南京大学
NANJING UNIVERSITY

自然语言处理

语言模型和词向量

吴震

南京大学人工智能学院
南京大学自然语言处理研究组

2023年3月

目录

南京大学
NANJING UNIVERSITY

01 语言模型

LANGUAGE MODEL

自然语言的合理性

智能拼音输入法

句子1： *我在学习自然语言处理* 语法通顺，表意清晰

句子2： *自然语言处理在学习我* 语法通顺，语义不常见

句子3： *学习在自然语言处理我* 不知所云

自然语言是一种上下文相关的信息表达和信息传递的方式

能不能用数学的方式来描述语言规律？

语言模型(LANGUAGE MODEL)

- 语言模型是衡量一句话出现在自然语言中的概率的模型

P(“*我在学习自然语言处理*”)=0.000001

P(“*自然语言处理在学习我*”)=1e-22

P(“*学习在自然语言处理我*”)=1e-40

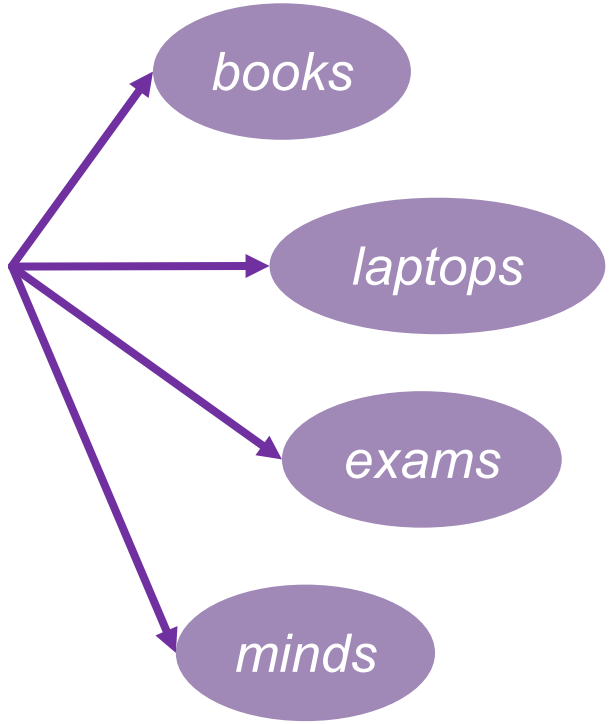
- 数学形式上，给定一句话$s = \{w_1, \dots, w_n\}$，它对应的概率为：

$$
\begin{aligned}
P(s) &= P(w_1, \dots, w_n) \\
&= P(w_1) \times P(w_2|w_1) \times \cdots \times P(w_n|w_1, \dots, w_{n-1}) \\
&= \prod_{i=1}^{n} P(w_i|w_1, \dots, w_{i-1})
\end{aligned}
$$

语言模型

- 语言模型的核心在于根据前文预测下一个词出现的概率

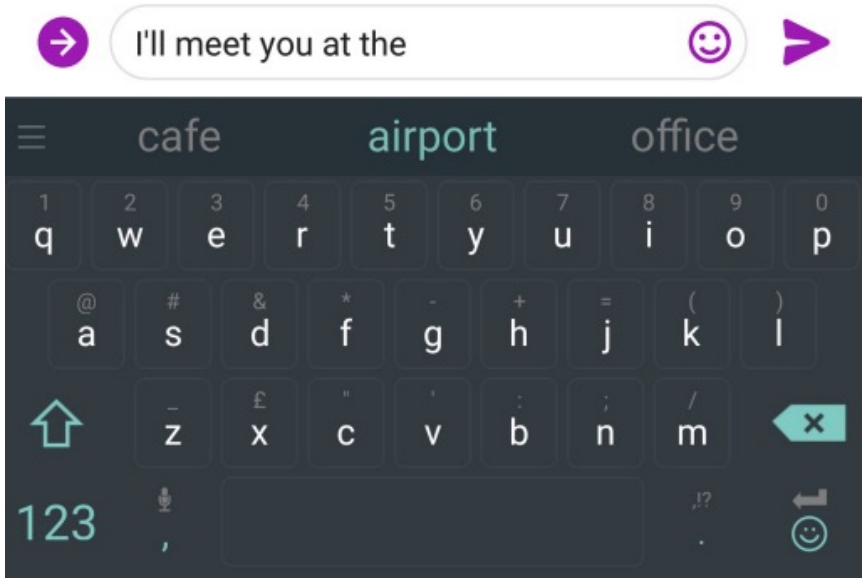
The students opened their ____

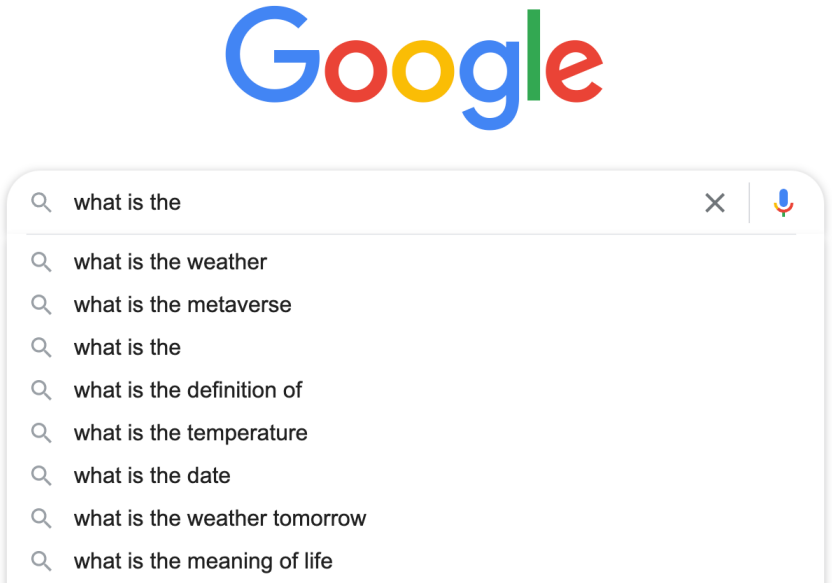


核心 $P(w_i|w_1,\dots,w_{i-1})$

$w_i \in V,\ V=\{w_1,\dots,w_{|V|}\}$

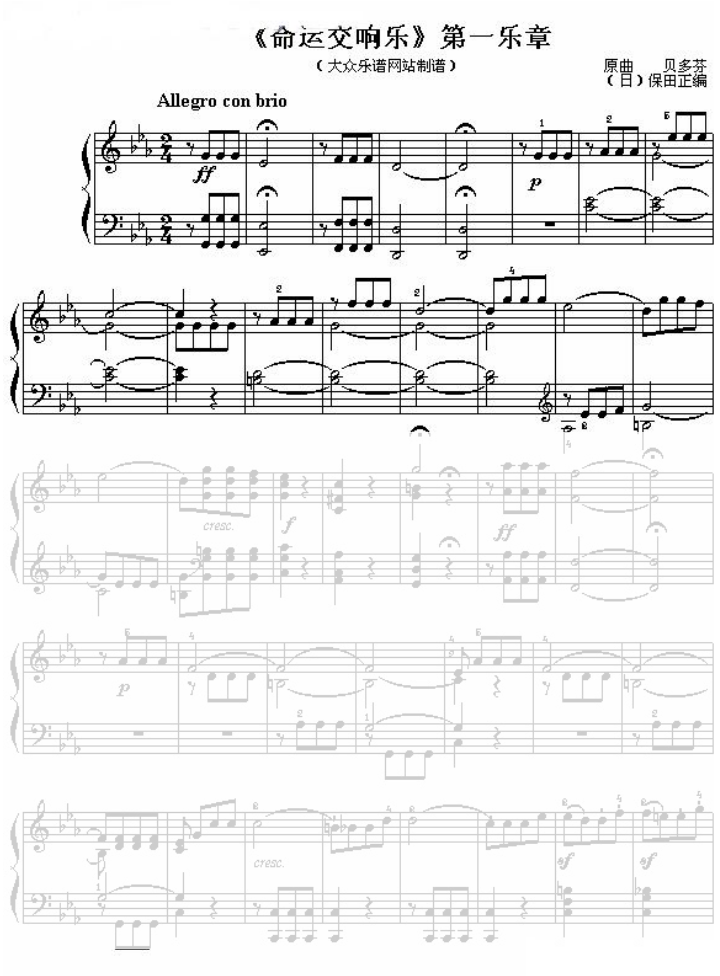
语言模型的应用场景


I'll meet you at the
cafe
airport
office

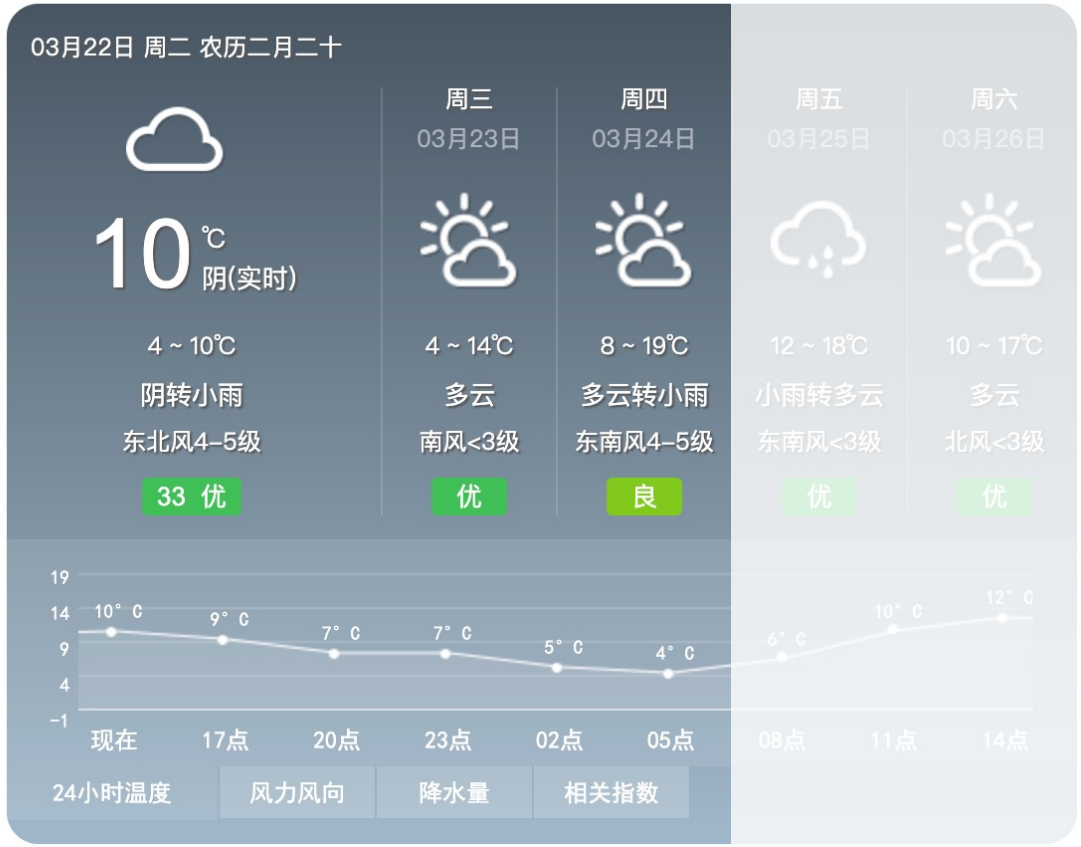

Google
what is the
what is the weather
what is the metaverse
what is the
what is the definition of
what is the temperature
what is the date
what is the weather tomorrow
what is the meaning of life


京东
这是
这是苹果吗也许是吧
这是理所当然的吗
这是我的
这是谁的家
这是什么声音
这是规定
秒杀
电器
/运营商/数码

另类“语言模型”



江苏南京天气预报_一周天气预报



https://openai.com/blog/musenet/

语言模型

- 如何计算语言模型的概率？

$$P(s) = P(w_1) \times P(w_2|w_1) \times P(w_2|w_1, w_2) \ldots \times P(w_n|w_1, \ldots, w_{n-1})$$

易算　　不麻烦　　还可以　　无法估算

马尔可夫假设(MARKOV ASSUMPTION)

- 马尔可夫链
  - 描述从一个状态到转换另一个状态的随机过程。该过程具备“无记忆”的性质，即当前时刻状态的概率分布只能由上一时刻的状态决定，和更久之前的状态无关。

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-1})$$

- K阶马尔可夫链
  - 当前时刻状态的概率分布只和前面k个时刻的状态相关

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$

统计语言模型

- 马尔可夫假设：当前词出现的概率只和它前面的k个词相关

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$ 马尔可夫假设

$$= P(w_i)$$ k=0, Unigram Model

$$= P(w_i|w_{i-1})$$ k=1, Bigram Model

$$= P(w_i|w_{i-2},w_{i-1})$$ k=2, Trigram Model

统计语言模型

- 用频率估计概率（大数定理）

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$ 马尔可夫假设

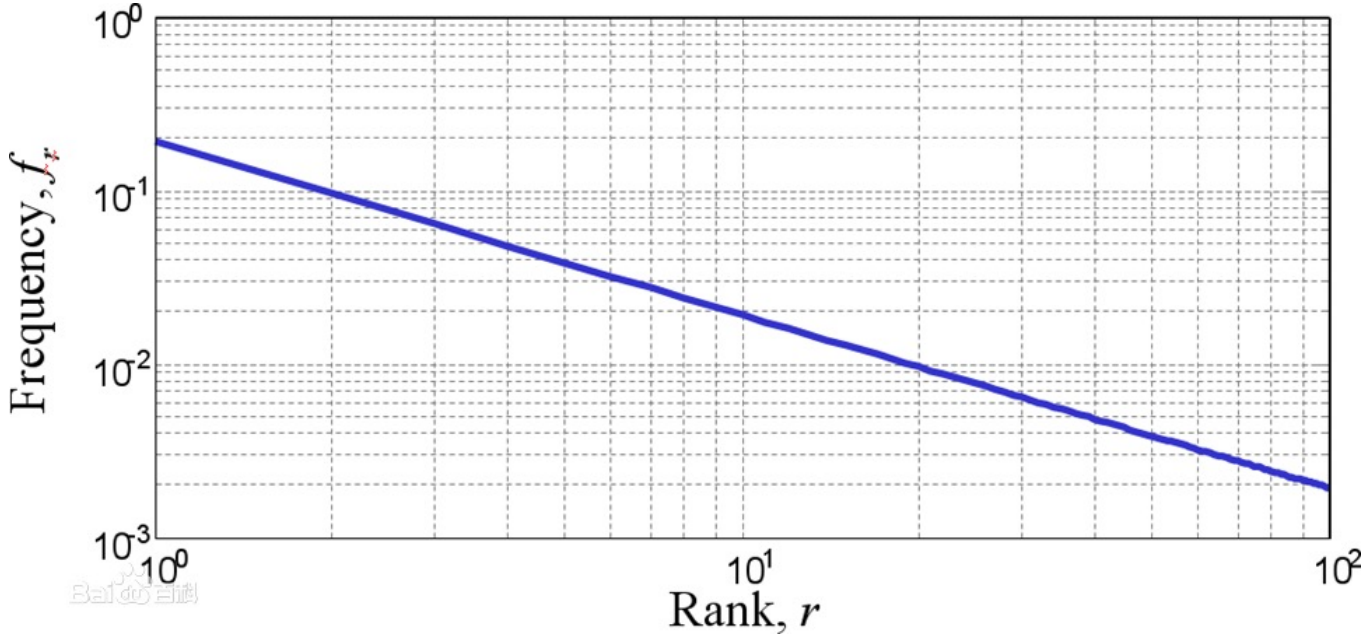
$$= \frac{P(w_{i-k},\dots,w_{i-1},w_i)}{P(w_{i-k},\dots,w_{i-1})}$$ 条件概率

$$\approx \frac{count(w_{i-k},\dots,w_{i-1},w_i)\Big/count(all\ grams)}{count(w_{i-k},\dots,w_{i-1})\Big/count(all\ grams)}$$

$$= \frac{count(w_{i-k},\dots,w_{i-1},w_i)}{count(w_{i-k},\dots,w_{i-1})}$$ 概率估计

统计语言模型

- Zipf Law
  - 如果以词频排序，词频和排位成反比（乘积接近一个常数）。



Zipf's law in *Tom Sawyer*

Word	Freq. (f)	Rank (r)	$f \cdot r$
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
Could	2	4000	8000
Applausive	1	8000	8000

ZIPF LAW

南京大学 NANJING UNIVERSITY

- 隐藏信息
  - 排位靠后的词的频率非常低，甚至未出现在语料中。

- 数据稀疏问题
  - 对于未出现在语料中的词或n-gram，无法估计其概率；
  - Balh等人用150万词的训练语料训练trigram模型，测试语料中有23%的trigram没有在训练语料中出现过。

统计语言模型

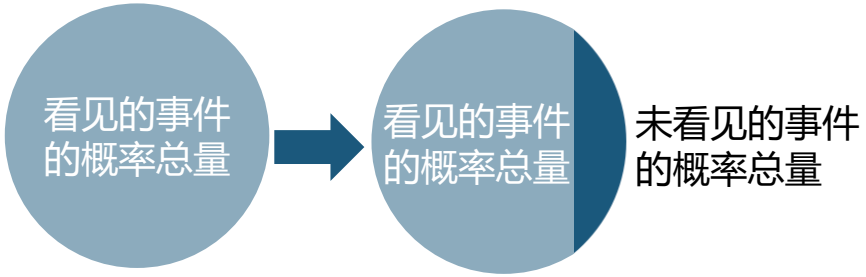
- 数据稀疏问题

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$

$$= \frac{P(w_{i-k},\dots,w_{i-1},w_i)}{P(w_{i-k},\dots,w_{i-1})}$$

未出现在语料中

$$\approx \frac{count(w_{i-k},\dots,w_{i-1},w_i)}{count(w_{i-k},\dots,w_{i-1})} \rightarrow 0$$



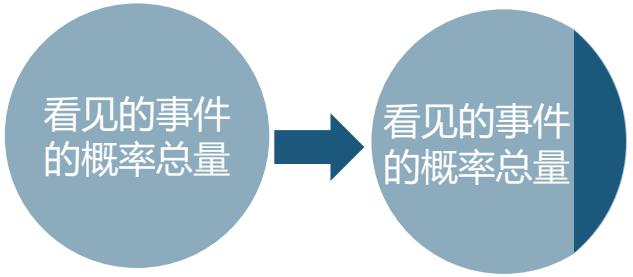
平滑技术

统计语言模型

- 数据稀疏问题

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$

$$= \frac{P(w_{i-k},\dots,w_{i-1},w_i)}{P(w_{i-k},\dots,w_{i-1})}$$



未看见的事件
的概率总量

$$\approx \frac{count(w_{i-k},\dots,w_{i-1},w_i)+1}{count(w_{i-k},\dots,w_{i-1})+|V|}$$

拉普拉斯平滑

平滑技术

古德-图灵平滑
插值平滑
Katz平滑
……

统计语言模型

- 数据稀疏问题

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$

$$= \frac{P(w_{i-k},\dots,w_{i-1},w_i)}{P(w_{i-k},\dots,w_{i-1})}$$

$$\approx \frac{count(w_{i-k},\dots,w_{i-1},w_i)}{count(w_{i-k},\dots,w_{i-1})} \longrightarrow 0$$

未出现在语料中

统计语言模型

- 数据稀疏问题

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$

$$= \frac{P(w_{i-k},\dots,w_{i-1},w_i)}{P(w_{i-k},\dots,w_{i-1})}$$

$$\approx \frac{count(w_{i-k},\dots,w_{i-1},w_i)}{count(w_{i-k},\dots,w_{i-1})}$$

→ students opened their

$$\approx \frac{count(w_{i-k},\dots,w_{i-1},w_i)}{count(w_{i-k+j},\dots,w_{i-1})}$$

回退策略

→ opened their

统计语言模型

- 参数规模问题

$$P(w_i|w_1,\dots,w_{i-1}) = P(w_i|w_{i-k},\dots,w_{i-1})$$

$$= \frac{P(w_{i-k},\dots,w_{i-1},w_i)}{P(w_{i-k},\dots,w_{i-1})}$$

$$\approx \frac{count(w_{i-k},\dots,w_{i-1},w_i)}{count(w_{i-k},\dots,w_{i-1})}$$

如何避免统计语言模型中的问题？

神经语言模型

词的表示

随着k的增大，参数数目呈指数增长，无法存储

k=1　　参数量：$|V|^2$

k=2　　参数量：$|V|^3$

……

k=n-1　参数量：$|V|^n$

语言模型评价方法

- 困惑度 (Perplexity)
  - 用来度量一个概率分布或概率模型预测样本的好坏程度；
  - 可以用来比较两个概率模型，低困惑度的概率模型能更好地预测样本。

$$Perplexity\,(s) = 2^{H(s)}$$

$$= 2^{-\frac{1}{n}\log_2 P(w_1,\ldots,w_n)}$$

在测试数据上

$$= 2^{\log_2 P(w_1,\ldots,w_n)^{-\frac{1}{n}}}$$

$$= P(w_1,\ldots,w_n)^{-\frac{1}{n}}$$

$$= \sqrt[n]{1/P(w_1,\ldots,w_n)}$$

南京大学
NANJING UNIVERSITY

02

词的表示

WORD REPRESENTATIONS

词库

- WordNet：一个包含同义词（ synonym ）和上位词（ hypernyms ）的知识库

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: good Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **good** (benefit) *"for your own good"; "what's the good of worrying?"*
- S: (n) **good**, goodness (moral excellence or admirableness) *"there is much good to be found in people"*
- S: (n) **good**, goodness (that which is pleasing or valuable or useful) *"weigh the good against the bad"; "among the highest goods of all are happiness and self-realization"*
- S: (n) commodity, trade good, **good** (articles of commerce)

Adjective

- S: (adj) **good** (having desirable or positive qualities especially those suitable for a thing specified) *"good news from the hospital"; "a good report card"; "when she was good she was very very good"; "a good knife is one good for*

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: panda Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: <lexical filename > (gloss)

Noun

- <noun.animal>S: (n) giant panda, **panda**, panda bear, coon bear, Ailuropoda melanoleuca (large black-and-white herbivorous mammal of bamboo forests of China and Tibet; in some classifications considered a member of the bear family or of a separate family Ailuropodidae)
- <noun.animal>S: (n) lesser panda, red panda, **panda**, bear cat, cat bear, Ailurus fulgens (reddish-brown Old World raccoon-like carnivore; in some classifications considered unrelated to the giant pandas)

词库的问题

- 缺少差异性
  - “proficient”也被视为“good”的同义词
  - 不够精确
- 缺少新词
  - wicked, badass, nifty, wizard, genius, ninja, bombes
  - 无法实时更新
- 主观性
- 人工标注

离散词表示

- One-hot表示：当前词维度为1，其它词的维度为0

单词表示

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 ... 0 0]

hotel = [0 0 0 0 0 0 0 0 1 0 0 0 0 ... 0 0]

词表维度：$|V|$

文本表示

The students opened their books

[0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 ... 0 0]

次数、频率、逆文档频率、TF-IDF，……

词袋模型（BOW，Bag of Words）

离散词表示

- 词频 (Term Frequency, TF)
  - 在文档中出现频率越高的词对当前文档可能越重要

$$f_{ij} = \frac{count(\text{term } i) \text{ in doc } j}{count(\text{all term}) \text{ in doc } j}$$

$$tf_{ij} = \frac{f_{ij}}{max_k(f_{kj})}$$ 归一化

离散词表示

- 逆文档频率 (Inverse Document Frequency, IDF)
  - 在很多文档中都出现的词可能不重要（如虚词）

$$df_i = \text{doc frequency of term } i$$

$$= \text{numbers of doc containing term } i$$

$$idf_i = log_2 \frac{N}{df_i}$$

N为文档总数目

离散词表示

- TF-IDF
  - 综合一个词在当前文档中的频率和所有文档中出现的次数来度量这个词对当前文档的重要性

$$tf_{ij} - idf_i = tf_{ij} * idf_i$$

$$= tf_{ij} * log_2 \frac{N}{df_i}$$

离散词表示的问题

- 语义鸿沟

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 … 0 0]

hotel = [0 0 0 0 0 0 0 0 1 0 0 0 0 … 0 0]

- 维度爆炸

$|V|$ ⟶ 500,000+ 存储、计算开销极大

分布式词表示

- 词向量（word vectors, word embeddings）：用一个低维稠密的向量表示单词的整体含义

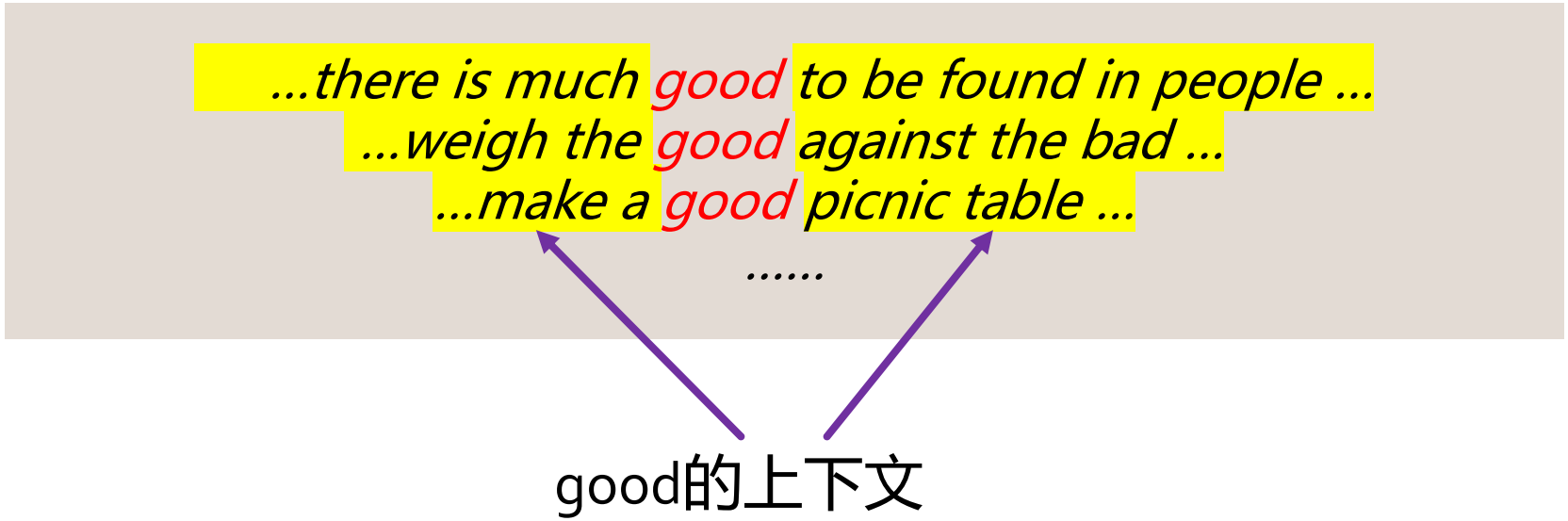
$$\text{motel} = \begin{bmatrix} \cdots \\ 0.286 \\ 0.792 \\ -0.177 \\ \cdots \end{bmatrix} \qquad \text{hotel} = \begin{bmatrix} \cdots \\ 0.271 \\ 0.695 \\ -0.172 \\ \cdots \end{bmatrix}$$

分布式词表示的特性

- 每一维不表示具体的含义
- 低维
  - 节省存储空间
  - 节省计算时间
- 可以度量词之词之间的语义相似性
  - “motel” vs “hotel”
  - “King” – “Queen” = “Man” – “Woman”？

分布式词表示

- 分布式词表示核心思想：一个词的含义能被这个词所在的上下文反映
  - “You shall know a word by the company it keeps” (J. R. Firth 1957: 11)
  - 现代自然语言处理研究的奠基思想



03

Word2Vec

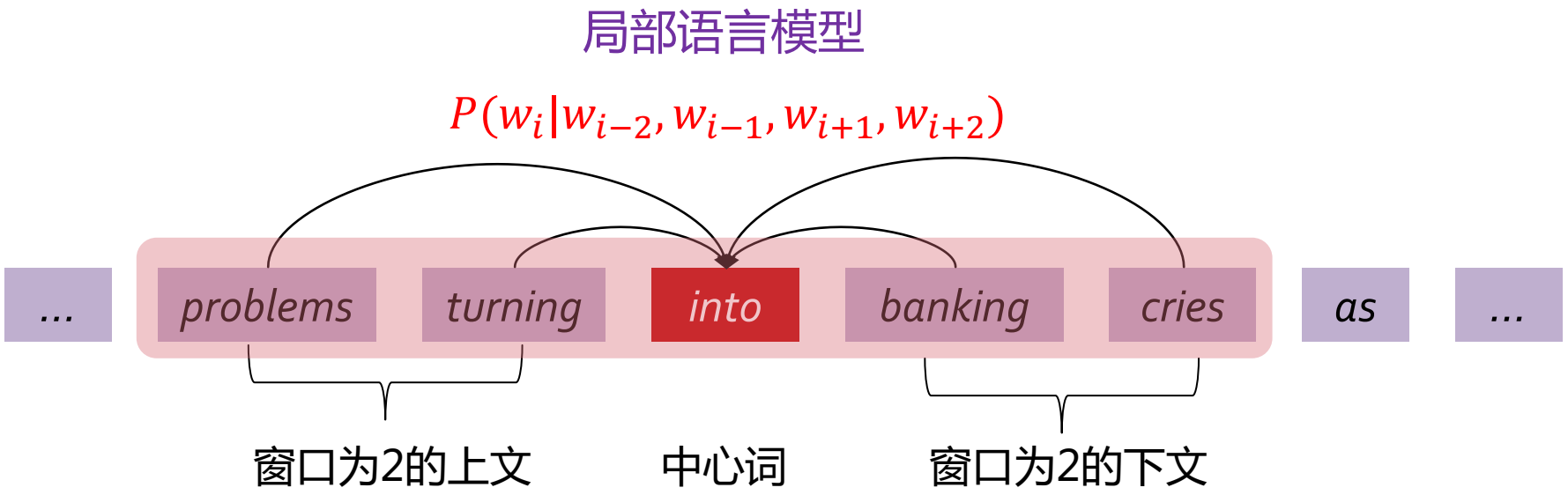
WORD2VEC

介绍

- Word2vec[Mikolov et al. 2013] 是一套学习词向量的算法框架
- 算法思想
  - 大量的自然语言文本（训练语料）
  - 为词表中的每个词随机初始化一个向量表示
  - 遍历文本中的每个单词c，其上下文单词为o
  - 使用单词c的上下文o预测单词c的概率分布（核心思想）
  - 更新词向量的表示使得单词c的预测概率最大化

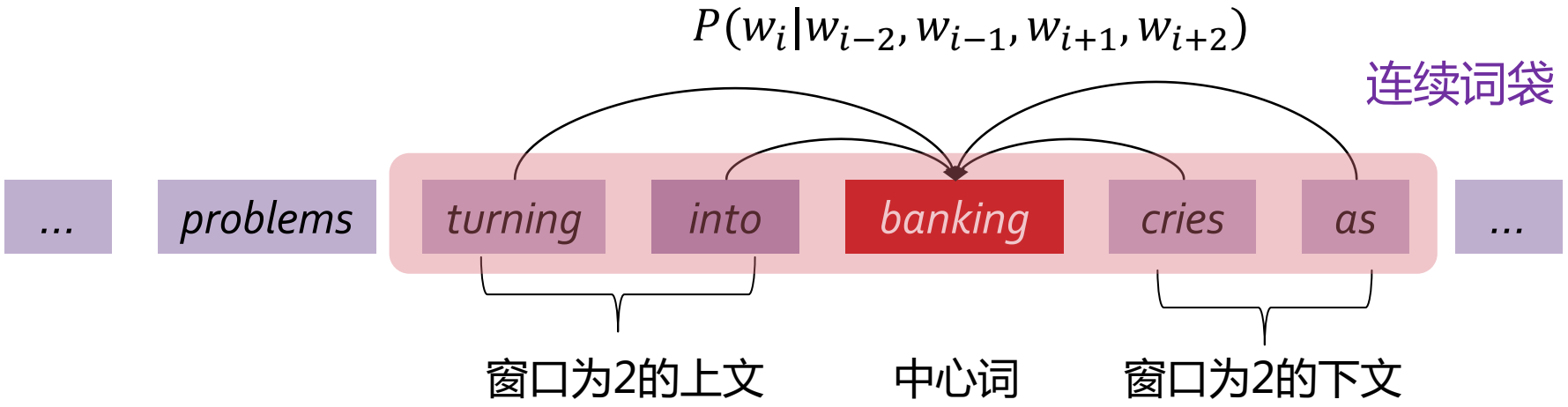
介绍

- 连续词袋模型（CBOW，Continuous Bag of Words）
  - 目的：获得词向量
  - 优化目标：局部语言模型



介绍

- 连续词袋模型（CBOW，Continuous Bag of Words）
  - 目的：获得词向量
  - 优化目标：局部语言模型



连续词袋模型

- 似然函数
  - 对于每个单词$w_i(i=1,\dots,n)$，给定其窗口为k的上下文$w_o=\{w_{i-k},\cdots w_{i-1},w_{i+1},\cdots w_{i+k}\}$，其数据似然为：

$$L(\theta)=\prod_{i=1}^{n}P(w_i|w_o;\theta)$$

模型参数

- 损失函数
  - 负对数似然的平均

$$J(\theta)=-\frac{1}{n}L(\theta)=-\frac{1}{n}\sum_{i=1}^{n}\log P(w_i|w_o;\theta)$$

优化目标

- 最小化损失函数

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log P(w_i|w_o;\theta)$$

- 如何计算$P(w_i|w_o;\theta)$?
  - 单词w_i的词向量v_i
  - 上下文w_o的词向量$v_o = \frac{v_{i-k}+\cdots+v_{i-1}+v_{i+1}+\cdots+v_{i+k}}{2k}$

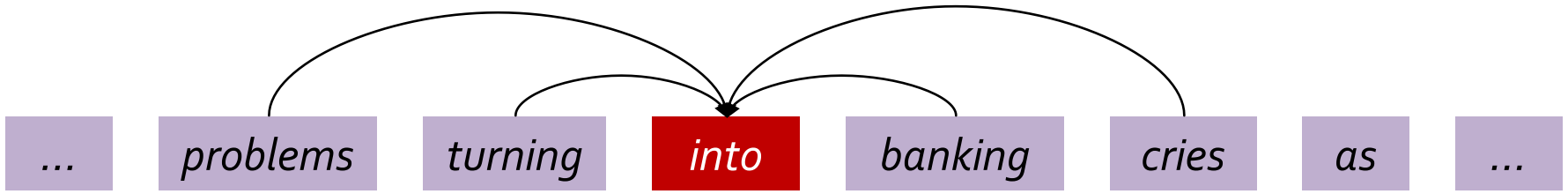
分布式词表示核心思想：单词的含义由其上下文反映 → 度量单词w_i和上下文w_o的相关性 → $v_o^T v_i$ 值越大，越相关

优化目标

- 如何计算$P(w_i|w_o;\theta)$?

$$P(w_i|w_o;\theta)=\frac{\exp(v_o^T v_i)}{\sum_{j=1}^{|V|}\exp(v_o^T v_j)}$$

归一化



in,
to,
out,
...

- Softmax函数
  - 将实数值转换成(0, 1)的概率值

模型优化

- 两套词向量（训练更稳定）
  - 单词w_i的词向量v_i
  - 上下文w_o的词向量$u_o = \frac{u_{i-k}+\cdots+u_{i-1}+u_{i+1}+\cdots+u_{i+k}}{2k}$

$$J(\theta) = -\log P(w_i|w_o;\theta)$$

$$= -\log \frac{\exp(u_o^T v_i)}{\sum_{j=1}^{|V|} \exp(u_o^T v_j)}$$

$$= -u_o^T v_i + \log \sum_{j=1}^{|V|} \exp(u_o^T v_j)$$

模型优化

- 参数梯度

$$\nabla_{u_o} J(\theta) = \nabla_{u_o}(-u_o^T v_i + \log \sum_{j=1}^{|V|} \exp(u_o^T v_j))$$

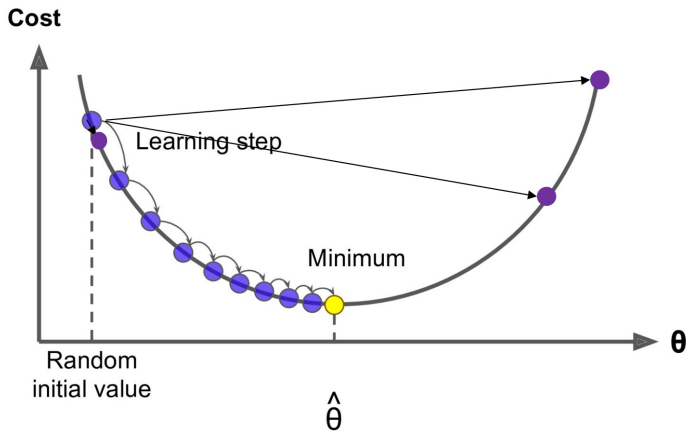
$$= \sum_{t=1}^{|V|} \frac{\exp(u_o^T v_t)}{\sum_{j=1}^{|V|} \exp(u_o^T v_j)} v_t - v_i$$

$$= \sum_{t=1}^{|V|} P(w_t|w_o;\theta)\, v_t - v_i$$

期望值–真实值

- 参数更新
  - 梯度下降（Gradient Descent）

$$\theta^{new} = \theta^{old} - \alpha \nabla_\theta J(\theta)$$



模型优化

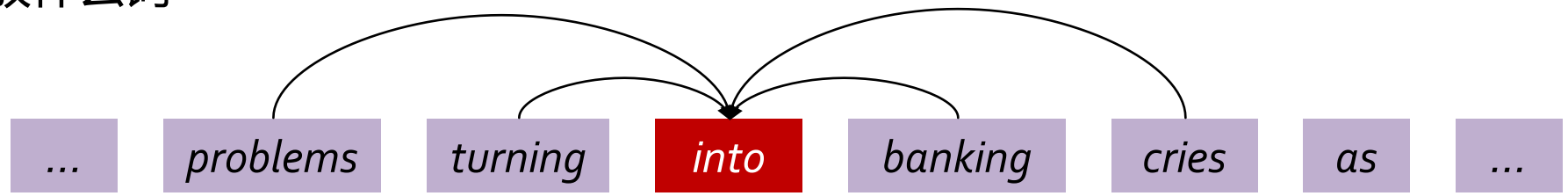
- Softmax的困境
  - 计算耗时

$$P(w_i|w_o;\theta) = \frac{\exp(v_o^T v_i)}{\sum_{j=1}^{|V|} \exp(v_o^T v_j)}$$

遍历词表中所有的词

模型优化

- 负采样（ Negative Sampling ）
  - 不再预测词表中的所有样本
  - 随机采样一些噪音词作为负样本，通过正样本和负样本之间的对比学习让模型知道当前位置应该是放什么词



in,
to,
out,
...

正样本：（ problems, turning, into, banking, cries ）
负样本：（ problems, turning, in, banking, cries ）
（ problems, turning, to, banking, cries ）
（ problems, turning, out, banking, cries ）
.....

模型优化

- 负采样

$$J(\theta) = -\log \frac{\exp(u_o^T v_i)}{\sum_{j=1}^{|V|} \exp(u_o^T v_j)}$$

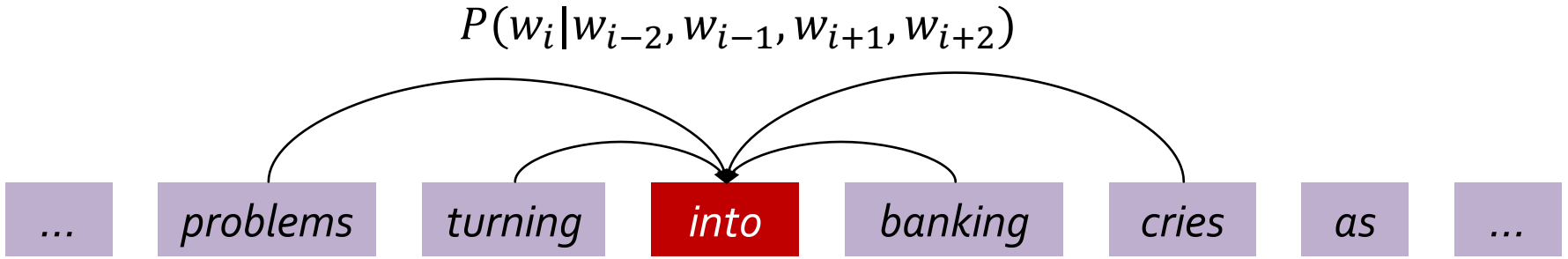
$$J(\theta) = -\log \sigma(u_o^T v_i) - \sum_{m \in \{M \text{ sampled indices}\}} \log \sigma(-u_o^T v_m)$$

正样本最大化

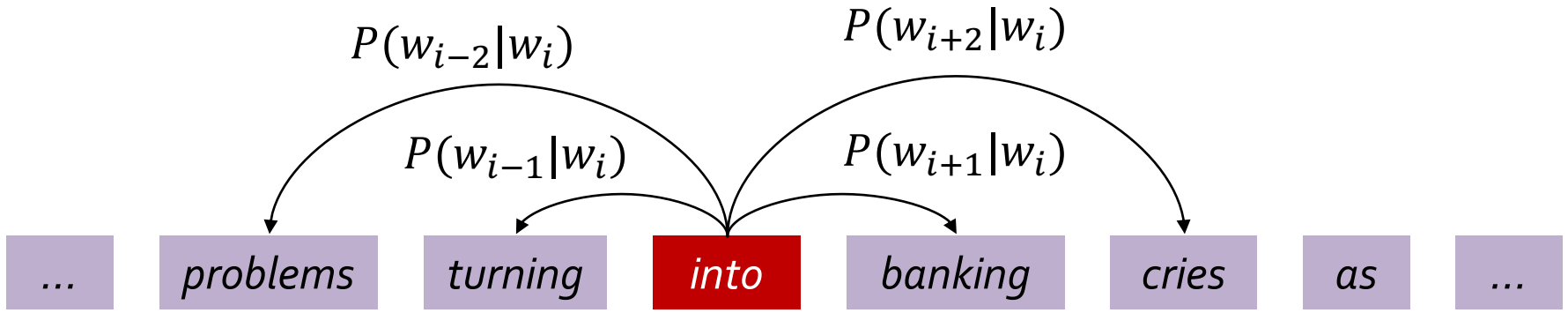
负样本最小化

WORD2VEC

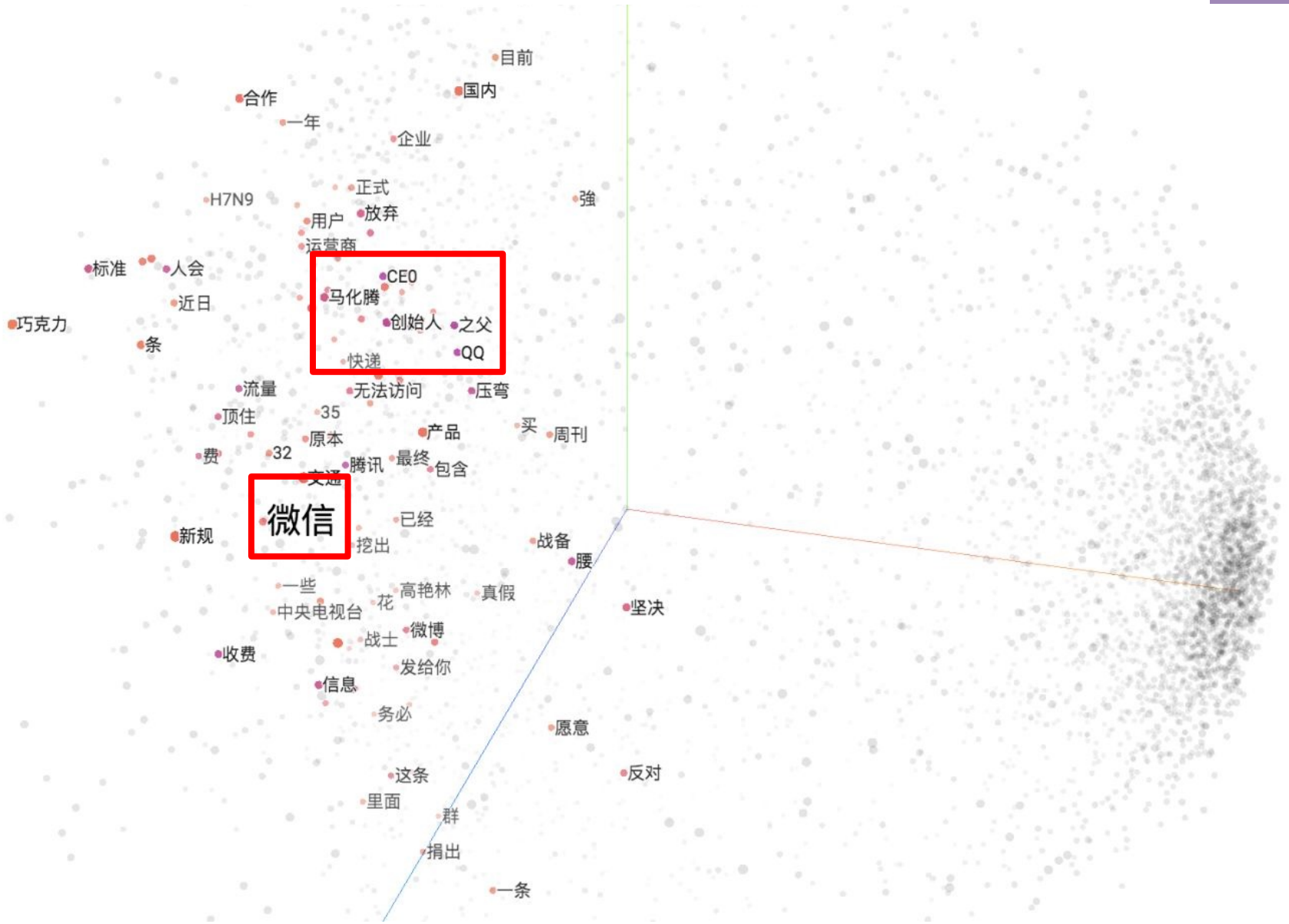
- 连续词典模型（CBOW）
  - 利用单词c的上下文o预测单词c



- Skip-gram
  - 利用单词c预测单词c的上下文o



WORD2VEC可视化



04

GloVe

GLOVE

共现矩阵

- 基于窗口的共现矩阵
  - 统计窗口内单词之间的共现信息
  - 类似于word2vec
  - 能够捕获一些句法和语义信息（局部信息）
- 基于文档的共现矩阵
  - 统计文档和单词之间的共现信息
  - LSA
  - 能够捕获话题信息（全局信息）

基于窗口的共现矩阵

- 样例
  - 窗口为2
  - 对称阵
  - 文本语料
    - I like deep learning
    - I like NLP
    - I enjoy flying

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

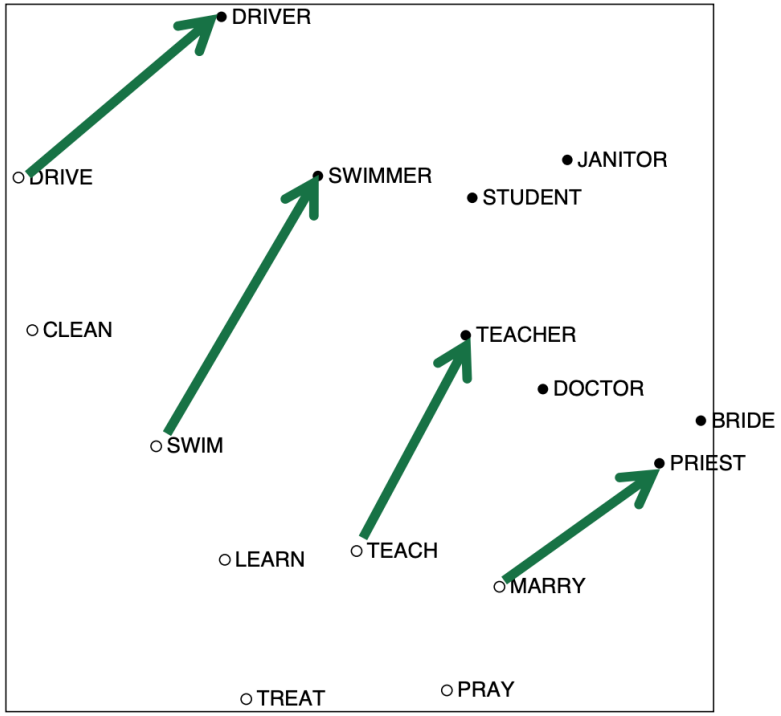
词向量

共现矩阵词向量

- 简单的计数矩阵
  - 维度爆炸
  - 存储困难
  - 稀疏性
- 低维向量
  - 选择矩阵中最有效的信息进行存储
  - SVD分解

共现矩阵词向量

- SVD分解性能不佳
  - 功能词（he, has）、停用词（a, the）出现次数多
  - SVD分解得到的信息很多和功能词、停用词相关
- 修复方法
  - 去除功能词、停用词
  - 将频数取对数
  - …



共现矩阵法 VS 直接学习法

- 共现矩阵
  - 代表方法： LSA, HAL (Lund & Burgess)
  - 优点：速度快、有效利用统计数据
  - 缺点：过分依赖单词共现性和数据量
- 直接学习
  - 代表方法： Skip-gram/CBOW (Mikolov et al)
  - 优点：能捕获语法和语义信息
  - 缺点：速度和数据规模相关、未有效利用统计数据

GLOVE

- 集两家之长
  - 共现概率矩阵X_{ij}
  - 单词w_i, w_j的词向量v_i, v_j
  - 以学习的方式，用词向量之间的语义关系来拟合共现概率矩阵

$$J = \sum_{i,j=1}^{|V|} f(X_{ij})(v_i^T v_j + b_i + b_j - \log X_{ij})^2$$

局部信息

全局统计信息

训练快
适应于大规模数据
在小规模数据上性能优秀

GLOVE可视化

- 最近邻

0. *frog*
1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

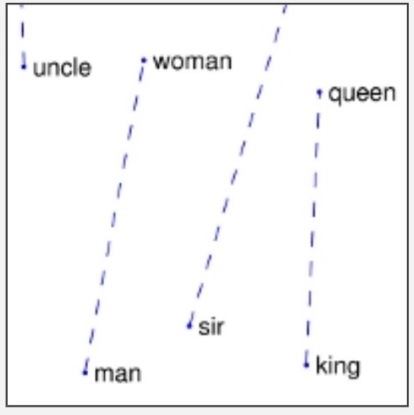

3. litoria


4. leptodactylidae

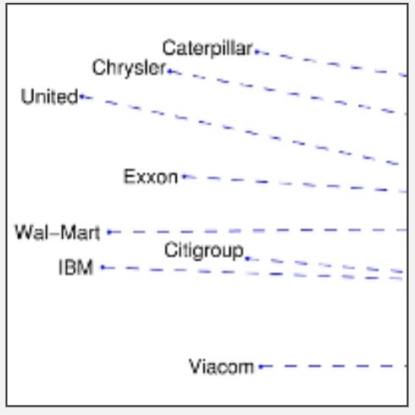

5. rana


7. eleutherodactylus

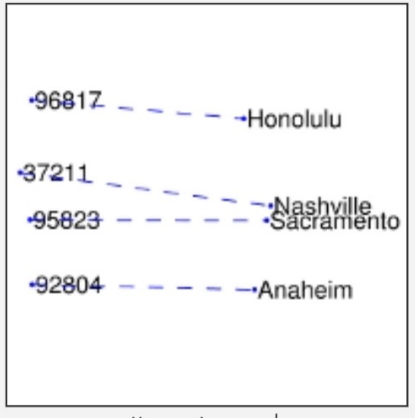
- 线性子结构



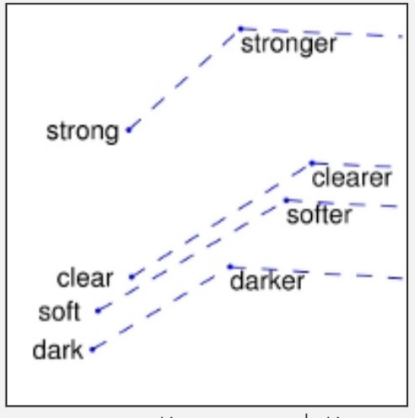
man - woman



company - ceo



city - zip code



comparative - superlative

REFERENCE

- Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3.
- Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26.
- Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

南京大学
NANJING UNIVERSITY

Q&A