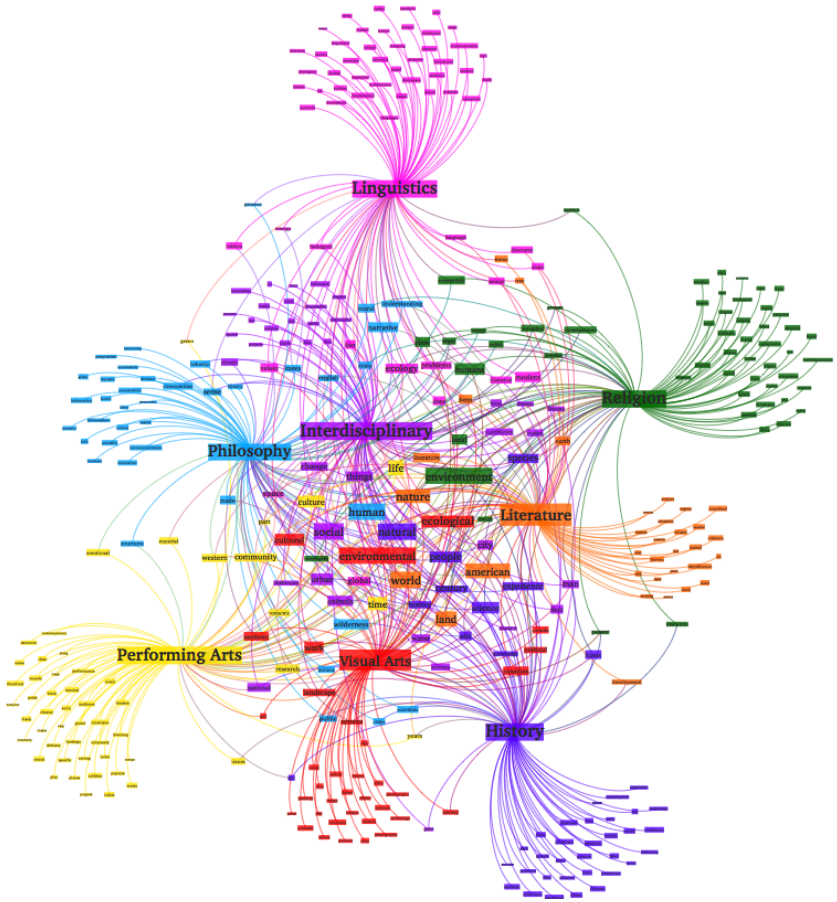
文本分类（text clustering）是根据文本内容将一组文本划分成不同的类别。

- 分类：事先知道有多少个类别，分类的过程就是将每一个给定文本自动划归为某个确定的类别；
- 聚类：事先不知道有多少个类别，需要根据某种标准将给定的文档集合划分为不同类别。



# 文本挖掘任务——主题模型

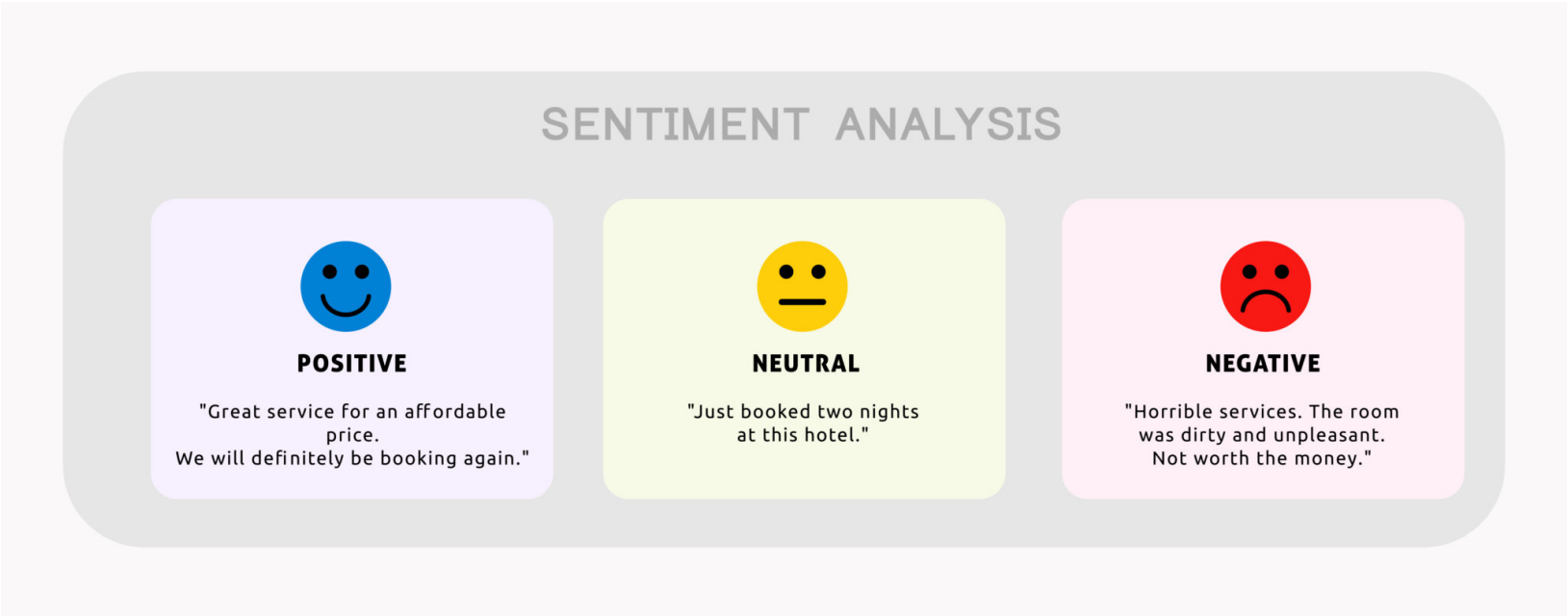
主题模型（topic model）是为了从文本中挖掘隐藏在词汇背后的主题和概念。



# 文本挖掘任务——情感分析与观点挖掘

文本情感分析（text sentiment analysis）是为了从文本中挖掘作者的观点和态度。

- 情感分类：根据本文本所表达的态度判断其褒贬极性；
- 例：分析某特殊事件发生后互联网上大量新闻报道和用户评论的主观倾向性；
- 例：从众多用户对一款新产品的网络评论中了解其主观评价。



# Table of Contents

# 文本噪声和非规范性表达

文本分析的主要数据来源是互联网，与规范书面语相比，网络文本存在大量非规范表述。

- “完了 芭比 q 了家人们”
- “人生无常，大肠包小肠”
- “YYDS”

# 歧义表达与语义隐藏性

- bank：银行？河岸？
- apple：苹果？苹果？
- (关于鲁迅) 的文章？关于 (鲁迅的文章)？

# Table of Contents

# 正则表达式——匹配字符串

在 Python 中，re 模块可以实现正则表达式 (regular expression)。

- re.search(regex,string): 检查字符串 string 是否匹配正则表达式 regex；
- 如果匹配到，返回一个 match 对象；
- 如果没有匹配到，返回 None。

# 正则表达式——匹配字符串

. 表示匹配任意单个字符。

- "a.c" 可以匹配到"abc","branch"，不能匹配到"add","crash"；
- "..t" 可以匹配到"bat","oat"，不能匹配到"it","table"。

# 正则表达式——匹配字符串

^ 表示匹配开始的字符串；$ 表示匹配结尾的字符串。

- "^a" 表示匹配所有以字母 a 开头的字符串；
- "a$" 表示匹配所有以字母 a 结尾的字符串。

# 正则表达式——匹配字符串

[] 表示匹配多个字符。

- "[bcr]at" 代表的匹配是"bat","cat" 以及"rat"。

# 正则表达式——抽取文本中的数字

- "[0-9]" 表示从 0 到 9 的所有数字，"[a-z]" 表示从 a 到 z 的所有小写字母。
- re.findall(regex,string) 返回能够与 regex 匹配的 string 中的那部分字符串。

# 正则表达式

re: Regular expression operations