2.1 语言学基础 & NLP基础任务

2.2 信息检索和搜索引擎中的统计模型：TF-IDF

2.1 语言学基础 & NLP基础任务

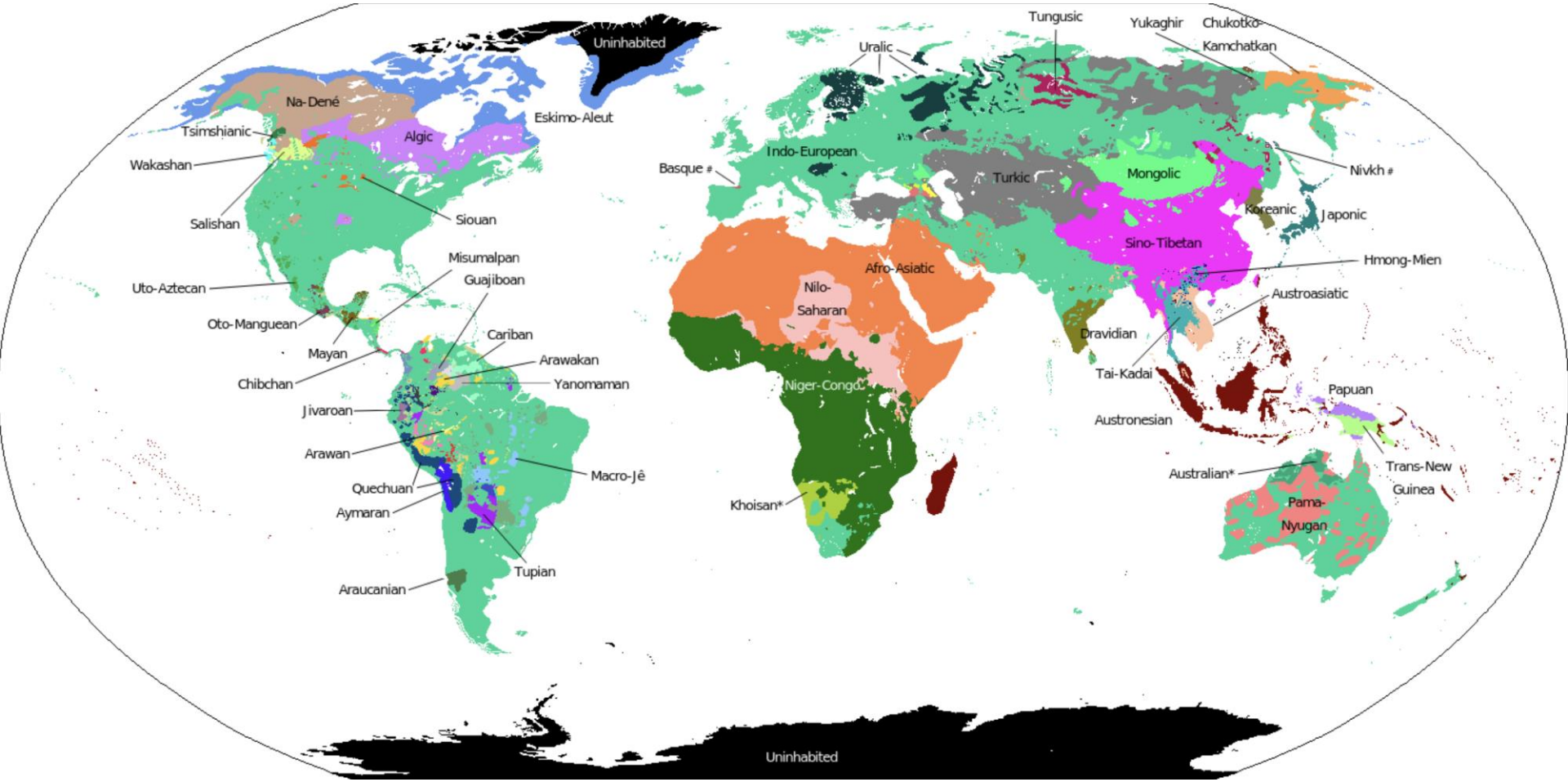
林洲汉
上海交通大学
2023年

目录

人类语言的独特地位

- **人类语言是独特的**
  - 相比于其他形式的交流系统，例如人类以外的动物所使用的动物语言。其他动物，比如蜂和猿所使用的交流系统都是封闭系统，其可表达的思想往往非常有限。
  - 而人类语言则相反，没有上限且富有创造性，允许人类从有限元素中产生大量话语，并创造新的词语和句子。
  - 人类语言异于动物的交流系统，还在于人类语言存在语法范畴，如名词和动词、现在和过去等，用来表达极其复杂的意义

世界语言

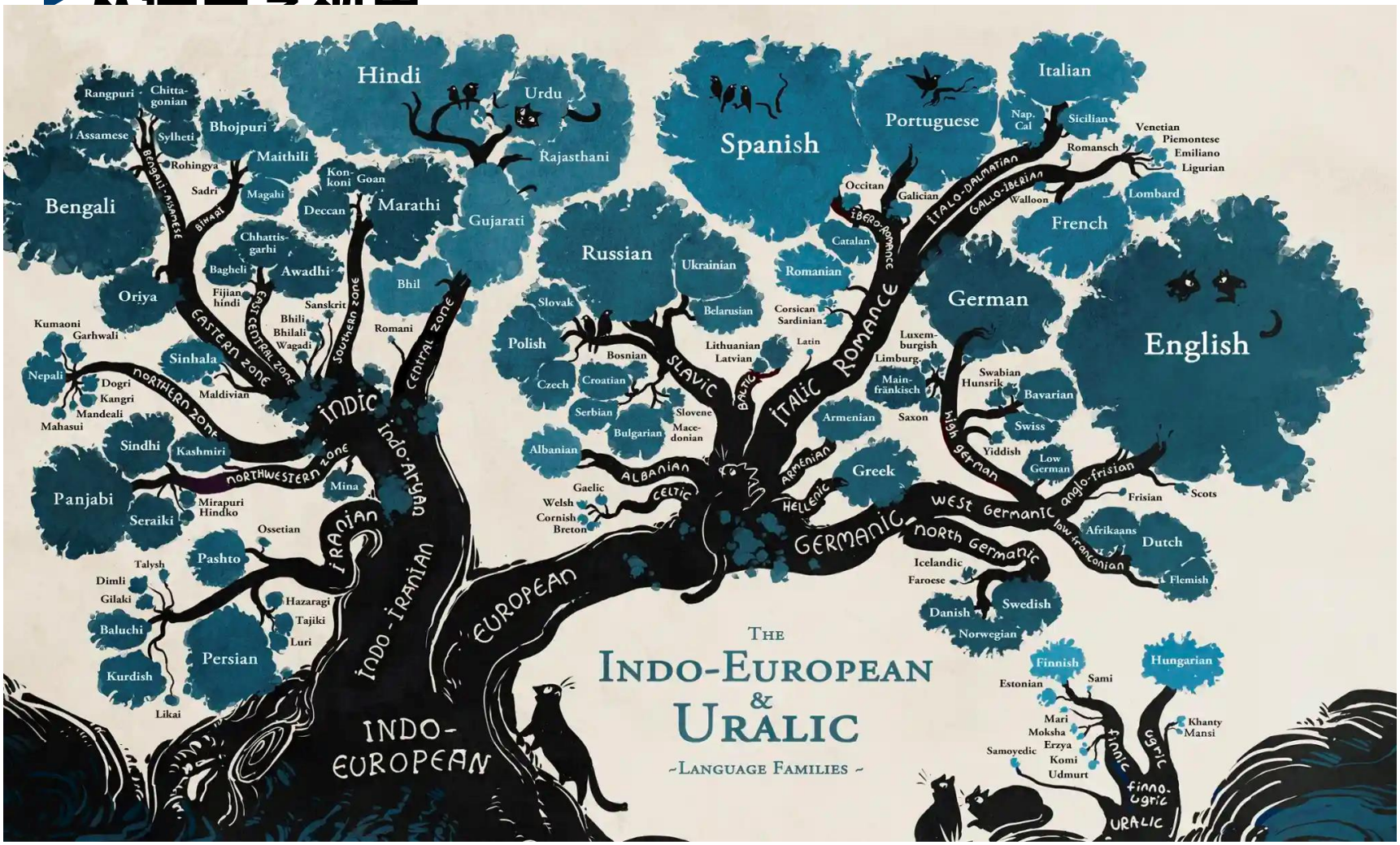

Uninhabited
Tungusic
Yukaghir
Chukotko-
Kamchatkan
Uralic
Na-Dené
Eskimo-Aleut
Tsimshianic
Algic
Wakashan
Basque
Indo-European
Turkic
Mongolic
Nivkh
Koreanic
Japonic
Salishan
Siouan
Sino-Tibetan
Misumalpan
Afro-Asiatic
Hmong-Mien
Guajiboan
Uto-Aztecan
Nilo-
Saharan
Austroasiatic
Oto-Manguean
Cariban
Dravidian
Mayan
Arawakan
Tai-Kadai
Chibchan
Yanomaman
Niger-Congo
Papuan
Jivaroan
Austronesian
Arawan
Trans-New
Guinea
Macro-Jê
Australian*
Quechuan
Khoisan*
Pama-
Nyugan
Aymaran
Tupian
Araucanian
Uninhabited

语言系属分类 （Language Family）

- **语言系属分类**
  - 是指根据语言的演化关系，对语言进行分类的方法，具有相同祖先的语言被归为一类，类似生物分类法。
  - 分类依据为各语言语音、词汇、语法之间的对应特征和演变规律。
- **语系（language family）**
  - 语族
    - 语支
- **英语属于**
  - 印欧语系
    - 日耳曼语族
      - 西日耳曼语支

以印欧语系为例

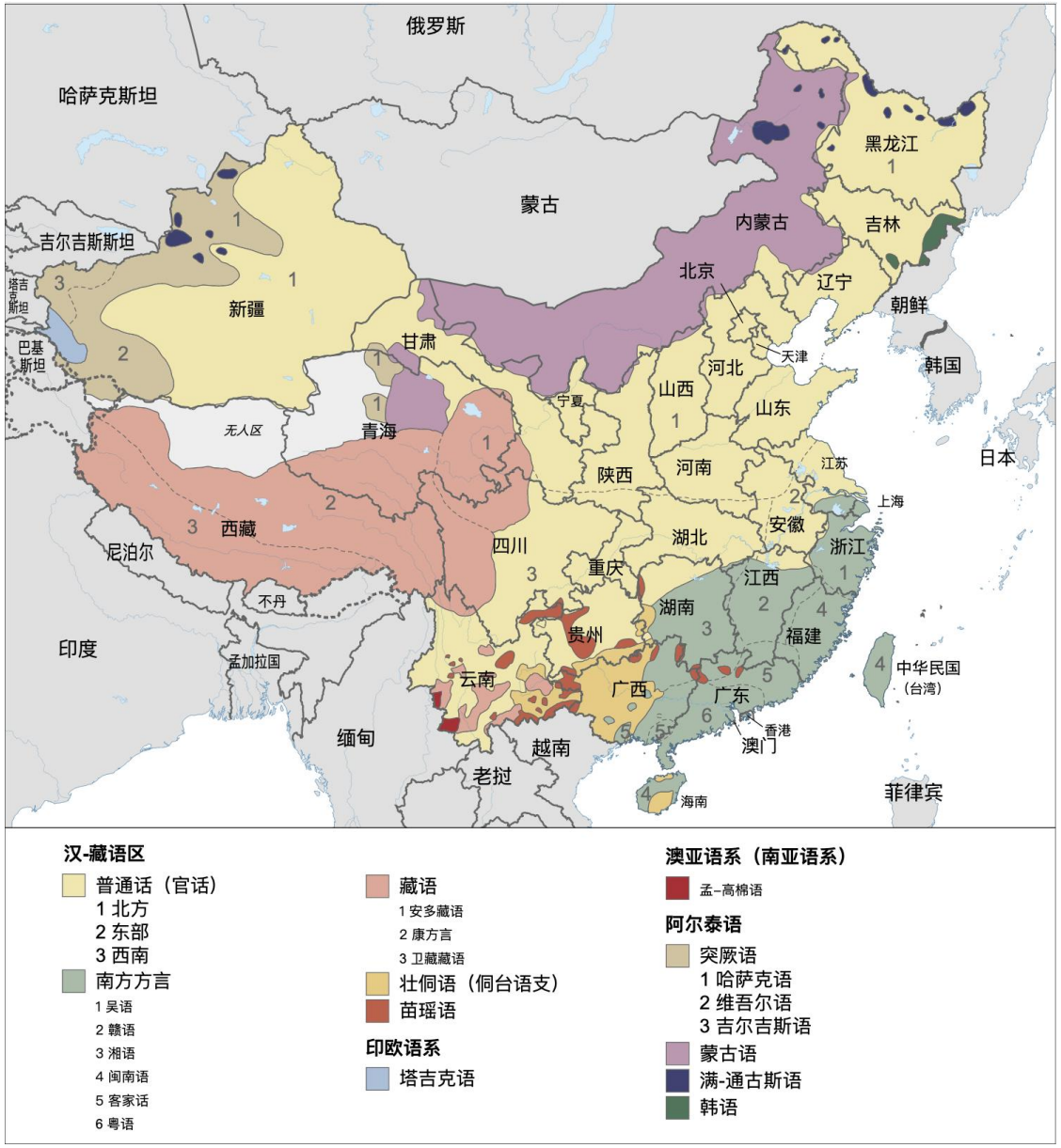
- 从语言学视角



以印欧语系为例

- **从地理的视角**

中国的语言


俄罗斯
哈萨克斯坦
蒙古
吉尔吉斯斯坦
塔吉克斯坦
巴基斯坦
新疆
甘肃
青海
无人区
宁夏
内蒙古
北京
天津
河北
山西
山东
辽宁
黑龙江
吉林
朝鲜
韩国
日本
陕西
河南
江苏
上海
安徽
湖北
浙江
西藏
四川
重庆
江西
湖南
福建
贵州
尼泊尔
不丹
印度
孟加拉国
云南
广西
广东
香港
澳门
中华民国（台湾）
缅甸
越南
老挝
海南
菲律宾
汉-藏语区
普通话（官话）
1 北方
2 东部
3 西南
南方方言
1 吴语
2 赣语
3 湘语
4 闽南语
5 客家话
6 粤语
藏语
1 安多藏语
2 康方言
3 卫藏藏语
壮侗语（侗台语支）
苗瑶语
印欧语系
塔吉克语
澳亚语系（南亚语系）
孟-高棉语
阿尔泰语
突厥语
1 哈萨克语
2 维吾尔语
3 吉尔吉斯语
蒙古语
满-通古斯语
韩语

这个对NLP有什么意义?

- **例如，对于机器翻译，语言间的相似性就很关键**

	Ar	Az	Be	De	He	It	Nl	Ro	Sk	Tr	Avg.
baselines	23.34	16.3	21.93	30.18	31.83	**36.47**	36.12	34.59	25.39	27.13	28.33
many-to-one	**26.04**	**23.68**	**25.36**	35.05	**33.61**	35.69	**36.28**	36.33	28.35	**29.75**	**31.01**
many-to-many	22.17	21.45	23.03	**37.06**	30.71	35.0	36.18	**36.57**	**29.87**	27.64	29.97

Table 5: X→En test BLEU on the 103-language corpus

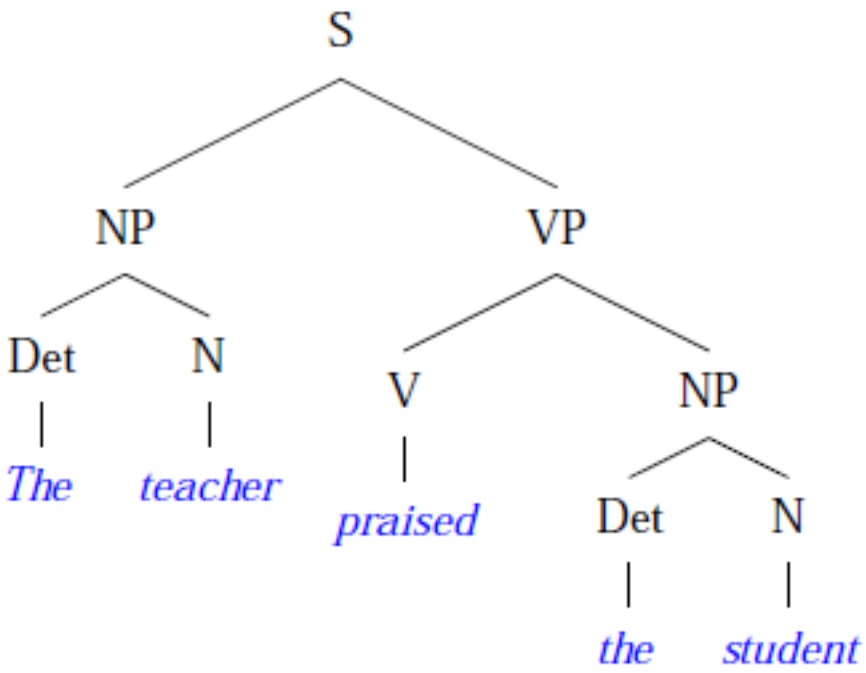
Massively Multilingual Neural Machine Translation, NAACL-2019

语言学

语言单位	语法	语义
词	词法	词义
句子	句法	句义
篇章	篇章语法	篇章语义

构成式语法

- **深层结构**
  - S-> NP + VP
  - NP-> Det + N
  - VP-> V + NP
  - NP -> Det + N
- **表层结构**
  - Det -> The
  - N -> Teacher
  - V -> Praise
  - Det -> the
  - N ->Student



依存式语法

- **依存句法**认为“谓语”中的动词是一个句子的中心，其他成分与动词直接或间接地产生联系。
- 依存语法存在一个共同的基本假设：句法结构本质上包含词和词之间的依存（修饰）关系。一个依存关系连接两个词，分别是核心词（head）和依存词（dependent）。依存关系可以细分为不同的类型，表示两个词之间的具体句法关系。

目录

词

- **词**是自然语言中能够独立运用的最小单位， 是自然语言处理的基本单位。
- **自动词法分析**就是利用计算机对自然语言的形态(morphology) 进行分析，判断词的结构和类别等。
- **词性或称词类**(Part-of-Speech, POS)是词汇最重要的特性，是连接词汇到句法的桥梁。

英语的形态分析：单词识别

◆英语单词的识别

例 (1) Mr. Green is a good English teacher.

(2) I'll see prof. Zhang home after the concert.

识别结果：

(1) Mr./ Green/ is/ a/ good/ English/ teacher/.

(2) I/ will/ see/ prof./ Zhang/ home/ after/ the/ concert/.

英语的形态分析：形态还原

- **有规律的：建立规则**
  - 例子：-ed 结尾的动词过去时，去掉 ed； (e.g., worked -> work)
- **无规律的：建立不规则变化词表**
  - 例子：choose, chose, chosen
- **对于表示年代、时间、百分数、货币、序数词的：数字形态还原**
  - 例：$20 去掉$，记录该数字为名词(20美圆);

词性标注 （2.4节：POS Tagging）

- **词性(part-of-speech, POS)标注(tagging)**的主要任务是 消除词性兼类歧义。在任何一种自然语言中，词性兼类问题都普遍存在。
- **英语**
  - 1) Time flies like an arrow.
  - 2) I want you to web our annual report.
- **汉语**
  - 形同音不同，如： “好(hao3，形容词)、好 (hao4，动词)” 这个人什么都好，就是好酗酒。
  - 同形、同音，但意义毫不相干，如： “会(会 议，名词)、会(能够、动词)” 每次他都会在会上制造点新闻。

词性标注集

- **UPenn Treebank 的词性标注集**
  - 33 类
  - NN （Noun, singular or mass） 名词、PRP （Personal pronoun ）代词，NNS复数名词，DT 定冠词 （Determiner ），IN （Preposition or subordinating conjunction ）介词

中文分词 （2.5节：中文分词）

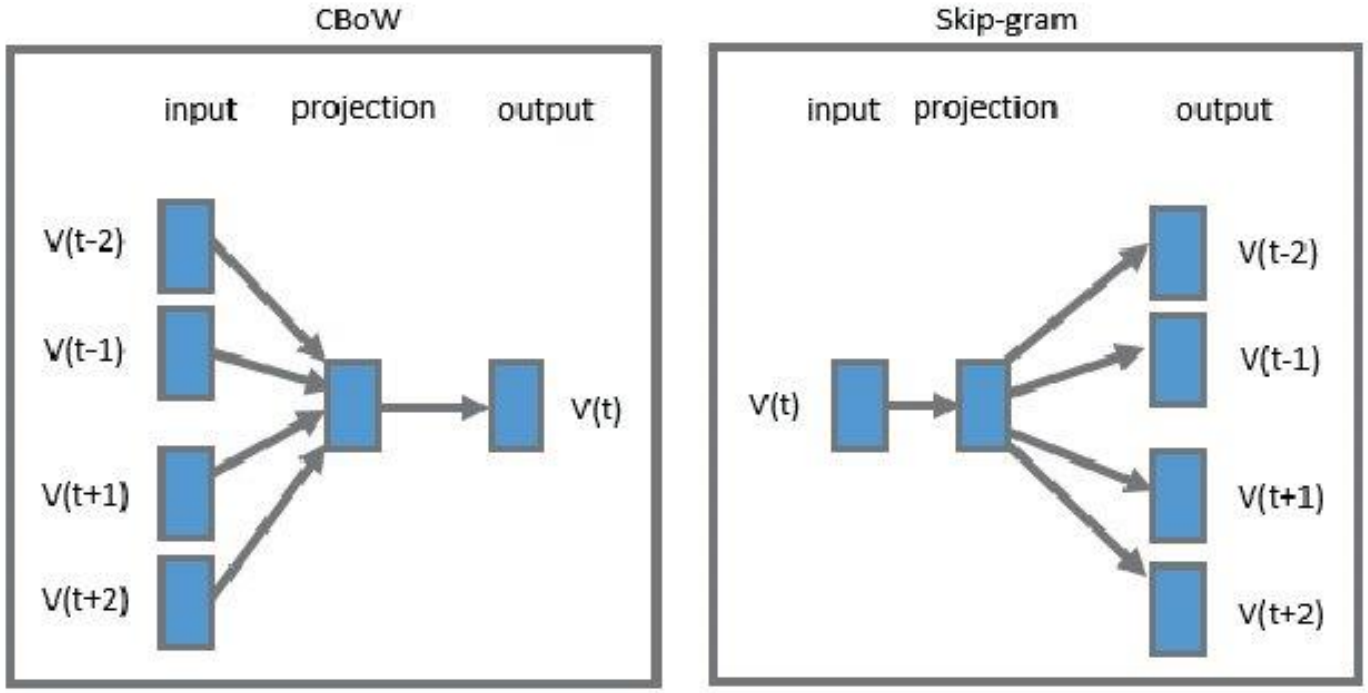
- **核心：歧义切分字段处理**
  - 例1：中国人为了实现自己的梦想 (交集型歧义)
    - 中国/ 人为/ 了/ 实现/ 自己/ 的/ 梦想
    - 中国人/ 为了/ 实现/ 自己/ 的/ 梦想
    - 中/ 国人/ 为了/ 实现/ 自己/ 的/ 梦想
  - 其他例子：“大学生”、“研究生物”、“从小学起”、“为人民工作”、“中国产品质量”、“部分

基本方法

- **词典切分**
  - 最大匹配
  - 最少分词法
- **基于统计的方法**
  - 基于语言模型的分词方法
  - 基于HMM的分词方法
  - 生成式方法

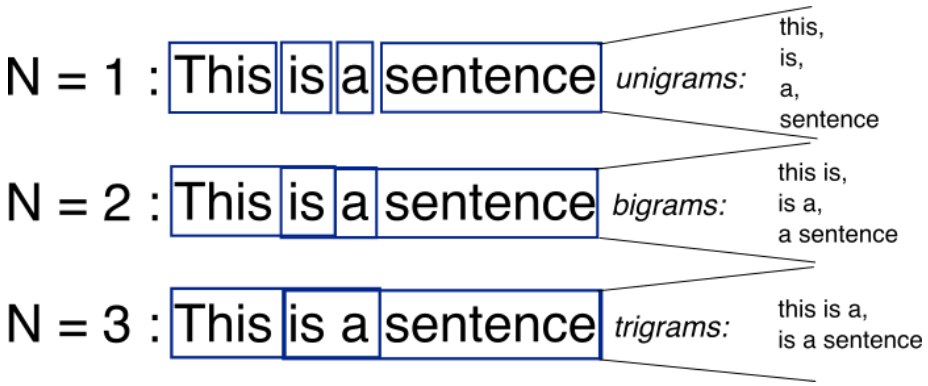
词表示 (3.1节：词向量表示)

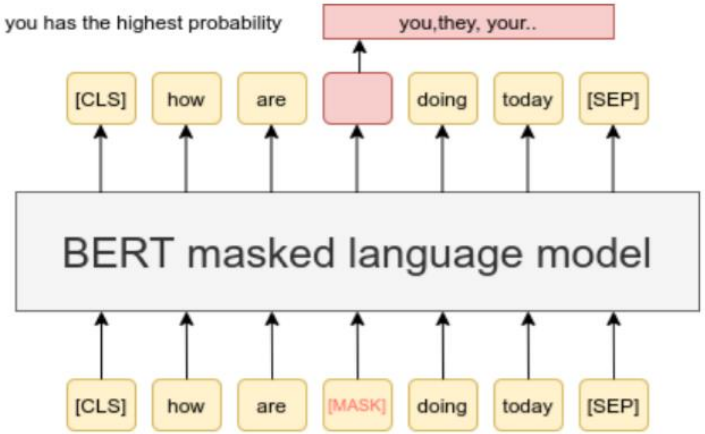
- **将词表示成空间维度上的向量**
  - 传统的潜在语义分析（Latent Semantic Analysis）
  - 基于神经网络的Word2Vec



词的预测：语言模型
(2.3节：n-gram语言模型 & 4.1节：预训练语言模型)

- **单向语言模型：给定上文中的单词，预测下一个单词**
- **双向语言模型：给定上下文中的单词，预测中间的词**





目录

句子级别的NLP任务

- **句子表示**
- **句子（文本）分类**
- **句子生成**
- **句法分析**

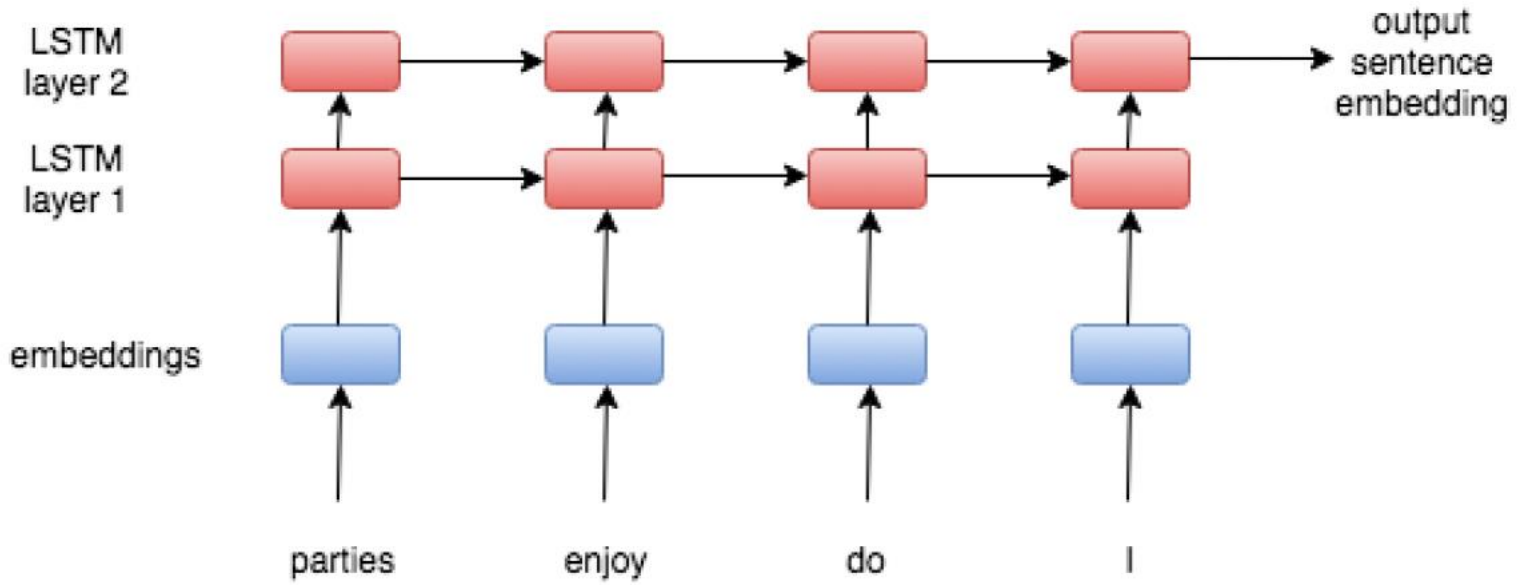
句子（文本）分类 （3.2节：神经语言模型及文本分类）

- **将表示好的句子（文本）分成各种类别**
  - 划分文本类别：例如专利数据归类
  - 情感分析：例如微博评论支持/反对某种观点

- *买没几天就降价一点都不开心，闪存跑分就五百多点点 ---* 
- *外观漂亮音质不错，现在电子产品基本上都是华为的了 ---* 
- *汽车不错，省油，性价比高 ---* 
- *这个政策好啊，利国利民 ---* 

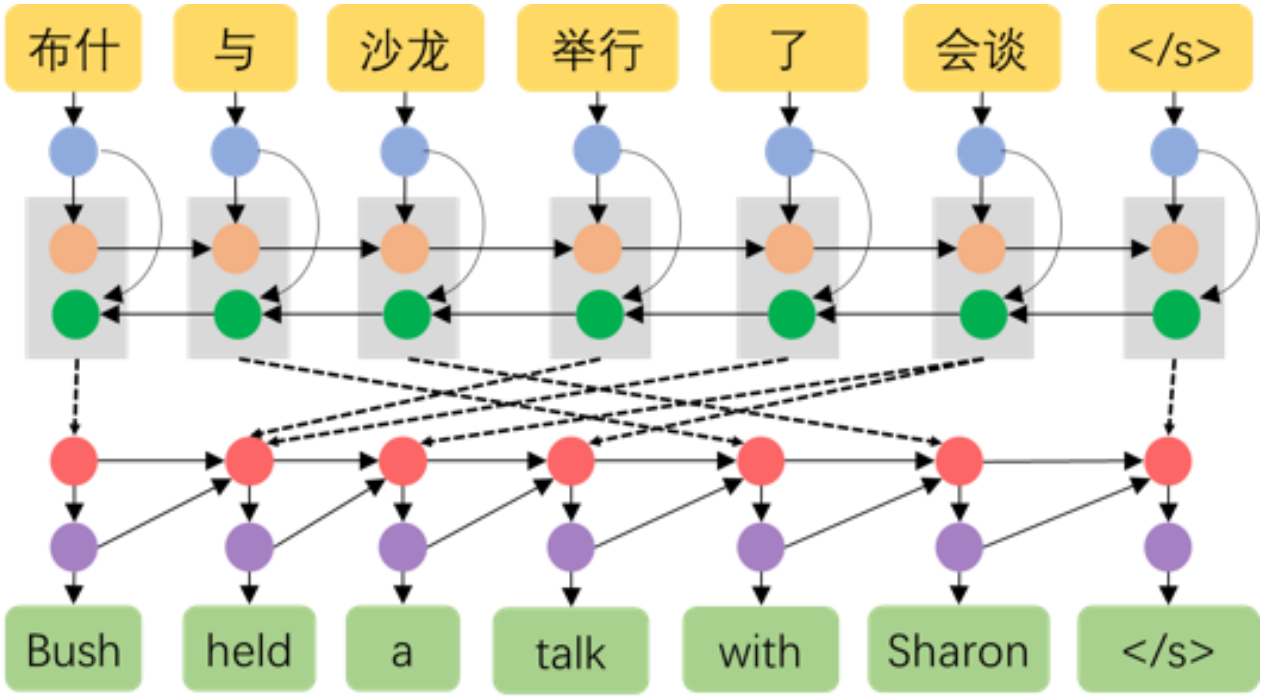
句子表示 （3.3节：序列标注任务）

- **句子级别的表示是词表示的集合，主要方式为：**
  - Bag-of-words
  - LSTM
  - Self-Attention



句子生成 （3.4节：机器翻译）

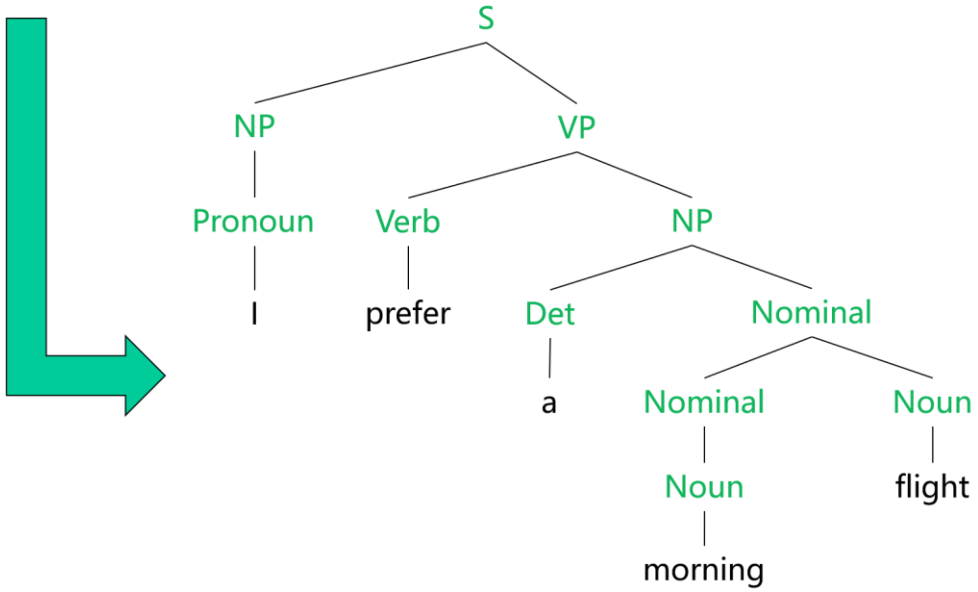
- **根据一段文本，生成一段新的句子**
  - 机器翻译
  - 转述 （paraphrase）



句法分析（2.6节：句法分析）

- **给定自然语言语句之后，由算法来分析语句的句法结构**

I prefer a morning flight.



目录

篇章级别

- **篇章分析 （discourse parsing）**
- **机器阅读理解 (machine reading comprehension)**
- **篇章机器翻译**

篇章分析 （discourse parsing）

- **分析各个句子之间的逻辑关系，例如**
  - elab = elaboration
  - QAP: Question-Answer-Pair”
  - Q·

(1) A: Anyone have sheep?

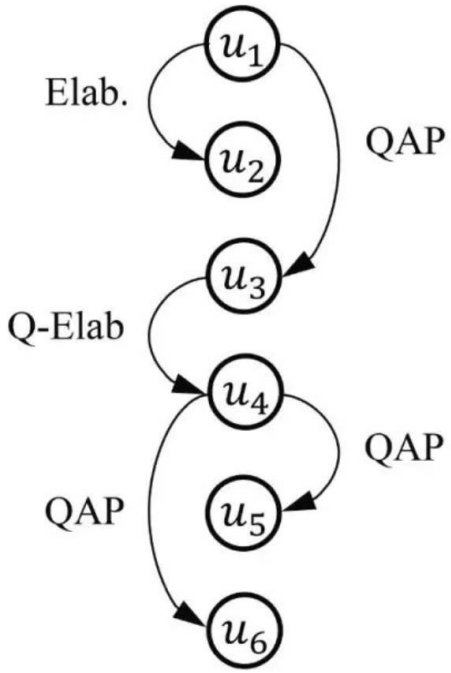
(2) A: I can give ore or wheat.

(3) B: I've got sheep as well.

(4) A: Need ore or wheat?

(5) C: I need wheat.

(6) B: Wheat.



Ground truth

机器阅读理解 （4.1节：智能问答 & 4.2节：对话系统）

- 机器阅读理解（MRC）是一项任务，用于测试机器通过要求机器根据给定的上下文回答问题来理解自然语言的程度。

段落

工商协进会报告，12 月消费者信心上升到 78.1，明显高于 11 月的 72。另据《华尔街日报》报道，2013 年是 1995 年以来美国股市表现最好的一年。这一年里，投资美国股市的明智做法是追着 “傻钱” 跑。所谓的 “傻钱” 策略，其实就是买入并持有美国股票这样的普通组合。这个策略要比对冲基金和其它专业投资者使用的更为复杂的投资方法效果好得多。

问题 1：什么是傻钱策略？

答案：买入并持有美国股票这样的普通组合

问题 2：12 月的消费者信心指数是多少？

答案：78.1

问题 3：消费者信心指数由什么机构发布？

答案：工商协进会

任务类别

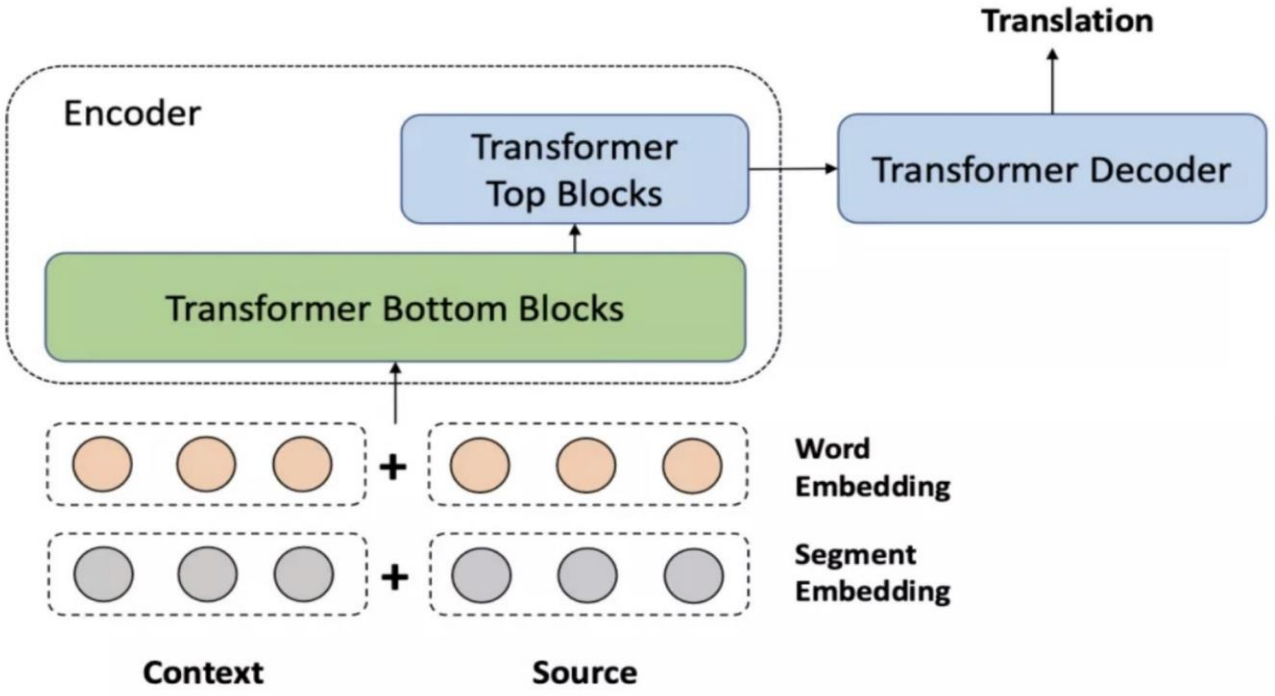
- 多项选择式，即模型需要从给定的若干选项中选出正确答案；
- 区间答案式，即答案限定是文章的一个子句，需要模型在文章中标明正确的答案起始位置和终止位置；
- 自由回答式，不限定模型生成答案的形式，允许自由

Cloze Test

CLOTH[93]	Context:	Comparisons were drawn between the development of television in the 20th century and the diffusion of printing in the 15th and 16th centuries. Yet much had happened __1__. As was discussed before, it was not __2__ the 19th century that the newspaper bec -ame the dominant pre-electronic __3__, following in the wake of the pamphlet and the book and in the __4__ of the periodical.· · ·
	Options:	1. A.between B.before C.since D.later 2. A.after B.by C.during D.until 3. A.means B.method C.medium D.measure 4. A.process B.company C.light D.form
	Answer:	1.A 2.D 3.C 4.B

篇章机器翻译

- 篇章级机器翻译致力于在神经网络中引入并利用上下文相关信息，使得神经网络能够捕获到相关信息并在翻译过程中保留语义现象从而使得翻译结果通顺且流畅。



2.2 信息检索和搜索引擎中的统计模型：TF-IDF

林洲汉
上海交通大学
2023年

搜索引擎

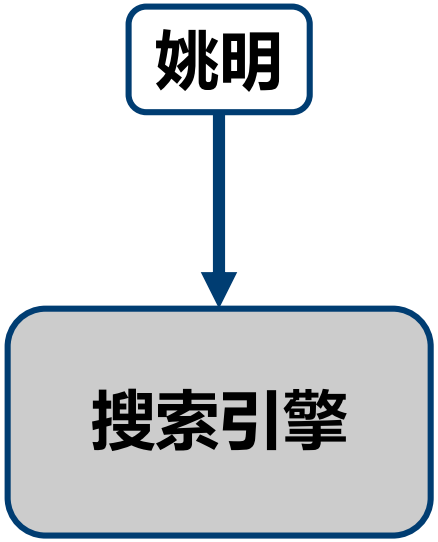

Google
姚明
全部
图片
新闻
视频
地图
更多
工具
找到约 20,100,000 条结果 （用时 0.57 秒）
https://zh.wikipedia.org › zh-hans › 姚明
姚明- 维基百科，自由的百科全书
姚明（1980年9月12日 - ），男，祖籍江蘇省苏州市吴江区震泽镇，生於上海，中国籃球運動員，曾為中國國家籃球隊隊員，曾效力于中国篮球职业联赛（CBA）上海大鯊魚籃球 ...
叶莉[编辑] · 奈史密斯篮球名人纪念堂[编辑] · 王治郅[编辑]
https://baike.baidu.com › item › 姚明
姚明（亚洲篮球联合会主席、中国篮球协会主席）_百度百科
姚明（Yao Ming），男，汉族，无党派人士，1980年9月12日出生于上海市徐汇区，祖籍江苏省苏州市吴江区震泽镇，前中国职业篮球运动员，司职中锋，现任亚洲篮球联合会 ...
生涯最高分： 41分
出生日期： 1980年9月12日
主要奖项： 8次NBA全明星（2003-2009；20...
出生地： 上海市徐汇区
早年经历 · 职业生涯 · NBA数据 · 公益活动
https://s.weibo.com › weibo › q=#姚明#
姚明 - 微博搜索- WEIBO
2007年全明星，艾弗森姚明纳什因伤缺阵没有上场，空气中弥漫的都是青春的味道，怀念那个美好懵懂的年代…… #姚明##艾弗森##NBA吐槽大会# L黑曼巴经典视频的微博视频 .
视频
15:01
2003年亚锦赛，姚明38分率队复仇韩国，直通雅典
YouTube · 篮球红绿灯
2020年12月15日
【NBA經典時刻】姚明狂砍41分絕殺對手！宰制內綫誰來都沒 ...

搜索引擎

- **文档检索**
  - 倒排
  - TF-IDF
- **重要性排序**
  - Page Rank

文档检索

- **方案1：**
  - 根据关键词，遍历所有文档，找出含有关键词的文档



文档检索

- **方案1：**
  - 根据关键词，遍历所有文档，找出含有关键词的文档

搜索引擎

网页数据库

匹配

文档检索

- **方案1：**
  - 根据关键词，遍历所有文档，找出含有关键词的文档

文档检索

- **方案1：**
  - 根据关键词，遍历所有文档，找出含有关键词的文档
  - 时间与计算复杂性极大

文档检索

- **方案1：**
  - 根据关键词，遍历所有文档，找出含有关键词的文档
  - 时间与计算复杂性极大

- **方案2：倒排**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配

- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配

搜索引擎

网页数据库

文档检索

- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配

- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配

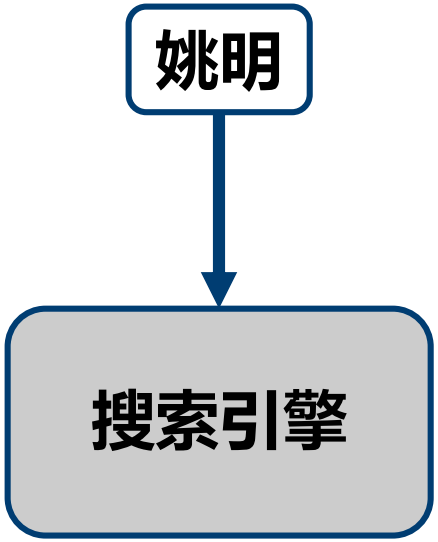
搜索引擎

网页数据库

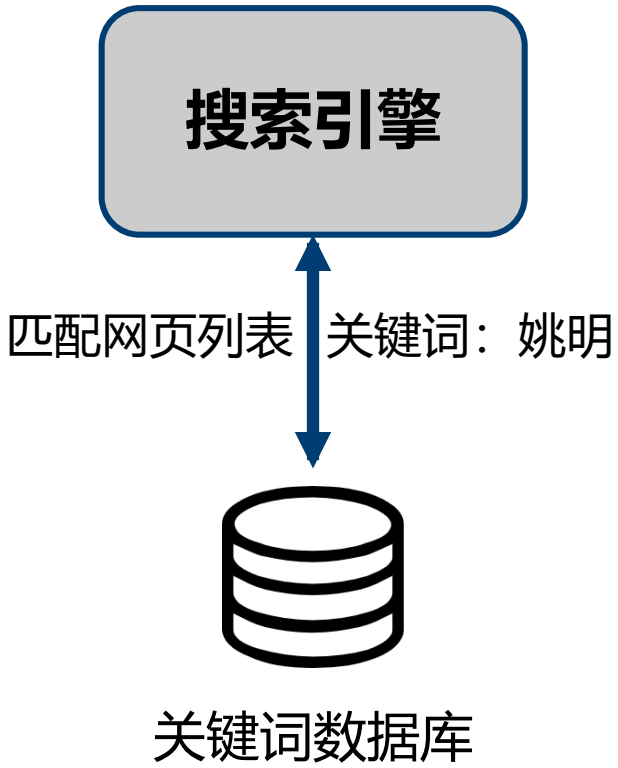
文档检索

- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配

- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配



- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配



文档检索

- **方案1：**
  - 根据关键词，遍历所有文档，找出含有关键词的文档
  - 时间与计算复杂性极大

- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配
  - 复杂性基本相当于一次查表

文档检索

- **方案1：**
  - 根据关键词，遍历所有文档，找出含有关键词的文档
  - 时间与计算复杂性极大

- **倒排：**
  - 提前遍历所有文档，储存出现的关键词，进行关键词匹配
  - 复杂性基本相当于一次查表
  - 问题：如何选取关键词？

如何选取关键词?

- **词频（Term Frequency）**
  - 词语在文档中出现的频率越大，则其关键程度越高

如何选取关键词?

- **词频（Term Frequency）**
  - 词语在文档中出现的频率越大，则其关键程度越高
  - 词语w_i在文档d_j中的词频为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中$n_{k,j}$是词语w_k在文档d_j中出现的次数

如何选取关键词?

词频

段落	前5个关键词（含并列）
姚明是中国最具影响力的人物之一，同时也是世界最知名的华人运动员之一[4][5]。2009年，姚明收购上海男篮，成为上海大鲨鱼篮球俱乐部老板[6]。2011年7月20日，姚明正式宣布退役[7][8]。2016年11月22日，姚明出任CBA联盟副董事长。2017年2月，姚明当选为中国篮球协会主席[9]。2016年4月4日，姚明与前NBA球星沙奎尔·奥尼尔和艾伦·艾弗森一同入选奈史密斯篮球名人纪念堂，他也是首位入选也是迄今为止唯一入选名人堂的亚洲球员。	姚明、年、月、是、日、入选、也、的
篮球，为一种由两队参与，在一个长方形篮球场进行的球类运动。每队出场5名队员，可将球向任何方向传、投、拍、滚或运，目的是将篮球投入对方球篮得分，并阻止对方获得控球权或得分。篮球是世界上最多人观看的球类活动之一[1]，从1936年起就是奥运会的正式竞赛项目之一[2]。在三分线内投进可得两分，若在三分线外投篮，可以得到三分。在比赛结束时，得分最多的球队获胜，但若二队平手，会进行延长赛。在球赛中，控球者可以用持续运球（将球弹到地上，再反弹到自己手中）的方式行进或奔跑，也可以将球传给其他队员，但若走步（掌控球者没有运球，且移动超过指定范围或指定方式）、翻球或二次运球均是违例。	在、的、或、可以、若、控球、篮球、球、运球、将、得分、是、三分
香蕉原产于热带的马来群岛及澳洲北部地区，最早可能是在巴布亚新几内亚驯化[3] [4]，未受人类驯化的野生蕉体型微小，难以打开果实且大部分为种子，与能剥开直接吃的食用蕉差异颇大，是人类以栽培的方式才获得黄色香蕉。至少有107个国家生产香蕉[5]。种植香蕉主要是为其果实，偶尔会用作纤维、香蕉酒、香蕉啤酒或园艺植物。2013年香蕉是产值第四大的食用植物，仅次于米、麦及玉米[6]。	香蕉、的、是、人类、驯化、及、大、蕉、果实、为

Q：通过词频选取的关键词合理吗?

如何选取关键词?

- **词频（Term Frequency）**
  - 特定词语在文件中出现的频率
  - 词语w_i在文件d_j中的词频为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中$n_{k,j}$是词语w_k在文件d_j中出现的次数

  - 缺点：
    - 一些常用词语会有极大的词频，但并没有有效“信息”
    - 例如：是、的等

Q：如何从关键词中去掉这些我们不关心的常用词?

如何选取关键词?

- **逆向文档频率（Inverse Document Frequency）**

$$idf_i = \log \frac{|D|}{|\{j: w_i \in d_j\}|}$$

  - 其中：
    - $|D|$：语料库中的文档总数
    - $|\{j: w_i \in d_j\}|$：包含词语w_i的文档数目

如何选取关键词?

- **逆向文件频率（Inverse Document Frequency）**

$$idf_i = \log \frac{|D|}{|\{j: w_i \in d_j\}|}$$

  - 其中：
    - $|D|$：语料库中的文件总数
    - $|\{j: w_i \in d_j\}|$：包含词语w_i的文件数目
  - 为每个词语重新分配权重，越通用的词语，权重越低

如何选取关键词?

逆向文件频率

段落	按词频计算的前5个关键词
姚明是中国最具影响力的人物之一，同时也是世界最知名的华人运动员之一[4][5]。2009年，姚明收购上海男篮，成为上海大鲨鱼篮球俱乐部老板[6]。2011年7月20日，姚明正式宣布退役[7][8]。2016年11月22日，姚明出任CBA联盟副董事长。2017年2月，姚明当选为中国篮球协会主席[9]。2016年4月4日，姚明与前NBA球星沙奎尔·奥尼尔和艾伦·艾弗森一同入选奈史密斯篮球名人纪念堂，他也是首位入选也是迄今为止唯一入选名人堂的亚洲球员。	姚明、年、月、是、日、入选、也、的
篮球，为一种由两队参与，在一个长方形篮球场进行的球类运动。每队出场5名队员，可将球向任何方向传、投、拍、滚或运，目的是将篮球投入对方球篮得分，并阻止对方获得控球权或得分。篮球是世界上最多人观看的球类活动之一[1]，从1936年起就是奥运会的正式竞赛项目之一[2]。在三分线内投进可得两分，若在三分线外投篮，可以得到三分。在比赛结束时，得分最多的球队获胜，但若二队平手，会进行延长赛。在球赛中，控球者可以用持续运球（将球弹到地上，再反弹到自己手中）的方式行进或奔跑，也可以将球传给其他队员，但若走步（掌控球者没有运球，且移动超过指定范围或指定方式）、翻球或二次运球均是违例。	在、的、或、可以、若、控球、篮球、球、运球、将、得分、是、三分
香蕉原产于热带的马来群岛及澳洲北部地区，最早可能是在巴布亚新几内亚驯化[3] [4]，未受人类驯化的野生蕉体型微小，难以打开果实且大部分为种子，与能剥开直接吃的食用蕉差异颇大，是人类以栽培的方式才获得黄色香蕉。至少有107个国家生产香蕉[5]。种植香蕉主要是为其果实，偶尔会用作纤维、香蕉酒、香蕉啤酒或园艺植物。2013年香蕉是产值第四大的食用植物，仅次于米、麦及玉米[6]。	香蕉、的、是、人类、驯化、及、大、蕉、果实、为

如何选取关键词?

逆向文件频率

候选词	逆向文件频率	候选词	逆向文件频率
若	0.4771	姚明	0.4771
日	0.4771	控球	0.4771
月	0.4771	可以	0.4771
将	0.4771	球	0.4771
篮球	0.1761	运球	0.4771
年	0.1761	入选	0.4771
也	0.1761	得分	0.4771
或	0.1761	三分	0.4771
在	0.1761	香蕉	0.4771
的	0	人类	0.4771
是	0	驯化	0.4771

Term Frequency – Inverse Document Frequency

$$tf\text{-}idf_{i,j} \overset{\text{def}}{=} tf_{i,j} \times idf_i$$

段落	前5个关键词（含并列）
姚明是中国最具影响力的人物之一，同时也是世界最知名的华人运动员之一[4][5]。2009年，姚明收购上海男篮，成为上海大鲨鱼篮球俱乐部老板[6]。2011年7月20日，姚明正式宣布退役[7][8]。2016年11月22日，姚明出任CBA联盟副董事长。2017年2月，姚明当选为中国篮球协会主席[9]。2016年4月4日，姚明与前NBA球星沙奎尔·奥尼尔和艾伦·艾弗森一同入选奈史密斯篮球名人纪念堂，他也是首位入选也是迄今为止唯一入选名人堂的亚洲球员。	姚明、月、日、入选、上海
篮球，为一种由两队参与，在一个长方形篮球场进行的球类运动。每队出场5名队员，可将球向任何方向传、投、拍、滚或运，目的是将篮球投入对方球篮得分，并阻止对方获得控球权或得分。篮球是世界上最多人观看的球类活动之一[1]，从1936年起就是奥运会的正式竞赛项目之一[2]。在三分线内投进可得两分，若在三分线外投篮，可以得到三分。在比赛结束时，得分最多的球队获胜，但若二队平手，会进行延长赛。在球赛中，控球者可以用持续运球（将球弹到地上，再反弹到自己手中）的方式行进或奔跑，也可以将球传给其他队员，但若走步（掌控球者没有运球，且移动超过指定范围或指定方式）、翻球或二次运球均是违例。	可以、若、控球、球、运球、将、得分、三分
香蕉原产于热带的马来群岛及澳洲北部地区，最早可能是在巴布亚新几内亚驯化[3] [4]，未受人类驯化的野生蕉体型微小，难以打开果实且大部分为种子，与能剥开直接吃的食用蕉差异颇大，是人类以栽培的方式才获得黄色香蕉。至少有107个国家生产香蕉[5]。种植香蕉主要是为其果实，偶尔会用作纤维、香蕉酒、香蕉啤酒或园艺植物。2013年香蕉是产值第四大的食用植物，仅次于米、麦及玉米[6]。	香蕉、人类、驯化、及、蕉、果实

Term Frequency – Inverse Document Frequency

$$tf\text{-}idf_{i,j} \stackrel{\text{def}}{=} tf_{i,j} \times idf_i$$

- **应用：**
  - 自动提取关键词
  - 评价文章相似性
  - 自动摘要生成

- **局限**
  - TF：没有考虑位置信息与语义信息
  - IDF：简单的假设文档频率越大的单词越无用

- **谢谢大家！**