数据挖掘与商务分析

# 自然语言处理与商务实践

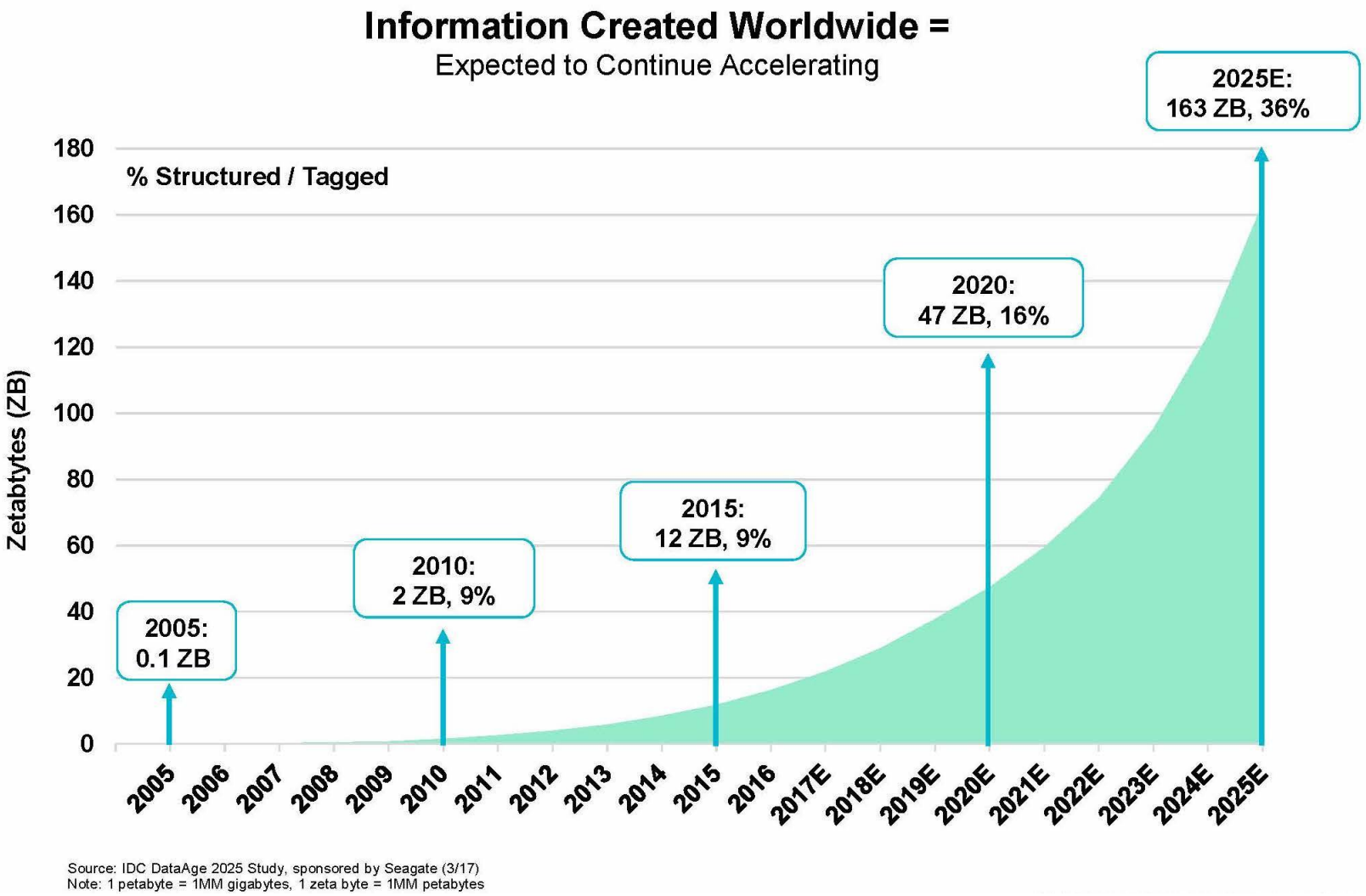
主讲教师：肖升生
xiao.shengsheng@shufe.edu.cn

# 知识回顾

- **数据类型：**
  - 数值
  - 文本
  - 位置
  - 声音
  - 视频
  - ……


资料来源: https://www.domo.com/

# 知识回顾




Source: IDC DataAge 2025 Study, sponsored by Seagate (3/17)
Note: 1 petabyte = 1MM gigabytes, 1 zeta byte = 1MM petabytes

# 项目案例分享

■ **问题：**

- 地址信息中存在着大量的非标准的、甚至错误的地址、地名信息；

■ **需求：**

- 将非标准地址数据校正为标准地址数据，可实现自动化的分发（分词+分级）
- 对新出现的地址，能够有效识别
- 在此基础上，根据标准地址实现自动分发

# 中文地址层级

中文地址的层级

| 地址要素的层级 | |
|---|---|
| 1 | 省 |
| 2 | 市 |
| 3 | 区、县 |
| 4 | 乡、镇、街道 |
| 5 | 社区、村 |
| 6 | 小区、组 |
| 7 | 开发区 |
| 8 | 商圈 |
| 9 | 主路 |
| 10 | 支路 |
| 11 | 主门牌 |
| 12 | 支门牌 |
| 13 | 地标 |
| 14 | 楼栋 |
| 15 | 单元 |
| 16 | 层 |
| 17 | 房 |
| 18 | 描述性信息 |

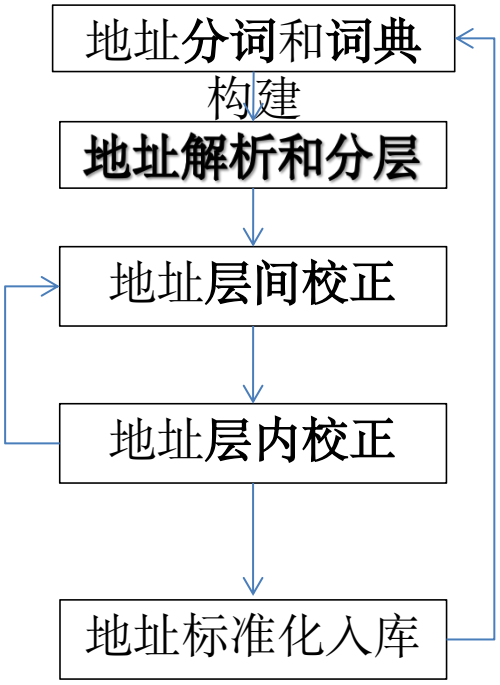
# 常见错误类型

1. **笔误（如同音字）**：“山东省青岛市绍新路59号-901”，“绍兴路”写成了“邵新路”。层内错误。
2. **元素重复**：“安徽省巢区世纪新疆合肥市居都c区号楼3单元401”，该地址中包含了“新疆”和“安徽省”两个省级地址。层内错误。
3. **地址逻辑错误**：“上海市杨浦区南京东路180号”，杨浦区不存在南京东路。层间错误。
4. 地址**层级次序**错误：“安徽省合肥市庐阳区221号长江中路巡警六大队”，该地址将路号“221号”放在了路名之前。层间错误。
5. **地址元素缺失**：“安徽省淮南市华宫公司”，没有具体的区县、道路、路号地址。层间错误。

# 纠错的主要挑战

1. 汉语输入是人工编码的间接输入，录入时的输入法不同，其错误的类型也大不一样：拼音、五笔、手写、联想等等。
2. 词之间没有分隔标志，在进行任何词和词以上层级的处理时，都必须先分词。分词结果直接影响到校对的查错率、召回率，但其本身尚有歧义切分、新词等难题。
3. 汉语的词类没有形态标志，而且和句法成分之间并没有简单的映射关系，再加上兼类、句法成分省略等的干扰，造成句法分析困难。

# 我们的技术路线




对原始地址分词，识别地址元素：基于标准地名词典和命名规则的分词方式，准确率高

按八级地址模型进行初步分层，考虑空间约束，基于有限状态机的地名树解析。

基于已经建成的标准基础地址库，使用模式规则和机器学习方法，对每一层和层间关系进行地址查重、查误处理，解决2、3、4类问题

- 针对地址层内错误，如文字错误、地址错误，或利用标准基础地址库也无法分层比对的新地址，集成应用NLP+机器学习方法来纠错（校正），包括n-gram、贝叶斯网络、隐马尔可夫和CRF等方法。
- 在层内校正基础上，再做层间校验，直到不再检测出层间和层内错误为止。

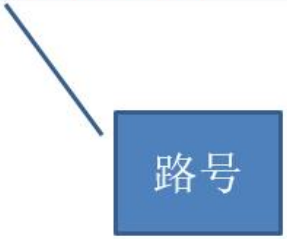
解决1-4类问题，大概率解决5类问题。

- 对于已完成以上校正处理的地址，标准化后存入标准基础地址库，并成为下一次迭代时的词典，通过不断的迭代和数据积累，
- 模型将具有自学习的能力，查全率和查准率可不断逼近100%。

# 结果和应用

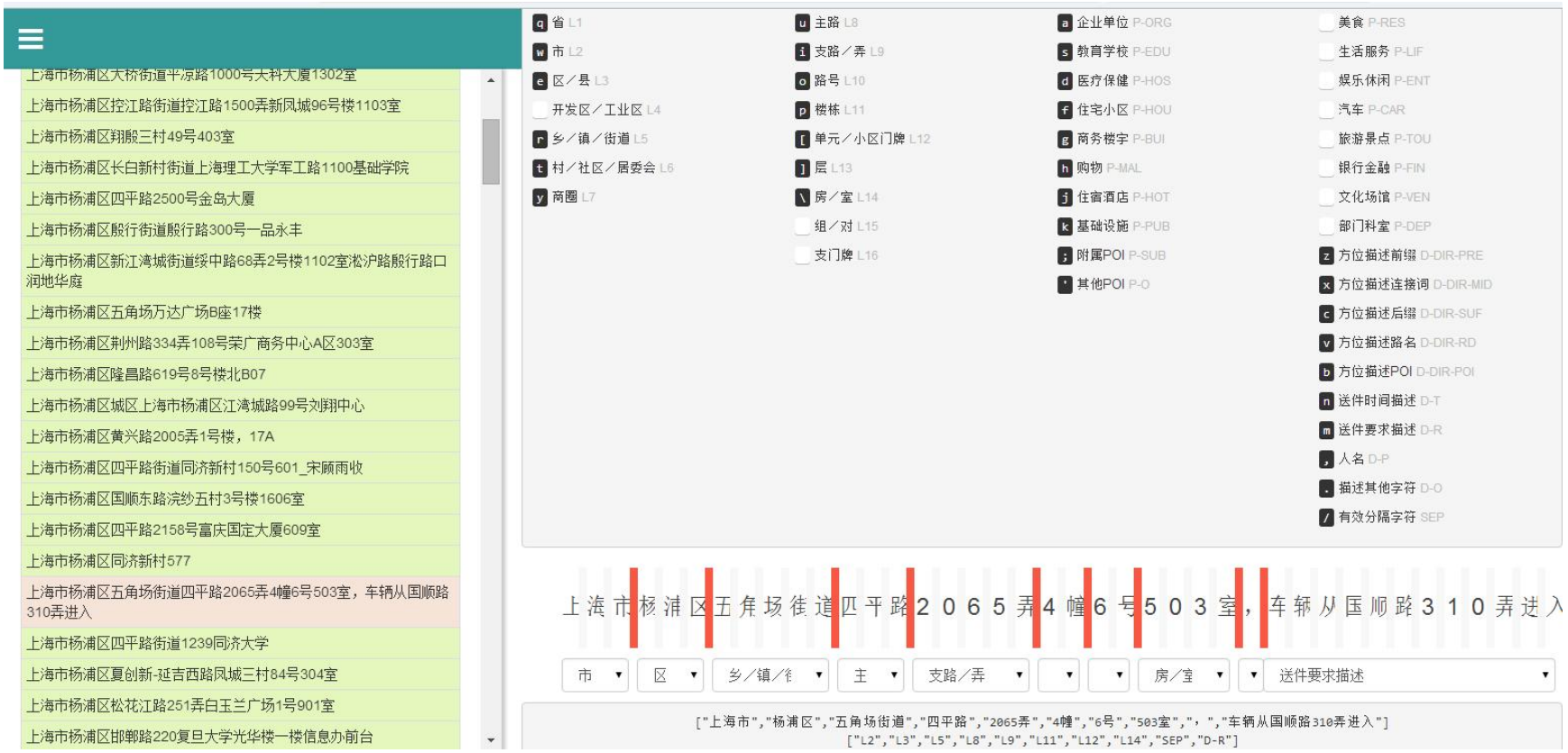
地址的解析结果样例

| | |
|---|---|
| **原始地址1** | 浙江省杭州市西湖区周日请勿配送谢谢转塘镇叶埠桥大美创意园 |
| **分级结果1** | 浙江省(1)杭州市(2)西湖区(3)周日请勿配送谢谢(18)转塘镇(5)叶埠桥（13）大美创意园(13) |
| **原始地址2** | 上海市金山区亭林镇长三角林吉路83311栋 |
| **分级结果2** | 上海市(1)金山区(3)亭林镇(5)长三角(13)林吉路(9)833(11)11栋(14) |




- 中文地址的标准化解析准确率达到**98%+**，覆盖率**91%**；
- 错误地址的查全率**90%+**，查准率**85%+**；
- 在复杂文本环境下（如自由文本），地址的识别准确率达到**90%+**。

# 结果和应用


上海市杨浦区大桥街道平凉路1000号大科大厦1302室
上海市杨浦区控江路街道控江路1500弄新凤城96号楼1103室
上海市杨浦区翔殷三村49号403室
上海市杨浦区长白新村街道上海理工大学军工路1100基础学院
上海市杨浦区四平路2500号金岛大厦
上海市杨浦区殷行街道殷行路300号一品永丰
上海市杨浦区新江湾城街道绥中路68弄2号楼1102室淞沪路殷行路口润地华庭
上海市杨浦区五角场万达广场B座17楼
上海市杨浦区荆州路334弄108号荣广商务中心A区303室
上海市杨浦区隆昌路619号8号楼北B07
上海市杨浦区城区上海市杨浦区江湾城路99号刘翔中心
上海市杨浦区黄兴路2005弄1号楼，17A
上海市杨浦区四平路街道同济新村150号601_宋顾雨收
上海市杨浦区国顺东路浣纱五村3号楼1606室
上海市杨浦区四平路2158号富庆国定大厦609室
上海市杨浦区同济新村577
上海市杨浦区五角场街道四平路2065弄4幢6号503室，车辆从国顺路310弄进入
上海市杨浦区四平路街道1239同济大学
上海市杨浦区夏创新-延吉西路凤城三村84号304室
上海市杨浦区松花江路251弄白玉兰广场1号901室
上海市杨浦区邯郸路220复旦大学光华楼一楼信息办前台
q 省 L1
w 市 L2
e 区／县 L3
开发区／工业区 L4
r 乡／镇／街道 L5
t 村／社区／居委会 L6
y 商圈 L7
u 主路 L8
i 支路／弄 L9
o 路号 L10
p 楼栋 L11
[ 单元／小区门牌 L12
] 层 L13
\ 房／室 L14
组／对 L15
支门牌 L16
a 企业单位 P-ORG
s 教育学校 P-EDU
d 医疗保健 P-HOS
f 住宅小区 P-HOU
g 商务楼宇 P-BUI
h 购物 P-MAL
j 住宿酒店 P-HOT
k 基础设施 P-PUB
; 附属POI P-SUB
' 其他POI P-O
美食 P-RES
生活服务 P-LIF
娱乐休闲 P-ENT
汽车 P-CAR
旅游景点 P-TOU
银行金融 P-FIN
文化场馆 P-VEN
部门科室 P-DEP
z 方位描述前缀 D-DIR-PRE
x 方位描述连接词 D-DIR-MID
c 方位描述后缀 D-DIR-SUF
v 方位描述路名 D-DIR-RD
b 方位描述POI D-DIR-POI
n 送件时间描述 D-T
m 送件要求描述 D-R
, 人名 D-P
. 描述其他字符 D-O
/ 有效分隔字符 SEP
上 海 市 杨 浦 区 五 角 场 街 道 四 平 路 2 0 6 5 弄 4 幢 6 号 5 0 3 室 ， 车 辆 从 国 顺 路 3 1 0 弄 进 入
市
区
乡／镇／街
主
支路／弄
房／室
送件要求描述
["上海市","杨浦区","五角场街道","四平路","2065弄","4幢","6号","503室","，","车辆从国顺路310弄进入"]
["L2","L3","L5","L8","L9","L11","L12","L14","SEP","D-R"]

# 讲授提纲

01 自然语言处理基本概念

02 文本数据预处理

03 文本表示向量空间模型

04 文本表示-主题模型&词向量模型

05 大语言模型及其进展

06 商务案例-在线用户评论分析

# 讲授提纲

01 自然语言处理基本概念

02 文本数据预处理

03 文本表示向量空间模型

04 文本表示-主题模型&词向量模型

05 大语言模型及其进展

06 商务案例-在线用户评论分析

# 基本概念

- 文本数据挖掘
  - Text Data Mining / Text Mining
- 从自然语言文本中挖掘用户所感兴趣的模式和知识的方法和技术


# 任务流程

文本预处理

文本表示

数据挖掘模型构建

数据挖掘模型评估

数据挖掘模型部署

- 输入：
  - Text mining is to identify useful information.
- 输出：

| text | mine | identify | useful | information | … |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 0 |

- 文本分类
  - 情感分析
- 文本聚类
- ……

# 实际应用 – 情感分析

Negative Positive

8-16 18:06 来自 iPhone

昨天去试驾了一下特斯拉，第一感受是内饰有些失望，虽然很简约，但是国产modle 3为了压低价格，内饰的硬塑材质让人有些大跌眼镜。其次标准续航里程市内通勤一周一充，平常也是足够了，不过续航电量打折力度还是比想象大。前后储物空间很大，尤其是前引擎盖内空间也可以储物，这点没想到。另外，科技智能感还是十足。总之，是一个大玩具。那么问题来了，哪个颜色好看？

**Tesla: The Key Reason The Shorts Are Getting Run Over**

DoctoRx • Yesterday, 8:00 AM • 412 Comments

**A Tesla Shareholder's Biggest Fear: Robotaxis**

Sean Chandler • Yesterday, 12:35 AM • 209 Comments

**Tesla: New Estimates Not Good**

Bill Maurer • Wed, Dec. 18 • 529 Comments

**Dissecting The Life-Cycle Profitability Of Tesla's Leased Cars**

Bill Cunningham • Tue, Dec. 17 • 291 Comments

**The Tesla Cybertruck Is No Ford F-150**

John Engle • Mon, Dec. 16 • 492 Comments

# 实际应用 – 市场预测

**Elon Musk** @elonmusk · 13h

Tesla stock price is too high imo

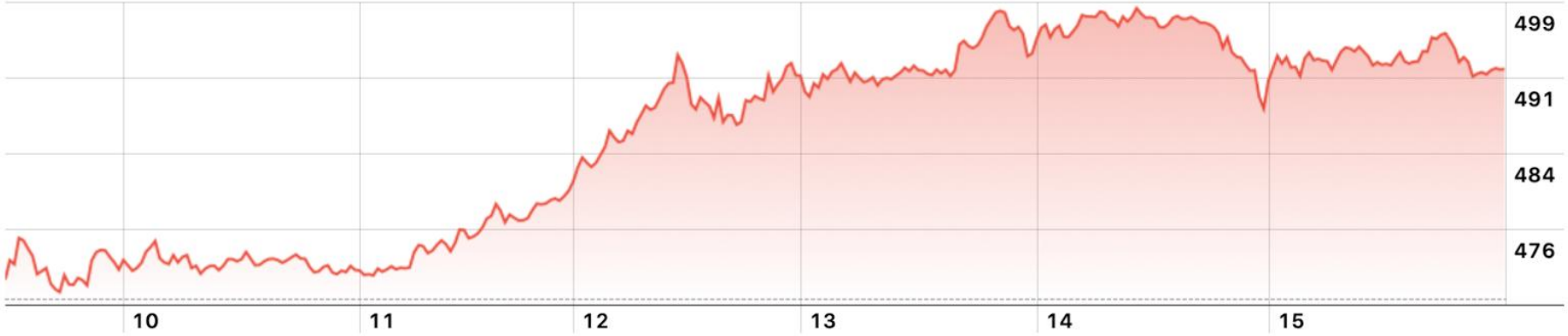
11.3K 18.1K 143.9K

**TSLA** Tesla, Inc.

492.14 +23.08 收盤

496.05 +3.91 盤後

1天 | 1週 | 1個月 | 3個月 | 6個月 | 1年 | 2年 | 5年 | 10年 | 全部




| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 開盤價 | 473.70 | 成交量 | 3056萬 | 52週最高價 | 498.49 | 殖利率 | – | 更多Yahoo資料 |
| 最高價 | 498.49 | 本益比 | – | 52週最低價 | 176.99 | 風險係數 | 0.65 | |
| 最低價 | 468.23 | 市值 | 845.3億 | 平均成交量 | 960.7萬 | EPS | -4.77 | |

# 实际应用 – 推荐系统


2020年版中国经济增长词典

· 民法典，为何而来
· 这些“钱”怎么来怎么花
· 快点！跟上这节奏
· 99秒“视”读政府工作报告

你喜欢

别出新财

线上教育的效果好吗？

回望过去这几个月的“网课历程”，无论是家长、孩子、老师还是教育机构，都可能是一场一言难尽的“奇幻漂流”历程。

**Fish's Wild Island Grill** Claimed

324 reviews Details

$$ · Seafood, Poke, Ramen Edit

Write a Review | Add Photo | Share | Saved

**COVID-19 Advisory: Business operations may be**

Due to ongoing precautionary measures, please contact the hours and availability.

**Popular Dishes**

**Chicken Katsu**
38 Photos · 64 Reviews

**Dot Island**
1 Photo · 37 Reviews

**Jiawei C.**
Davis, CA
0 friends
1 review

9/30/2016
Lemon pepper steamed shrimp is awesome.
Useful Funny Cool

**Megan D.**
San Francisco, CA
99 friends
4 reviews
2 photos

10/21/2019
One of the healthiest, friendliest, fastest and tastiest places in Davis! When I first moved here I avoided it thinking a cheap seafood place couldn't be good but I was so wrong! I get their grilled fish bowls or poke bowls weekly. Their poke is a hidden gem, it's one of the best in Davis in terms of fish quality, freshness, and ability to customize every aspect of it. The small bowl is a perfect size without being overwhelming and super affordable.

# 面临的困难

- **文本噪声或非规范性表达**
  - 烫烫烫
  - 太厉厉厉害了!
- **歧义表达与文本语义的隐蔽性**
  - 关于“鲁迅的文章”；“关于鲁迅”的文章
  - 请问张三的爸爸的儿子今年几岁？
- **样本收集和标注困难**
- ……

# 讲授提纲

01 自然语言处理基本概念

02 文本数据预处理

03 文本表示向量空间模型

04 文本表示-主题模型&词向量模型

05 大语言模型及其进展

06 商务案例-在线用户评论分析

# 文本预处理

## Shanghai University of Finance and Economics

From Wikipedia, the free encyclopedia
(Redirected from SUFE)

*Not to be confused with Shanghai University or Shanghai Finance University.*
*"SUFE" redirects here. For other uses, see SUFE (disambiguation).*

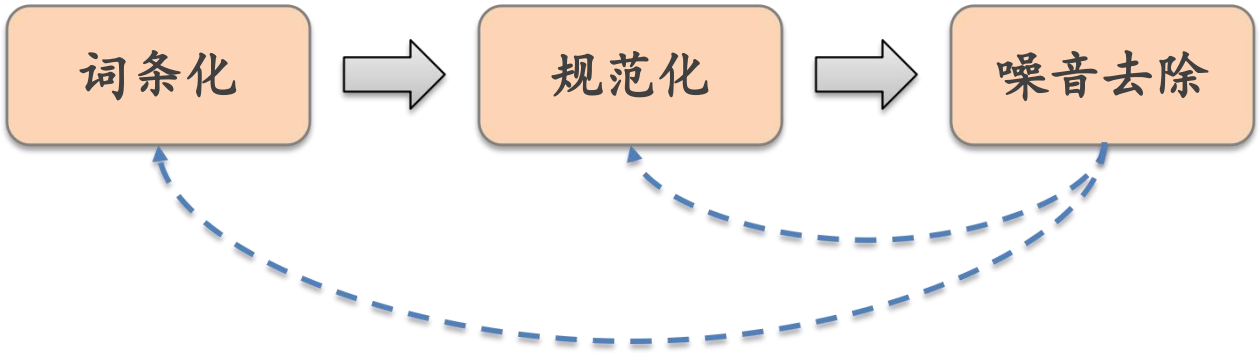
The **Shanghai University of Finance and Economics** (**SUFE**; Chinese: 上海财经大学; pinyin: *Shànghǎi Cáijīng Dàxué*), founded in 1917, is a top-ranked, world-renowned finance- and economics-oriented research university located in Shanghai, the People's Republic of China. The university is under the

## 上海财经大学

Shanghai Univ. Of Fin. & Econ

上海财经大学（Shanghai University of Finance and Economics）是中华人民共和国教育部直属的一所以经济管理学科为主，经、管、法、文、理、哲等多学科协调发展的研究型重点大学，国家首批世界一流学科建设高校，国家"211工程"、"985工程优势学科创新平台"重点建设高校，入选国家"111计划"、卓越法律人才教育培养计划、国家级大学生创新创业训练计划、国家经济学基础人才培养基地、国家海外高层人才创新创业基地、全国高校实践育人创新创业基地、教育部人文社会科学重点研究基地、国家建设高水平大学公派研究生项目、中国政府奖学金来华留学生接收院校、全国深化创新创业教育改革特色典型经验高

# 一般流程




- 词条化（Tokenization）
- 规范化（Normalization）
- 噪音去除（Noise Removal）

# 词条化

- **定义：**
  - 将给定的文本切分成为词汇单位的过程
  - 分词（Segmentation）
- **相关因素：**
  - 语言、情境、切分粒度等
- **以英文为例：**
  - 输入: It's not straight-forward to perform so-called "tokenization".
  - 输出#1: 'It's', 'not', 'straight-forward', 'to', 'perform', 'so-called', '"tokenization".'
  - 输出#2: 'It', '"', 's', 'not', 'straight', '-', 'forward, 'to', 'perform', 'so', '-', 'called', '"', 'tokenization', '"', '.'

# 词条化

- **定义：**
  - 将给定的文本切分成为词汇单位的过程
- **相关因素：**
  - 语言、情境、切分粒度等
- **以中文为例：**
  - 输入：我来到上海财经大学上课
  - 输出#1：我 / 来到 / 上海财经大学 / 上课
  - 输出#2：我 / 来到 / 上海 / 财经 / 大学 / 上课

# 规范化

- **定义：**
  - 对词汇单位的不同形态进行归并
- **以英文为例：**
  - 大小写
  - 字母数字混合
    - p2p
  - 字母标点混合
    - wi-fi => wifi
  - 词形还原 (lemmatization)
  - 词干提取 (stemming)

# 词形还原

- **定义：**
  - 把任意变形的词汇还原成为原形
  - 单复数：
    - ladies => lady, ……
  - 进行式：
    - referring => refer, ……
  - 过去式：
    - forgotten => forget, ……
  - ……
- **实际中，可以考虑词性以提高词形还原准确率**
  - ground => grind, ……

# 词干提取

- 定义：
  - 去除词缀得到词根的过程

- 不一定能够表达完整的语义
  - fisher => fish
  - effective => effect
  - ……

# 噪音去除

- 数据获取时混入的网页标签或文件头
  - <body>麻烦</body>
  - Sunday, June 28, 2020
- 拼写错误
  - fhsadlfh
- 非正式用语
  - 山寨, 雨女无瓜, ……
  - b4, 2morrow, Soooo, ……
- ……

# 停用词去除

- **停用词（Stopwords）：**
  - 主要指功能词，通常指在各类文档中频繁出现的、附带极少文本信息的助词、介词、连词、语气词等高频词。
- **以英文为例：**
  - the, is, at, which ……
- **以中文为例：**
  - 的、了、是 ……

| Nouns | Verbs | Adjectives | Prepositions | Others |
|---|---|---|---|---|
| 1. time | 1. be | 1. good | 1. to | 1. the |
| 2. person | 2. have | 2. new | 2. of | 2. and |
| 3. year | 3. do | 3. first | 3. in | 3. a |
| 4. way | 4. say | 4. last | 4. for | 4. that |
| 5. day | 5. get | 5. long | 5. on | 5. I |
| 6. thing | 6. make | 6. great | 6. with | 6. it |
| 7. man | 7. go | 7. little | 7. at | 7. not |
| 8. world | 8. know | 8. own | 8. by | 8. he |
| 9. life | 9. take | 9. other | 9. from | 9. as |
| 10. hand | 10. see | 10. old | 10. up | 10. you |
| 11. part | 11. come | 11. right | 11. about | 11. this |
| 12. child | 12. think | 12. big | 12. into | 12. but |
| 13. eye | 13. look | 13. high | 13. over | 13. his |
| 14. woman | 14. want | 14. different | 14. after | 14. they |
| 15. place | 15. give | 15. small | 15. beneath | 15. her |
| 16. work | 16. use | 16. large | 16. under | 16. she |
| 17. week | 17. find | 17. next | 17. above | 17. or |
| 18. case | 18. tell | 18. early | | 18. an |
| 19. point | 19. ask | 19. young | | 19. will |
| 20. government | 20. work | 20. important | | 20. my |
| 21. company | 21. seem | 21. few | | 21. one |
| 22. number | 22. feel | 22. public | | 22. all |
| 23. group | 23. try | 23. bad | | 23. would |
| 24. problem | 24. leave | 24. same | | 24. there |
| 25. fact | 25. call | 25. able | | 25. their |

# 样例

## 原始文本

- 文档1: Text mining is to identify useful information.
- 文档2: Useful information is mined from text.
- 文档3: Text mining is a subset of data mining.

## 文本预处理结果

- 文档1: text, mine, identify, useful, information
- 文档2: useful, information, mine, text
- 文档3: text, mine, subset, data, mine

# 讲授提纲

01 自然语言处理基本概念

02 文本数据预处理

03 文本表示向量空间模型

04 文本表示-主题模型&词向量模型

05 大语言模型及其进展

06 商务案例-在线用户评论分析

# 文本表示

## 形式化表示

- 使计算机能够高效处理

## 向量化表示方法

- 词集模型（Set of Words）
- 词袋模型（Bag of Words）
- 向量空间模型（Vector Space Model）
- 主题模型（Topic Models）
- 词向量模型（Word Embedding）

# 词集模型

- **定义：**
  - 单词构成的集合，考虑单词在文档中是否出现
- **文本预处理结果**
  - 文档1: text, mine, identify, useful, information
  - 文档2: useful, information, mine, text
  - 文档3: text, mine, subset, data, mine

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| 文档1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 文档2 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 文档3 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |

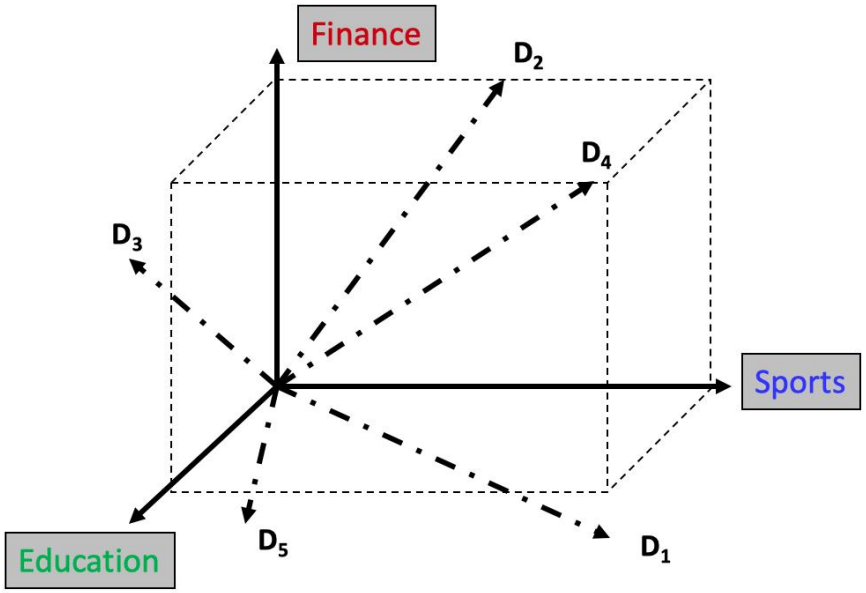
# 词袋模型

- **定义：**
  - 如果一个单词在文档中多次出现，统计其频数
- **文本预处理结果**
  - 文档1: text, mine, identify, useful, information
  - 文档2: useful, information, mine, text
  - 文档3: text, mine, subset, data, mine

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| 文档1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 文档2 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 文档3 | 1 | 2 | 0 | 0 | 0 | 1 | 1 |

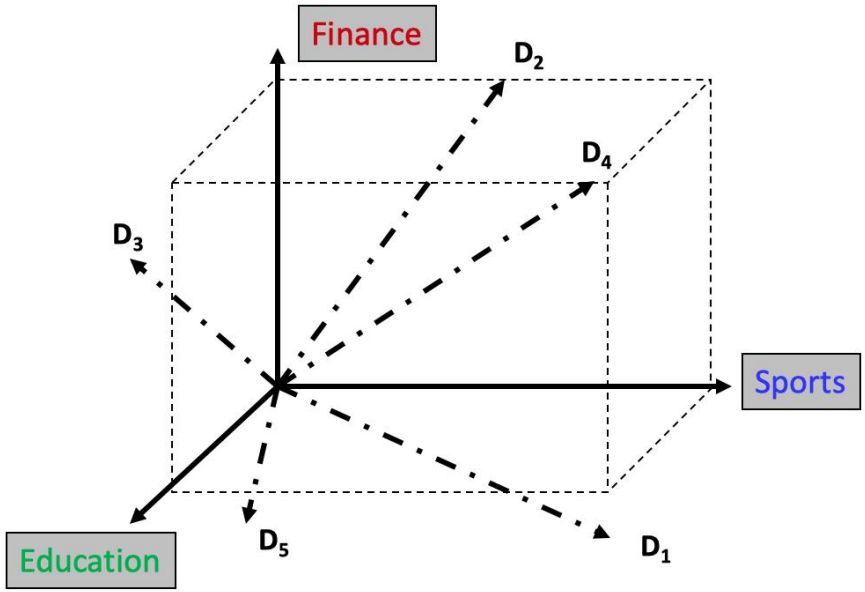
# 向量空间模型

- **目的：**
  - 将文档表示成实数特征向量
- **特征项：**
  - 语言单元，可以是字、词、词组、短语等
- **特征项权重：**
  - 表示特征项在文本中的重要性和相关性



# 特征项

- **各特征项之间：**
  - 互不相同
  - 相互独立
- **一元语法（Unigram）**
  - 词集&词袋模型
- **N元语法（N-gram）**




- **以2-gram为例**
  - 文档1: text-mine, mine-identify, identify-useful, useful-information
  - 文档2: useful-information, information-mine, mine-text
  - 文档3: text-mine, mine-subset, subset-data, data-mine

# 特征项权重

- **目的：**
  - 真实地反映文档的内容
  - 对不同文档有较好的区分能力
- **布尔权重**

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| … | … | … | … | … | … | … | … |
| 文档3 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |

- **特征频率（Term Frequency, TF）**

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| … | … | … | … | … | … | … | … |
| 文档3 | 1 | 2 | 0 | 0 | 0 | 1 | 1 |

# 特征项权重

## 特征频率（Term Frequency, TF）

- 该特征项在当前文档中出现的次数

$$TF(t,d) = N(t,d)$$

| | text | mine | identify | useful | information | subset | data | length |
|---|---|---|---|---|---|---|---|---|
| 文档1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 5 |
| 文档2 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 4 |
| 文档3 | 1 | 2 | 0 | 0 | 0 | 1 | 1 | 5 |

- 归一化
  - $TF(t,d) = \frac{N(t,d)}{\sum_t N(t,d)}$
  - $TF(t,d) = a + (1-a)\frac{N(t,d)}{\max\limits_t N(t,d)}, \text{if } N(t,d) > 0$

# 特征项权重

- **特征频率（Term Frequency, TF）**
  - 该特征项在当前文档中出现的次数
  - $TF(t,d) = N(t,d)$
  - 归一化
    - $TF(t,d) = \frac{N(t,d)}{\sum_t N(t,d)}$
    - $TF(t,d) = a + (1-a)\frac{N(t,d)}{\max_t N(t,d)}, \text{if } N(t,d) > 0$

- 当$a$=0.5时

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| 文档1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 文档2 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 文档3 | 0.75 | 1 | 0 | 0 | 0 | 0.75 | 0.75 |

# 特征项权重

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| 文档1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 文档2 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| 文档3 | 1 | 2 | 0 | 0 | 0 | 1 | 1 |

- **倒文档频率（Inverse Document Frequency, IDF）**
  - 文档频率（Document Frequency, DF）表示语料中包含特征项的文档数目
  - 一个特征项的DF越高，其包含的有效信息量往往越低
  - $IDF(t) = 1 + \log(\frac{N}{DF(t)})$

# 特征项权重

## 倒文档频率（Inverse Document Frequency, IDF）

- 文档频率（Document Frequency, DF）表示语料中包含特征项的文档数目
- 一个特征项的DF越高，其包含的有效信息量往往越低
- $IDF(t) = 1 + \log(\frac{N}{DF(t)})$

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| 文档1 | 1 | 1 | 2.099 | 1.405 | 1.405 | 2.099 | 2.099 |
| 文档2 | 1 | 1 | 2.099 | 1.405 | 1.405 | 2.099 | 2.099 |
| 文档3 | 1 | 1 | 2.099 | 1.405 | 1.405 | 2.099 | 2.099 |

# 特征项权重

## 特征频率-倒文档频率（TF-IDF）权重：

- 定义为TF和IDF的乘积
- $w(t,d) = TF(t,d) \times IDF(t)$

| | text | mine | identify | useful | information | subset | data |
|---|---|---|---|---|---|---|---|
| 文档1 | 1 | 1 | 2.099 | 1.405 | 1.405 | 0 | 0 |
| 文档2 | 1 | 1 | 0 | 1.405 | 1.405 | 0 | 0 |
| 文档3 | 1 | 2 | 0 | 0 | 0 | 2.099 | 2.099 |

- 对识别文档最有意义的特征项：
  - 在当前文档中出现频率足够高
  - 而在文档集合的其他文档中出现频率足够小

# 向量空间模型

- **优点：**
  - 简单、方便、快捷
  - 在语料充足的条件下，对简单的文本挖掘任务效果不错
    - 使罕见单词更加突出并且有效地忽略常用单词
- **缺点：**
  - 高维、高度稀疏
  - 假设词语之间相互独立，无法关注词语之间的顺序

# 讲授提纲

01 自然语言处理基本概念

02 文本数据预处理

03 文本表示向量空间模型

04 **文本表示-主题模型&词向量模型**

05 大语言模型及其进展

06 商务案例-在线用户评论分析

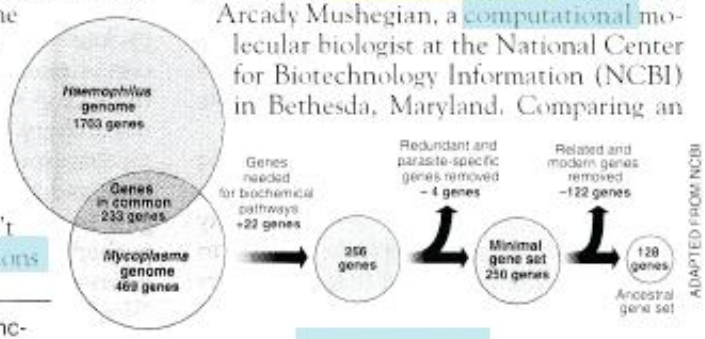
# 主题模型

- 目的：
  - 从文本语料中发现隐藏在词汇表面之下的潜在语义

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



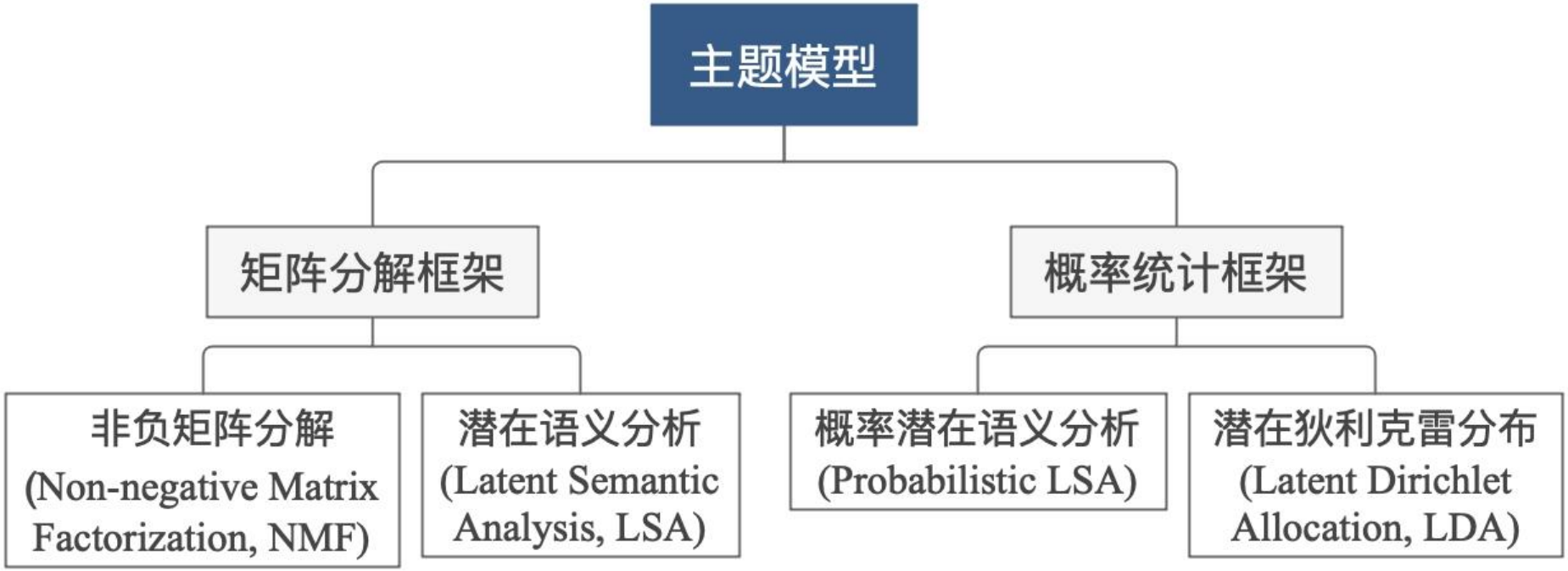

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

# 主题模型

- **本质：**
  - 高维词项空间 => 低维主题空间



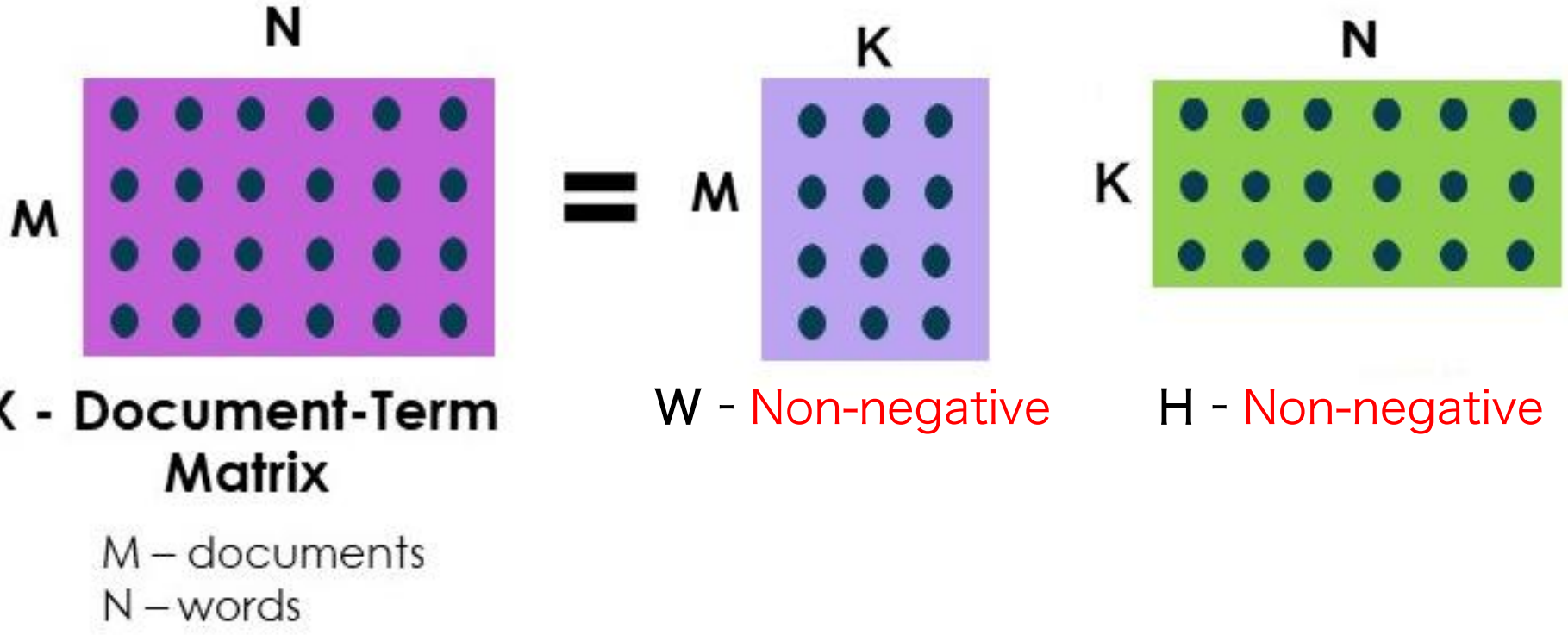
# NMF模型-非负矩阵分解

## 基本思想:

- 将非负的大矩阵分解成两个非负的小矩阵

$$\hat{X} = WH$$

- $\min_{W,H} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - (WH)_{ij})^2, W_{ia} \geq 0, H_{bj} \geq 0$



# LSA模型-潜在语义分析

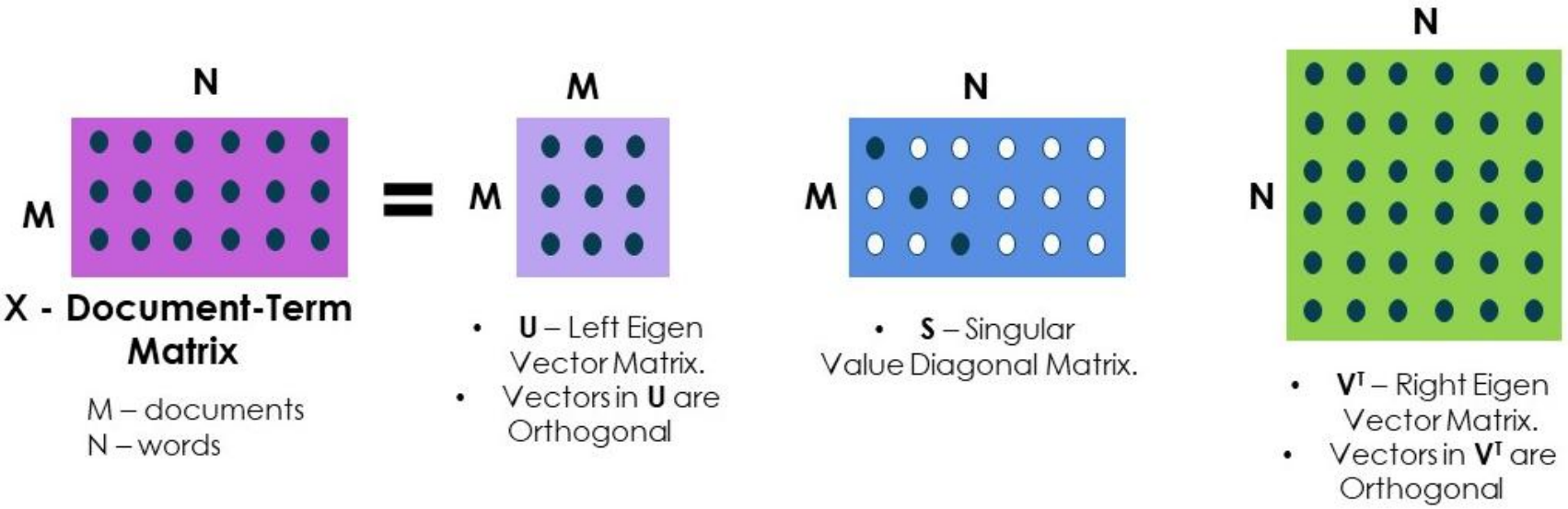
## 基本思想：

- 奇异值分解（Singular Value Decomposition, SVD）

$$X = USV^T,$$

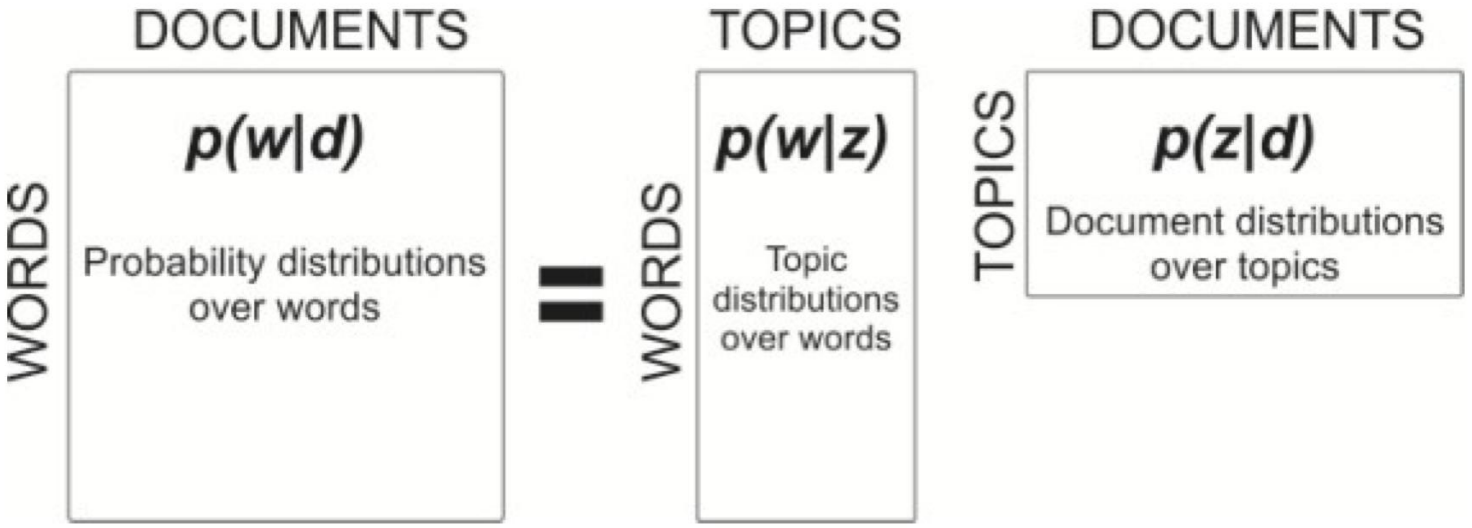
其中，$U^TU = I$、$V^TV = I$、S除了主对角线上的元素以外全为0

$$\hat{X} = U_K S_K V_K^T$$



# PLSA模型-概率潜在语义分析

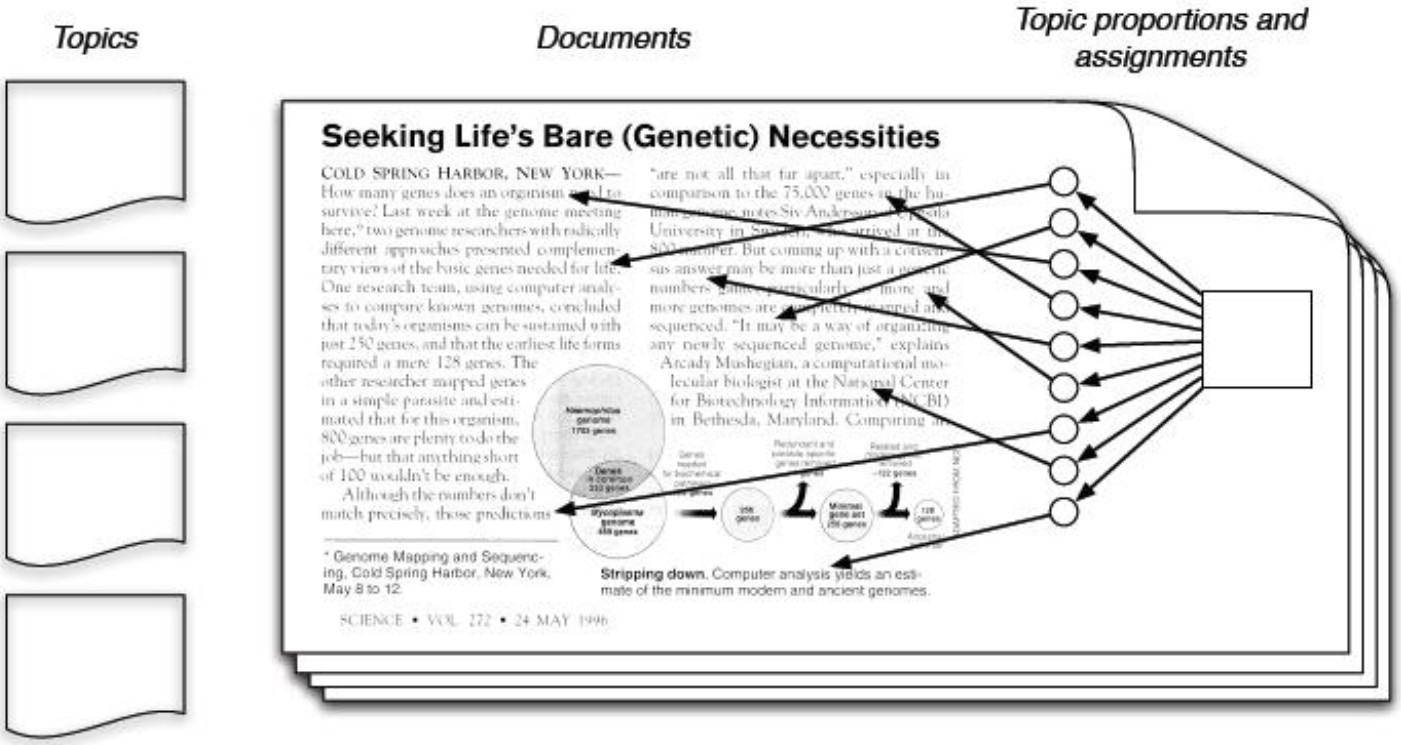
- **基本思想：**
  - 对于给定的观测数据，PLSA模型基于最大似然估计学习参数 $p(w_j|z_k)$ 和 $p(z_k|d_i)$ 的取值



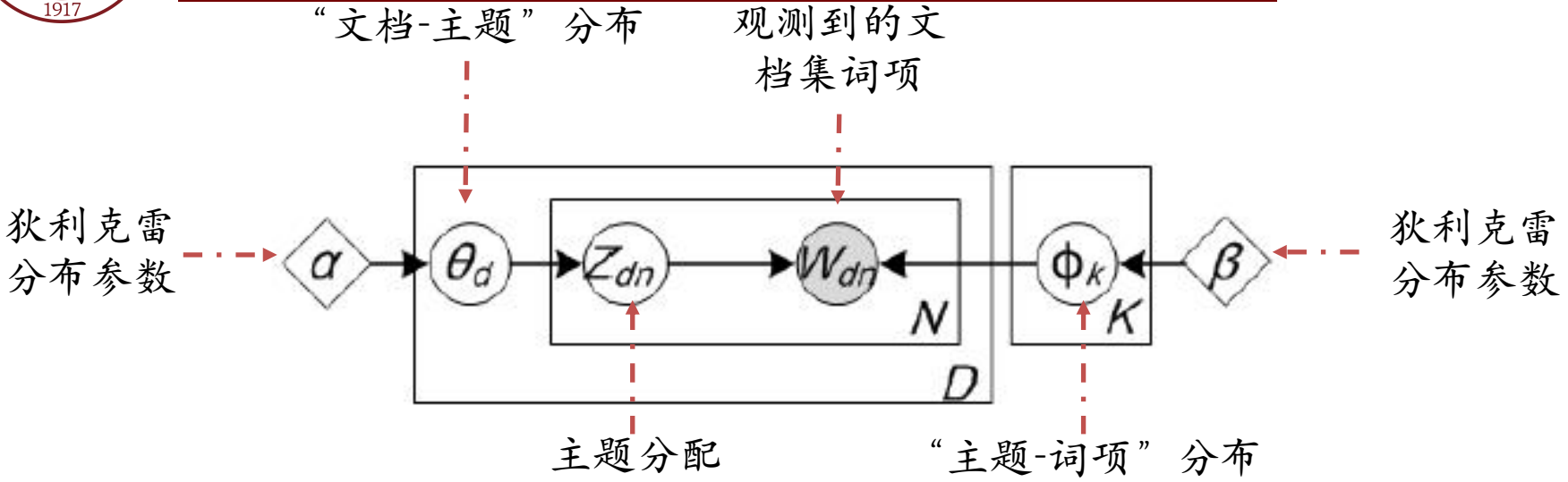
# LDA模型-潜在狄利克雷分布

## 基本思想：

- 将$\varphi_{kj} = p(w_j|z_k)$和$\theta_{ik} = p(z_k|d_i)$视为随机变量，并以狄利克雷分布作为参数的先验分布
- 最大似然估计 => 贝叶斯估计

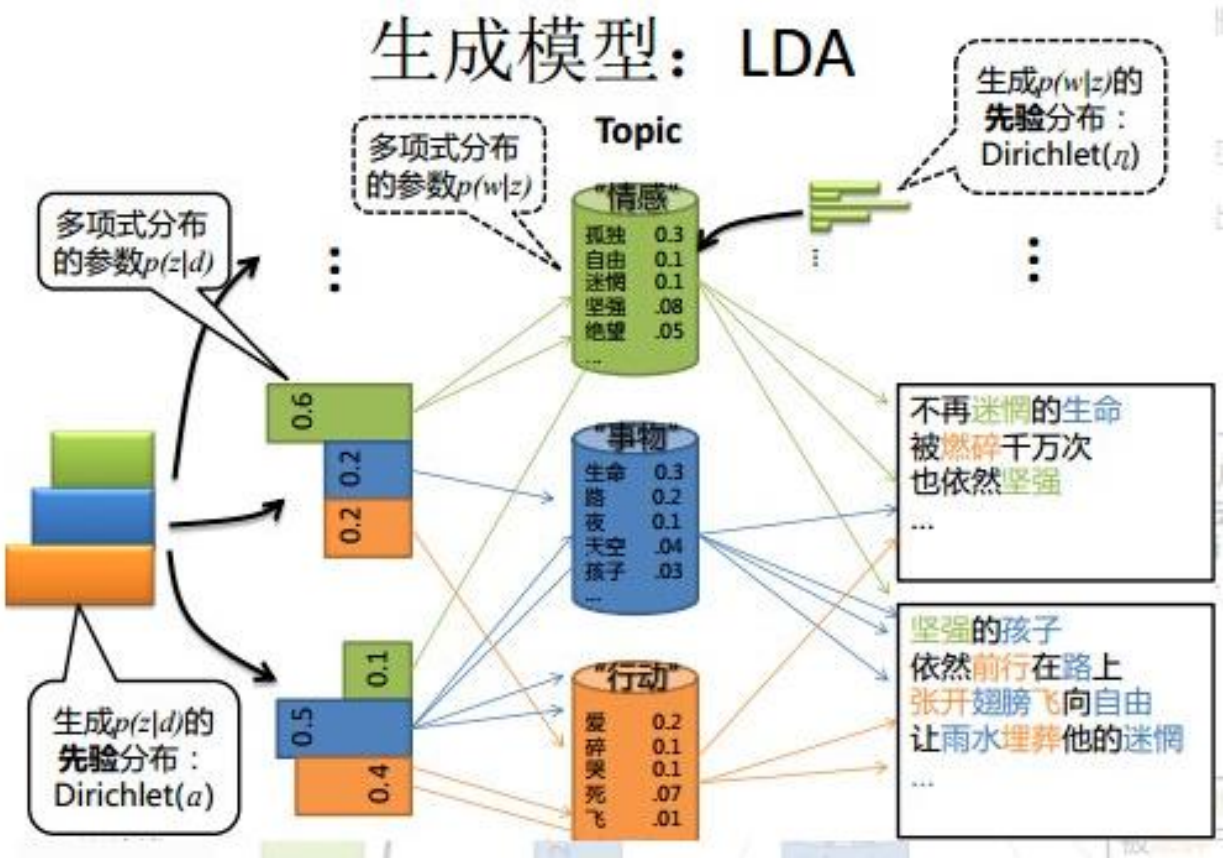


# LDA模型-生成过程




- 对每个主题 $z_k$
  - 生成“主题-词项”分布参数 $\phi_k \sim Dir(\beta)$
- 对每个文档 $d$
  - 生成“文档-主题”分布参数 $\theta_d \sim Dir(\alpha)$
  - 对当前文档的每个位置 $w$
    - 生成当前位置的所属主题 $z_{dn} \sim Multi(\theta_d)$
    - 根据当前位置的主题，以及“主题-词项”分布参数，生成当前位置对应的词项 $w_{dn} \sim Multi(\phi_{z_{dn}})$

# 示例


生成模型： LDA
Topic
多项式分布的参数p(w|z)
多项式分布的参数p(z|d)
生成p(w|z)的先验分布：Dirichlet(η)
生成p(z|d)的先验分布：Dirichlet(α)
"情感"
孤独 0.3
自由 0.1
迷惘 0.1
坚强 .08
绝望 .05
"事物"
生命 0.3
路 0.2
夜 0.1
天空 .04
孩子 .03
"行动"
爱 0.2
碎 0.1
哭 0.1
死 .07
飞 .01
0.6
0.2
0.2
0.1
0.5
0.4
不再迷惘的生命
被燃碎千万次
也依然坚强
...
坚强的孩子
依然前行在路上
张开翅膀飞向自由
让雨水埋葬他的迷惘
...

# 参数推断

- **潜变量：**
  - 主题分配 $z$
- **待推断变量：**
  - “主题-词项”分布 $\phi$ 和“文档-主题”分布 $\theta$
- **后验：**

$$p(\theta, \phi, Z \mid W, \alpha, \beta) = \frac{p(\theta, \phi, Z, W \mid \alpha, \beta)}{p(W \mid \alpha, \beta)}$$

- **联合分布：**

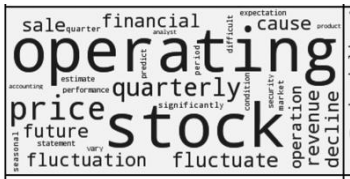
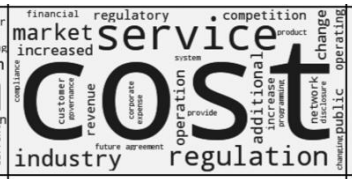
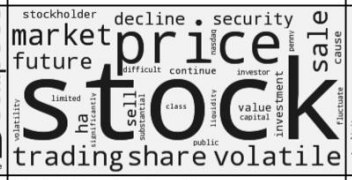
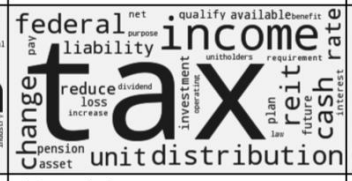
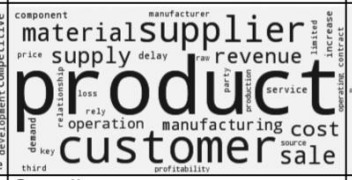
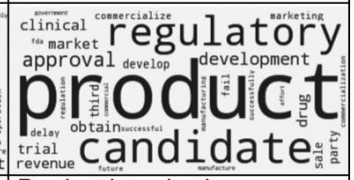
$$p(\Theta, \Phi, \mathbf{Z}, \mathbf{W} \mid \alpha, \beta) = \prod_{k=1}^{K} p(\Phi_k \mid \beta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \Big( \prod_{n=1}^{N} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \Phi) \Big)$$

# 近似后验推断

- 变分期望最大化算法
  - Variational Expectation Maximization (Blei et al., 2003)
- 期望传播算法
  - Expectation Propagation (Minka Lafferty, 2002)
- Gibbs采样算法
  - Collapsed Gibbs sampling(Griffiths and Steyvers, 2004)
- 在线变分推断算法
  - Online Variational Inference (Hofman et al., 2010)
- 分布式Gibbs采样算法
  - Distributed Gibbs sampling (Ahmed et al., 2012, Yuan et al., 2015)
- ……

# 实际应用

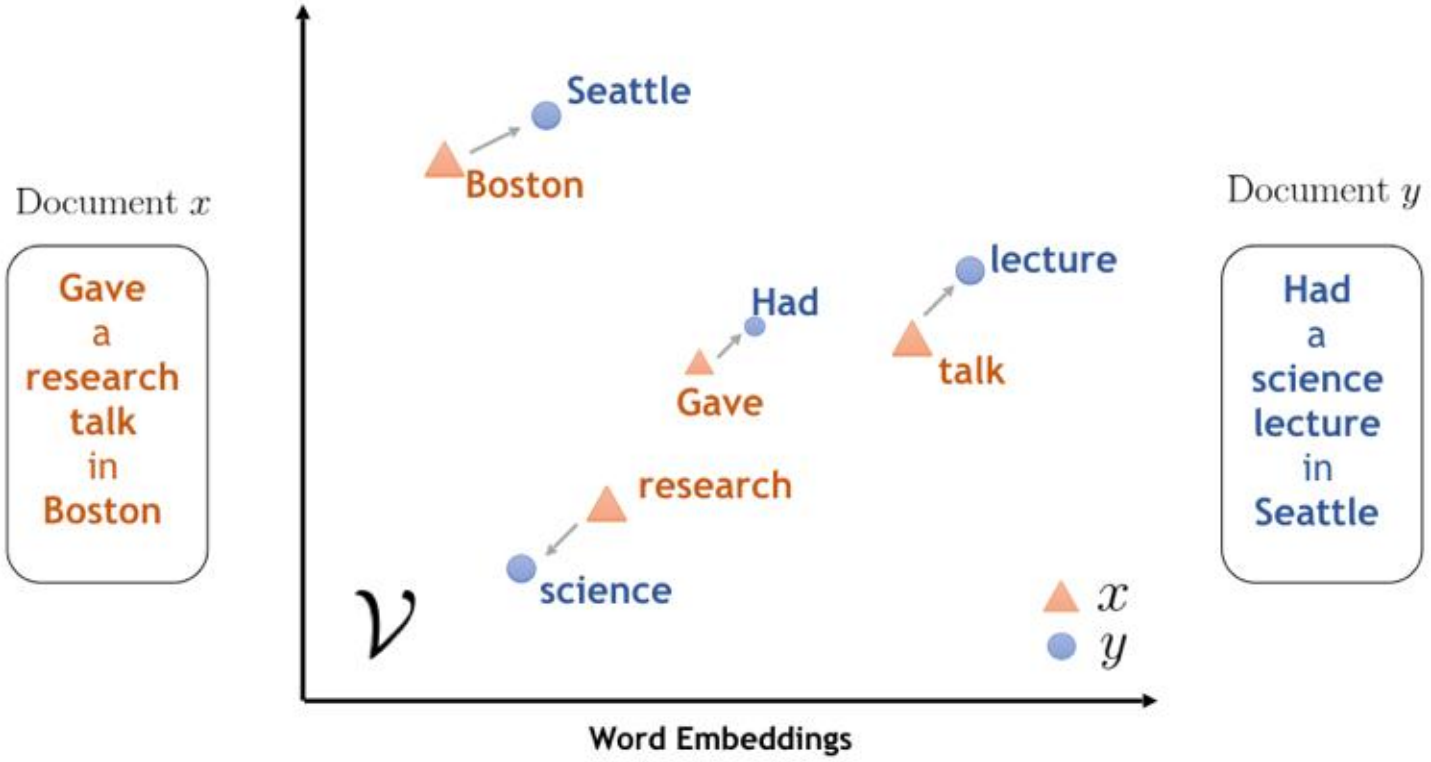
- Textual Risk Disclosures => Risk Types
  - Bao and Datta (2014)

| | | | | |
|---|---|---|---|---|
|  |  |  |  |  |
| *Volatile stock price* | *Shareholder's interest* | *Macroeconomic cyclical industry* | ***Cost risks** | *Rely on large customers* |
|  |  |  |  |  |
| *Competition* | *Volatile stock price* | ***Debt risks** | *Funding* | *Financial condition risks* |
|  |  |  |  |  |
| ***Property** | ***Investment** | *Regulation changes* | ***Tax risks** | *International risks* |
|  |  |  |  |  |
| ***Credit risks** | *Volatile demands product introduction* | *Suppliers* | ***Accounting risks** | *Product introduction* |

# 词向量模型

## 基本思想：

- 通过对词的上下文信息进行建模，将每个词映射为一个低维连续的实数向量
- 词嵌入（Word Embedding）

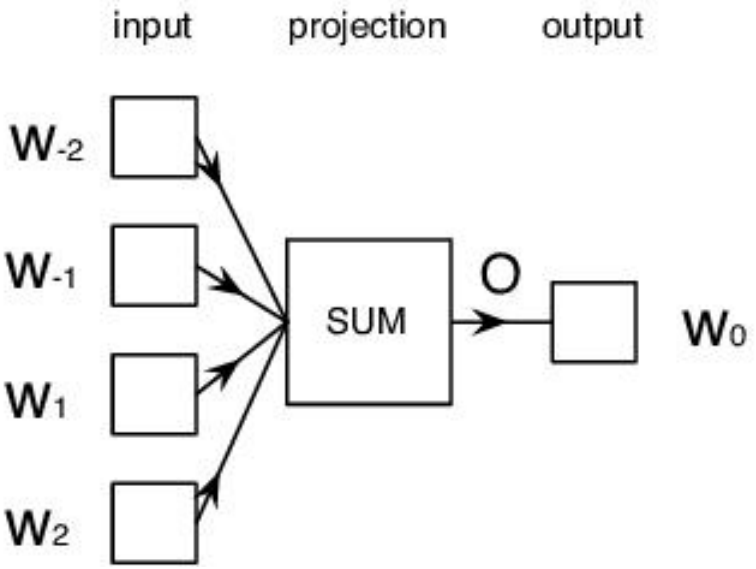


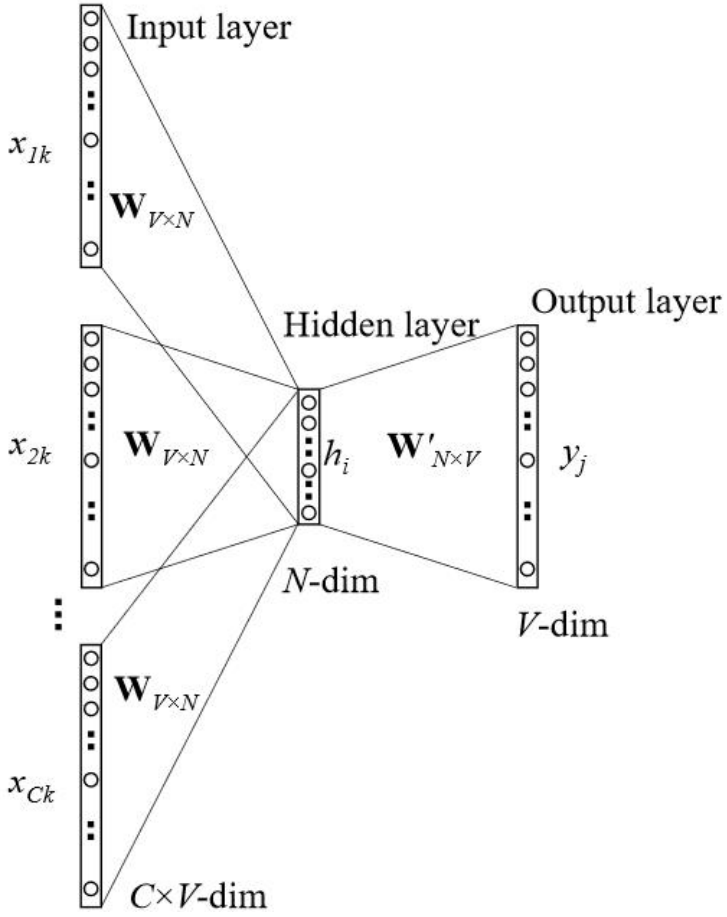
# 词向量模型

- **特点：**
  - 低维、稠密
- **word2vec**
  - 具有代表性的词向量模型之一
  - Efficient Estimation of Word Representation in Vector Space by Mikolov et al. (2013)
  - Continuous Bag-of-Words (CBoW)
    - 根据上下文单词预测中心词
  - Continuous Skip-gram (Skip-gram)
    - 通过中心词预测上下文单词

# CBoW

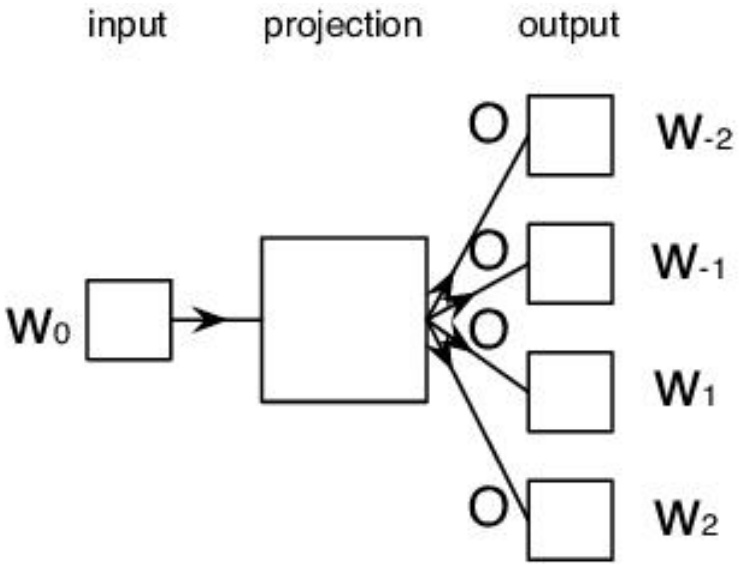
- Continuous Bag-of-Words (CBoW)
  - 通过附近的词预测中心词

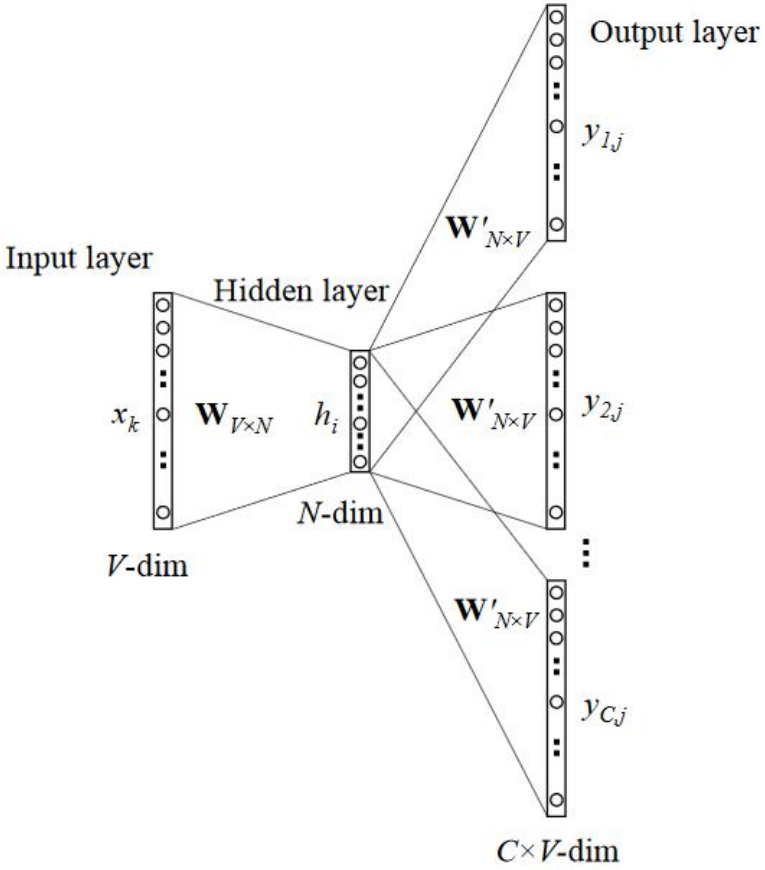


# Skip-gram

- Continuous Skip-gram (Skip-gram)
  - 通过中心词预测上下文单词



# 发展现状

- **GloVe: Global Vectors for Word Representation**
  - Pennington et al. (2014)
  - https://nlp.stanford.edu/projects/glove/
- **FastText**
  - Mikolov et al. (2018) <Advances in Pre-Training Distributed Word Representations>
  - https://fasttext.cc/
- **ELMo**
  - Peters et al. (2018) <Deep Contextualized Word Representations>
  - https://allennlp.org/elmo
- **BERT**
  - Devlin et al. (2019) <BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding>
  - https://arxiv.org/abs/1810.04805

# 讲授提纲

01 自然语言处理基本概念

02 文本数据预处理

03 文本表示向量空间模型

04 文本表示-主题模型&词向量模型

05 **大语言模型及其进展**

06 商务案例-在线用户评论分析

# 示例

写一个介绍chatgpt的文章

ChatGPT是一种基于深度学习的聊天机器人技术，它支持自然语言处理和语义理解，可以与用户进行自然语言对话。ChatGPT可以帮助公司构建聊天自动机器人，可以解答用户的提问，并且可以根据用户的问题和语境而提供自动响应，从而给用户带来友好而专业的服务。ChatGPT拥有强大和智能的语言处理技术，可以帮助企业构建智能聊天机器人，改善客户体验，提高工作效率，提供一流的客户服务。

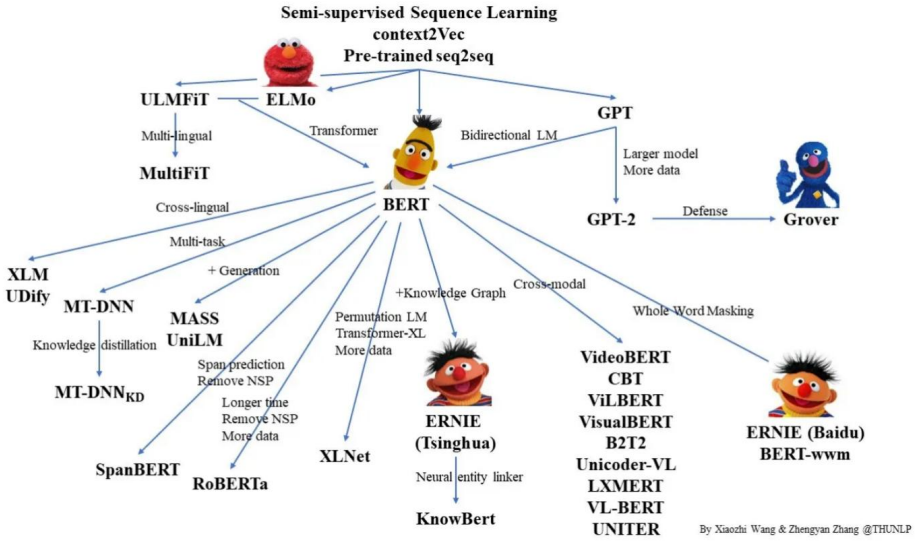
chatgpt的技术介绍

ChatGPT是一种基于深度学习的聊天机器人技术，它可以帮助企业构建智能聊天机器人，支持自然语言处理和语义理解，可以与用户进行自然语言对话，可以解答用户的提问，并且可以根据用户的问题和语境而提供自动响应。ChatGPT使用模型预测、大规模的多模态输入和多模态输出策略等技术，可以实现更为智能和准确的聊天机器人。此外，ChatGPT也具有连续领域知识的表示能力，它可以更准确地理解复杂的上下文，更好地服务于用户，提高工作效率，提供一流的客户服务。

# 发展阶段

- 语言模型本质： 出现的语句是否合适合理?
- 三个发展阶段:
  - 专家语法规则模型 （至 80年代）
    - 在计算机初始阶段，随着计算机编程语言的发展，归纳出的针对自然语言的语法规则。但是自然语言本身的多样性、口语化，在时间、空间上的演化，及人本身强大的纠错能力，导致语法规则急剧膨胀，不可持续。
  - 统计语言模型 （至 00年）
    - 统计语言模型通过对句子的概率分布进行建模，就是计算一个句子的概率大小的模型，P(w1；w2；…；wm)
  - 神经网络语言模型 （till Now）
    - 词向量代替 ngram，采用连续变量(具有一定维度的实数向量)来进行单词的分布式表示，解决了维度爆炸的问题，同时通过词向量可获取词之间的相似性。
    - 通过网络的叠加和特征的逐层提取，表征除词法外，相似性，语法，语义，语用等多方面的表示。

# 预训练语言模型




预训练语言模型的脉络

经典模型：ELMO，GPT，Bert……

训练范式：

| | Unidirectional language model | Bidirectional language model | Sequence-to-sequence model |
|---|---|---|---|
| Architecture | Transformer decoder | Transformer encoder | Transformer |
| Pre-training | Language modeling (2) | Mask language modeling (3) | Sequence-to-sequence learning |
| Tasks | Language generation | Language understanding | Sequence-to-sequence |
| Models | GPTs[3,25,26] | BERT,[8] RoBERTa,[17] ALBERT,[14] XLNet,[36] Electra[7] | BART,[15] T5[24] |

# 预训练语言模型 -> 大语言模型

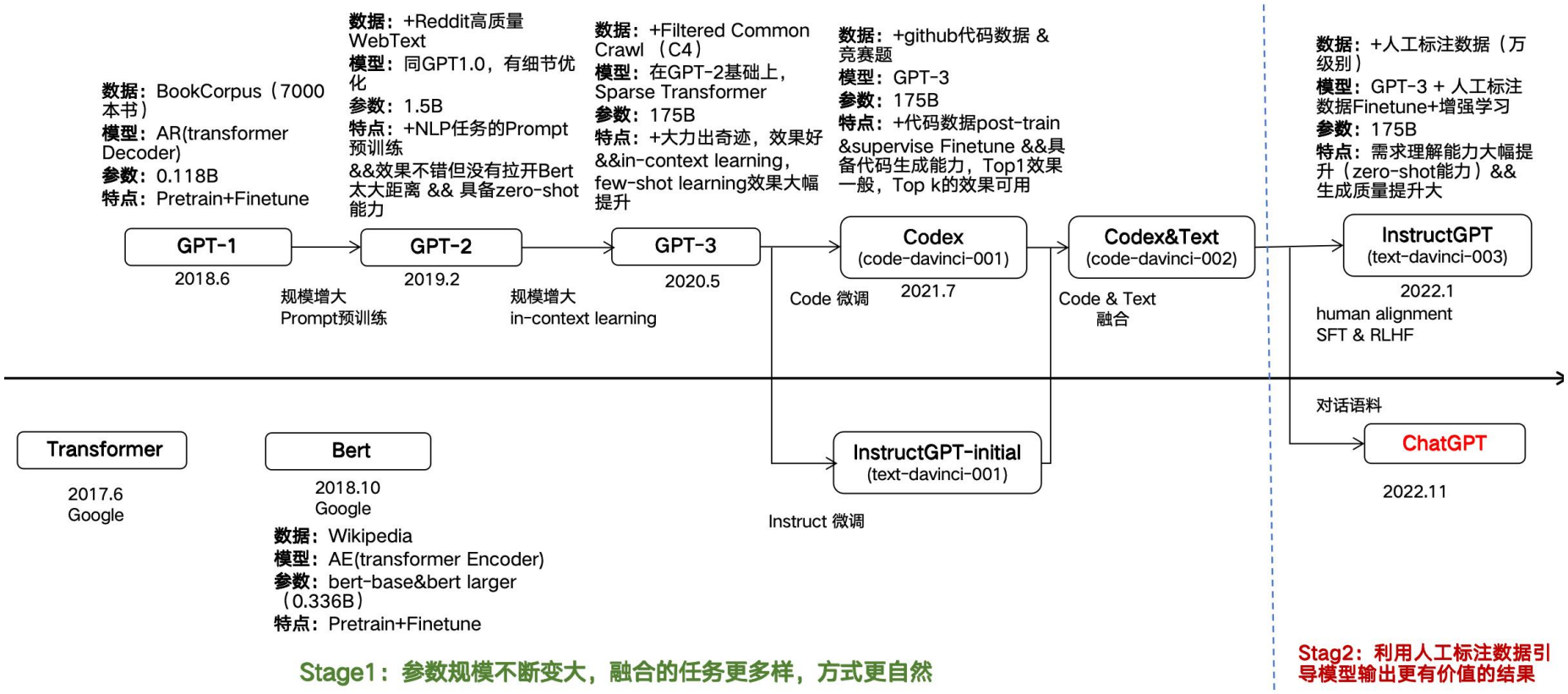
- **大规模语言模型：** 通常指参数量超过 10B 的模型
  - 更多的计算量、推理开销更大
  - 泛化性能更强，出现涌现能力

| | 预训练语言模型<br>（小模型、常规模型） | 大规模生成式语言模型 |
|---|---|---|
| 典型模型 | ELMo，BERT，GPT-2 | GPT-3、ChatGPT、LLaMA |
| 模型结构 | BiLSTM，Transformer | Transformer |
| 注意力机制 | 双向、单向 | 单向 |
| 训练方式 | 去噪自编码模型 | 自回归生成 |
| 擅长任务类型 | 理解、判断 | 生成 |
| 模型规模 | 1-10亿级参数 | 10-1000亿级参数 |
| 下游任务应用方式 | 微调 | 微调 & 提示学习 |
| 涌现能力 | 小数据领域迁移 | 上下文学习，思维链提示 |

# 本质变化

- **量变到质变：从过拟合（overfitting）风险到欠拟合（underfitting）风险**
- **训练数据的变化：** 多元化
  - 不再仅仅是自然语言文本，而是多种数据的组合：自然语言文本、编程代码、化学分子式，乃至基因序列，甚至图像
- **训练方式的变化：** 从判别式预训练（BERT为典型）全面转向生成式预训练（GPT为典型）
- **模型架构的变化：** 从双向Transformer转向单向Transformer（Decoder-only）
- **应用方式的变化：** 从微调走向更为友好的提示学习
  - 样本更少，从必须一定的标注样本，到少样本，乃至零样本
  - 提示学习的工作形式逼近人机对话形式

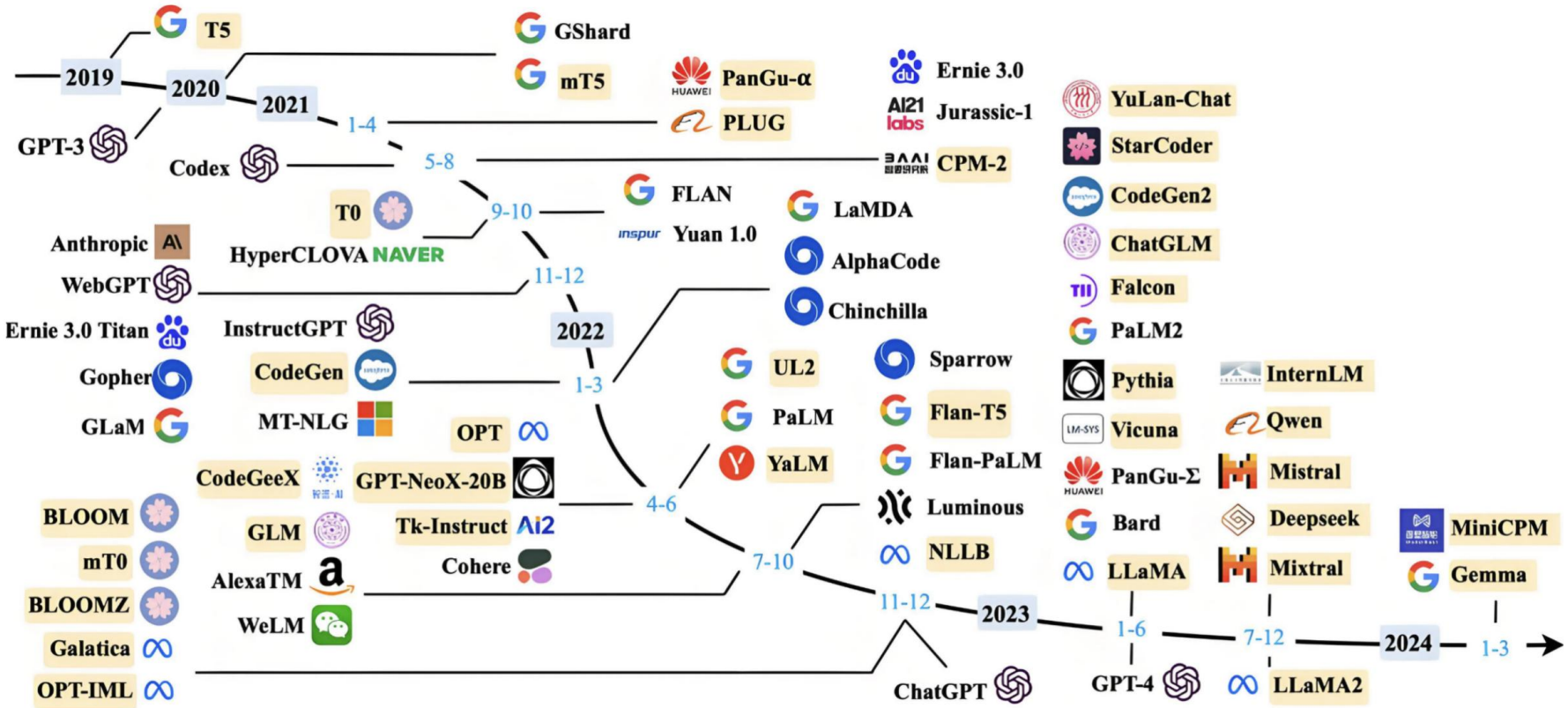
# ChatGPT 的演进概览



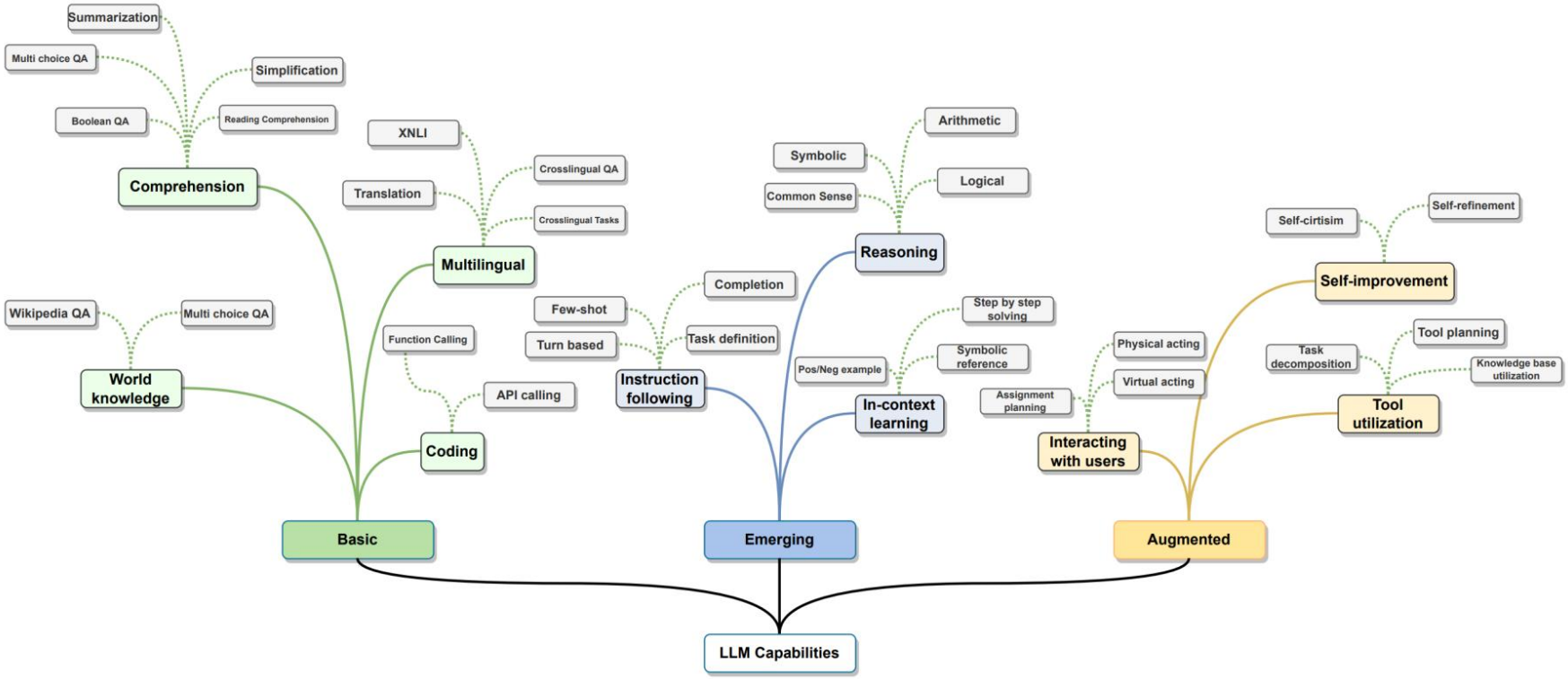
# 大语言模型的进化

| 模型 | 机构 | 参数量 | 训练数据量 | 方法和结论 | 文献 |
|---|---|---|---|---|---|
| GPT3 | OpenAI | 0.1B~175B | 约500B tokens | Transformer Decoder | Language Models are Few-Shot Learners |
| LaMDA | Google | 137B | 1.56T words | Transformer Decoder三大目标：质量、安全和根基性（事实正确性）。质量分为合理性、特异性和趣味性；主要根据以上评测指标来约束生成，将生成和排序融合到一起，同时增加了两个任务来融入知识（输入对话上下文，输出知识查询语句；输入知识查询语句，输出生成的最终结果） | LaMDA: Language Models for Dialog Applications |
| WebGPT | Open AI | 760M、13B、175B | Demonstraions: 6209 Comprisons:21548 | 其核心思想是使用GPT3模型强大的生成能力，学习人类使用搜索引擎的一系列行为，通过训练奖励模型来预测人类的偏好，使WebGPT可以自己搜索网页来回答开放域的问题，而产生的答案尽可能满足人类的喜好。 | WebGPT: Browser-assisted question-answering with human feedback |
| Sparrow | Deep Mind | 70B | / | 核心为从人类反馈中学习，创造更安全的对话助手。 | Improving alignment of dialogue agents via targeted human judgements |
| FLAN-T5 | Google | 540B | 1800个任务 | 任务的指令 与数据进行拼接。统一的输入输出格式（4种类型），引入chain-of-thought，大幅提高任务数量，大幅提高模型体积； | Scaling Instruction-Finetuned Language Models |
| Gopher | Deep Mind | 44M~ 280B | 10.5TB | 堆参数的大模型 | Scaling Language Models: Methods, Analysis & Insights from Training Gopher |
| PaLM | Google | 8B、62B、540B | 780B tokens 包括网页、书籍、维基百科、代码、社交对话 | Transformer Decoder | PaLM: Scaling Language Modeling with Pathways |
| InstructGPT | Open AI | 1.3B、6B、175B | 微调数据1w+，Reward Model 4w+，PPO无标注数据4w+ | GPT3 Finetune+RLHF指令微调 | Training language models to follow instructions with human feedback |
| ChatGPT | Open AI | / | 推测和InstructGPT差不多 | GPT3.5 （codex基础上）Finetune+RLHF+解决对齐问题 | / |

# 大语言模型的产业实践


2019
2020
2021
2022
2023
2024
1-4
5-8
9-10
11-12
1-3
4-6
7-10
11-12
1-6
7-12
1-3
T5
GPT-3
Codex
GShard
mT5
PanGu-α
PLUG
Ernie 3.0
Jurassic-1
CPM-2
T0
HyperCLOVA
FLAN
Yuan 1.0
LaMDA
AlphaCode
Chinchilla
Anthropic
WebGPT
Ernie 3.0 Titan
Gopher
GLaM
InstructGPT
CodeGen
MT-NLG
OPT
CodeGeeX
GPT-NeoX-20B
GLM
Tk-Instruct
AlexaTM
Cohere
WeLM
BLOOM
mT0
BLOOMZ
Galatica
OPT-IML
UL2
PaLM
YaLM
Sparrow
Flan-T5
Flan-PaLM
Luminous
NLLB
LLaMA
ChatGPT
YuLan-Chat
StarCoder
CodeGen2
ChatGLM
Falcon
PaLM2
Pythia
Vicuna
PanGu-Σ
Bard
GPT-4
InternLM
Qwen
Mistral
Deepseek
Mixtral
LLaMA2
MiniCPM
Gemma

# 大语言模型的能力版图




Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, Jianfeng Gao. Large Language Models: A Survey. https://arxiv.org/pdf/2402.06196.pdf

# 讲授提纲


01 自然语言处理基本概念

02 文本数据预处理

03 文本表示向量空间模型

04 文本表示-主题模型&词向量模型

05 大语言模型及其进展

06 商务案例-在线用户评论分析

# 在线用户评论分析

- 整理自Yelp官方公开的商户、点评和用户数据
  - https://www.yelp.com/dataset
- 所有位于多伦多的餐馆截至2017年7月的评论数据(review_res.txt)

| 字段名称 | 字段描述 |
| --- | --- |
| user_id | 用户ID |
| business_id | 商户ID |
| date | 用户评论日期 |
| text | 用户评论内容 |
| stars | 用户评分星级，1星到5星 |

# Python – nltk

- https://www.nltk.org/

## Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

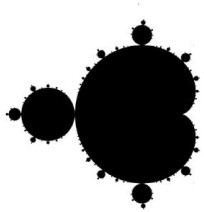
Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

# Python – TextBlob

- https://textblob.readthedocs.io/en/dev/



TextBlob

Star 6,936

TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

## Useful Links

TextBlob @ PyPI
TextBlob @ GitHub
Issue Tracker

## Stay Informed

Follow @sloria

## Donate

If you find TextBlob useful, please consider supporting its author:

## TextBlob: Simplified Text Processing

Release v0.15.2. (Changelog)

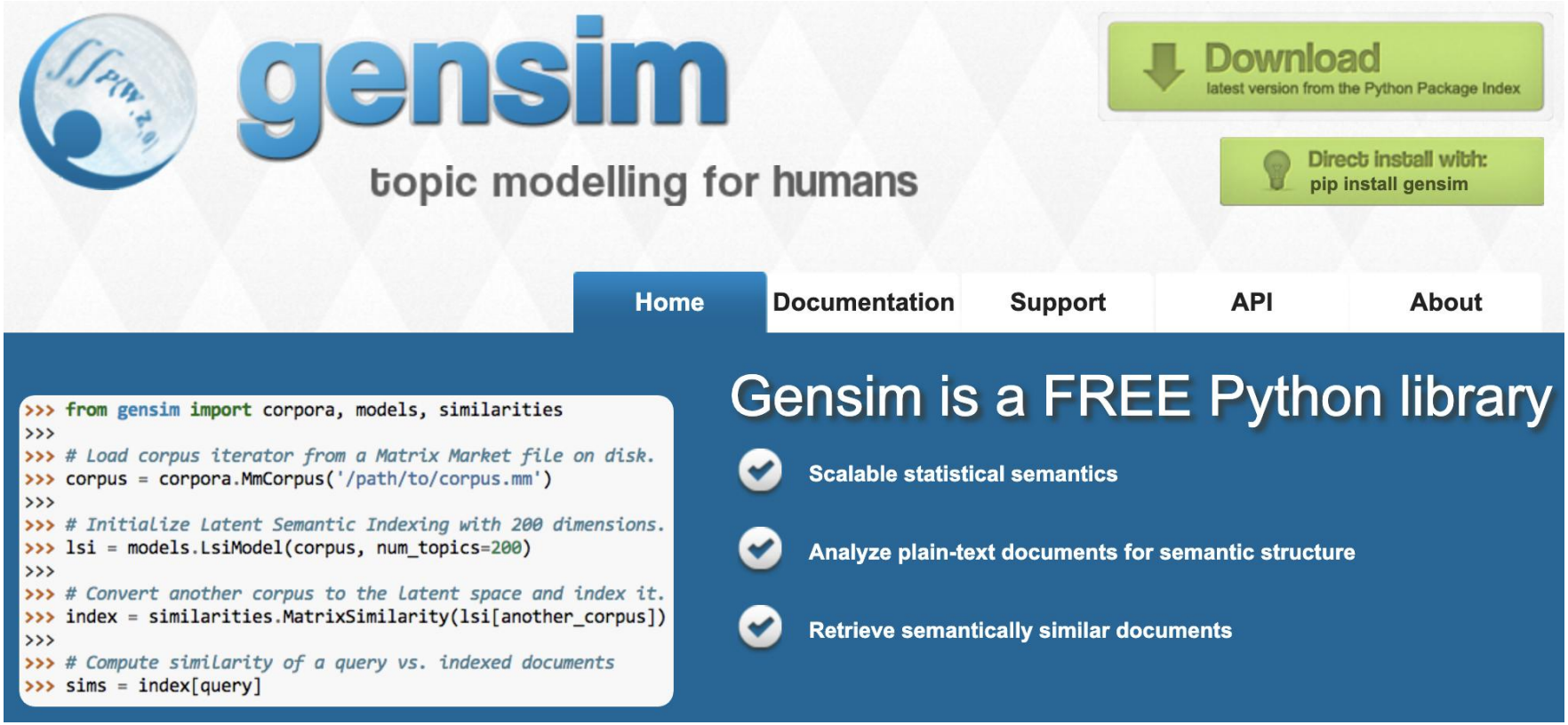
*TextBlob* is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

```
from textblob import TextBlob

text = '''
The titular threat of The Blob has always struck me as the ultimate movie
monster: an insatiably hungry, amoeba-like mass able to penetrate
virtually any safeguard, capable of--as a doomed doctor chillingly
describes it--"assimilating flesh on contact.
Snide comparisons to gelatin be damned, it's a concept with the most
devastating of potential consequences, not unlike the grey goo scenario
proposed by technological theorists fearful of
artificial intelligence run rampant.
'''

blob = TextBlob(text)
blob.tags           # [('The', 'DT'), ('titular', 'JJ'),
                    #  ('threat', 'NN'), ('of', 'IN'), ...]

blob.noun_phrases   # WordList(['titular threat', 'blob',
                    #            'ultimate movie monster',
                    #            'amoeba-like mass', ...])

for sentence in blob.sentences:
    print(sentence.sentiment.polarity)
# 0.060
# -0.341

blob.translate(to="es")  # 'La amenaza titular de The Blob...'
```

TextBlob stands on the giant shoulders of NLTK and pattern, and plays nicely with both.

# Python – gensim

- https://radimrehurek.com/gensim/



# Python - jieba

- https://github.com/fxsjy/jieba

"结巴"中文分词：做最好的 Python 中文分词组件

"Jieba" (Chinese for "to stutter") Chinese text segmentation: built to be the best Python Chinese word segmentation module.

- *Scroll down for English documentation.*

## 特点

- 支持四种分词模式：
  - 精确模式，试图将句子最精确地切开，适合文本分析；
  - 全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；
  - 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。
  - paddle模式，利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注。paddle模式使用需安装paddlepaddle-tiny，`pip install paddlepaddle-tiny==1.6.1`。目前paddle模式支持jieba v0.40及以上版本。jieba v0.40以下版本，请升级jieba，`pip install jieba --upgrade` 。PaddlePaddle官网
- 支持繁体分词
- 支持自定义词典
- MIT 授权协议