西安交通大学
XI'AN JIAOTONG UNIVERSITY

自然语言理解与机器翻译

第三章：语言模型 – 向量空间模型

刘均（liukeen@xjtu.edu.cn）

陕西省天地网技术重点实验室
智能网络与网络安全教育部重点实验室

1 3.1.1 语言模型概述

2 3.1.2 BoW模型、属性权重

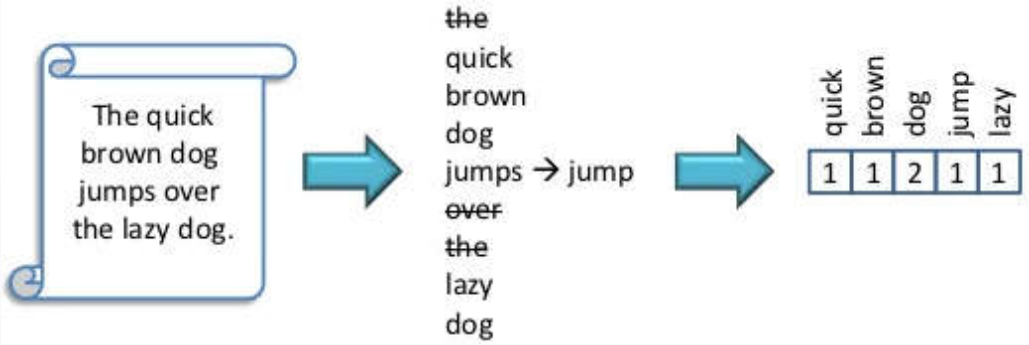
3 3.1.3 TF-IDF

4 3.1.4 NLU实例：情感分析、文本聚类

基本要求：①掌握BoW模型与TF-IDF；②了解情感分析、文本聚类两项NLU任务与典型算法。

3.1.1 语言模型

- 各类NLU任务大量使用了不同的机器学习、数据挖掘算法，这些算法很难直接处理自然语言中的原始文本
  - 自然语言特点：非规范、歧义、动态演化
  - 需求：结构化、向量长度固定、数值型、可计算
- 语言模型：自然语言在不同语言单位上的数学模型，旨在实现自然语言的可计算性。
- 词袋模型：用文档中出现的词汇来表示文档
  - 两个问题：词汇表、重要性度量
  - 优点：简单有效；缺点：忽略词序、上下文、稀疏



3.1.1 语言模型

- **概率语言模型：根据给定词汇序列来预测下一个词汇的概念模型**

$$S = w_1, w_2, \ldots, w_k$$

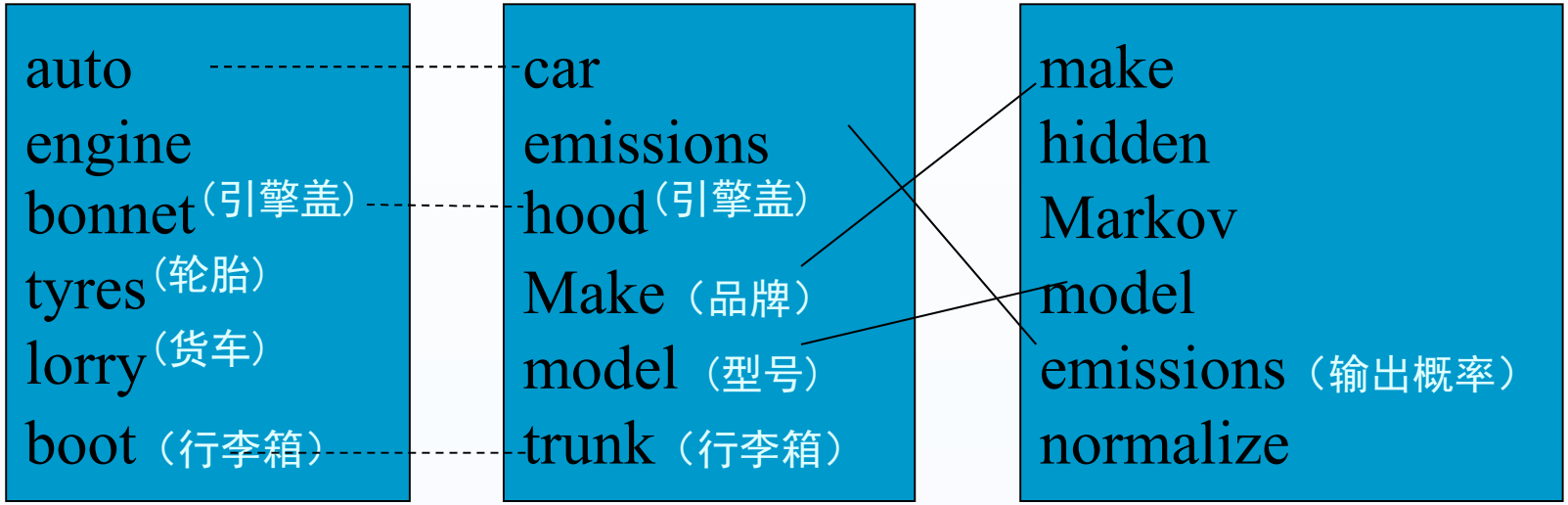
$$P(S) = p(w_1)p(w_2|w_1)p(w_3|w_1w_2)\ldots p(w_k|w_1w_2\ldots w_{k-1})$$

简化：$$P_n(S) = \prod_{i=1}^{k} P(w_i \mid w_{i-1}, w_{i-2}, w_{i-3}, \ldots, w_{i-n+1})$$

- **N元文法：$n = 1$：unigram; $n = 2$：bigram; $n = 3$：trigram**
- **n>4的情况很少， 随着n增加，复杂度增高，数据稀疏性问题严重，需要更大的语料库**
- **词序影响不大的NLU应用，如IR，取$n = 1$**
- **词序影响较大大的NLU应用，如MT，取$n = 3$，4**

3.1.1 语言模型

- Synonymy (一义多词)：poor recall
- Polysemy (一词多义)：poor precision



Synonymy

Will have small cosine but are related!

Polysemy

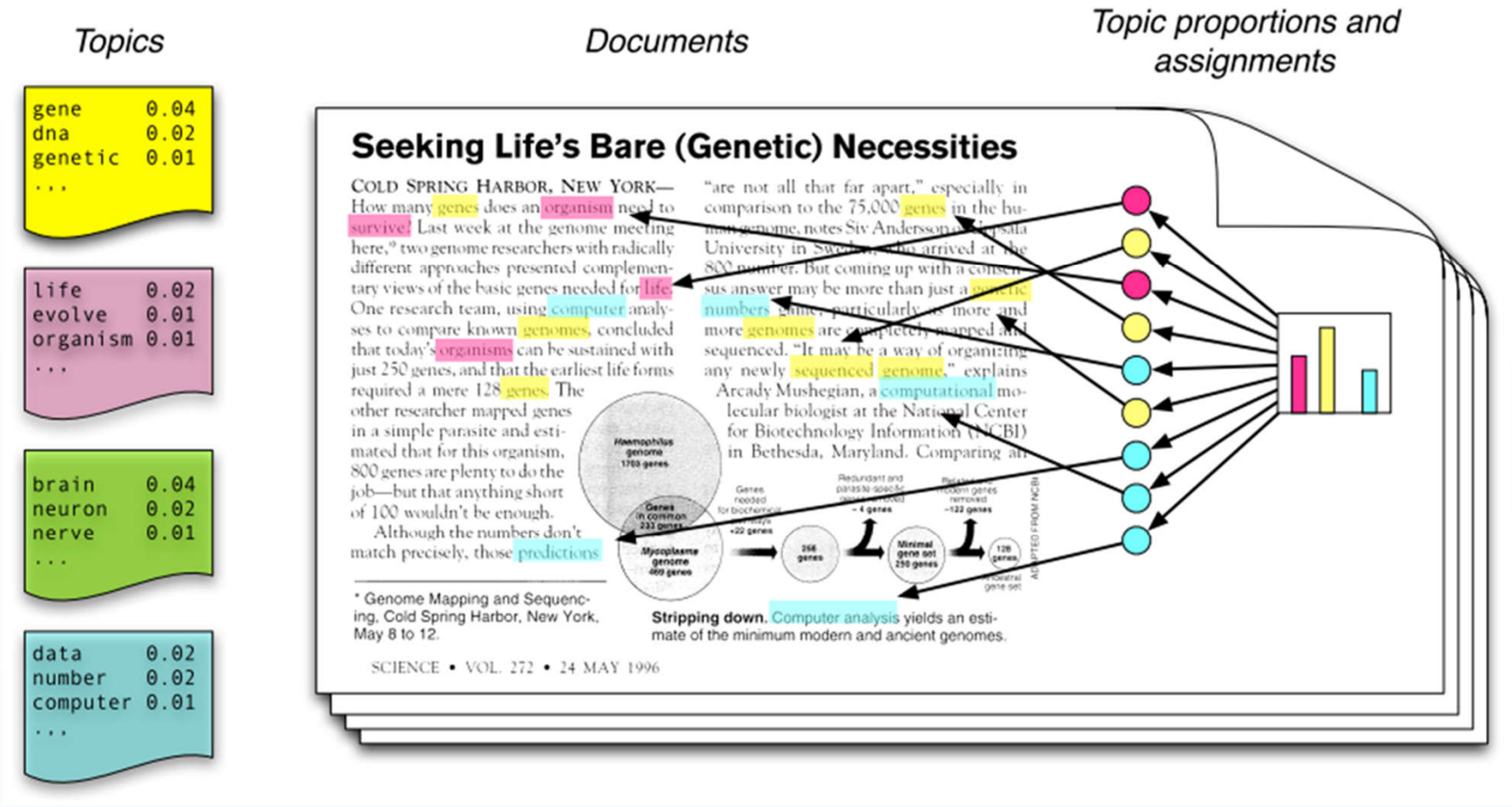
Will have large cosine but not truly related

from Lillian Lee

3.1.1 语言模型

- **主题模型：利用非监督方法获得文档中隐含的主题**
  - **非监督：数据集的类别(标签)未知，识别数据的簇与隐含模式**
  - **隐含主题所在空间是低维的**
- **分布假设（Distributional Hypothesis）：经常出现在相同上下文环境中的两个词具有语义上的相似性**
- **主题模型假设文档中的词是通过以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语生成的。**

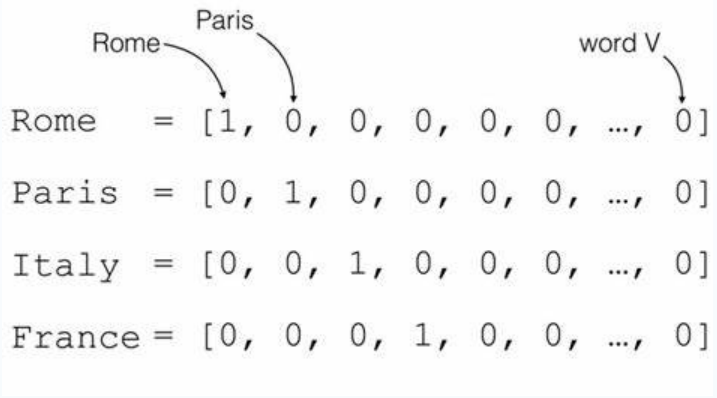
3.1.1 语言模型

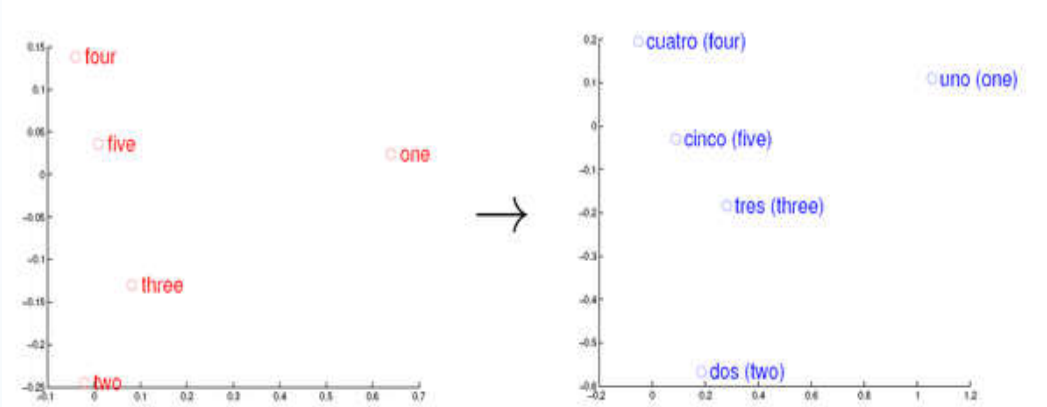


https://bigdata.oden.utexas.edu/wp-content/uploads/2015/01/LDA-concept.png

3.1.1 语言模型

- **神经网络语言模型：利用神经网络学习词汇、句子、字符等的分布式表示。**
  - One-hot Encoding：高维、稀疏，没有距离的概念?
  - 分布式表示（Distributed representation）：将语义信息分不到不同相互独立的维度上，低维、稠密，语义计算





陕西 − 西安 = 河南 − 郑州

3.1.2 BoW模型

- **Bag of Words：** 文本（段落或者文档）被看作是无序的词汇集合
  - ✓ 基本假设：词与词之间概率分布条件独立（在给定类别后每个词的概率分布与其他词无关）
  - ✓ 忽略语法、单词顺序
  - ✓ 简单，但是有效

John is quicker than Mary
Mary is quicker than John
From Christopher's ppt

McDonald's slims down spuds

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.
NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.
But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.
But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.
Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

"Bag of Words"

- 14 × McDonalds
- 12 × fat
- 11 × fries
- 8 × new
- 7 × french
- 6 × company, said, nutrition
- 5 × food, oil, percent, reduce, taste, Tuesday

3.1.3 TF-IDF

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

T_k = term k in document D_i

tf_{ik} = frequency of term T_k in document D_i

idf_k = inverse document frequency of term T_k in C

N = total number of documents in the collection C

n_k = the number of documents in C that contain T_k

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

Slide from Mitch Marcus

3.1.3 TF-IDF

TFIDF例子

D1 : This is a database system textbook

D2 : Oracle database sells for $1000 this year

D3 : My oracle database textbook for my database class

Raw frequencies:

	Database	System	Textbook	Oracle	Sells	Year	class
D1	1	1	1	0	0	0	0
D2	1	0	0	1	1	1	0
D3	2	0	1	1	0	0	1

3.1.3 TF-IDF

■ TFIDF例子

D1 : This is a database system textbook

D2 : Oracle database sells for $1000 this year

D3 : My oracle database textbook for my database class

Normalized frequencies:

	Database	System	Textbook	Oracle	Sells	Year	class
D1	1	1	1	0	0	0	0
D2	1	0	0	1	1	1	0
D3	1	0	0.5	0.5	0	0	0.5

3.1.3 TF−IDF

■ TFIDF例子

D1 : This is a database system textbook

D2 : Oracle database sells for $1000 this year

D3 : My oracle database textbook for my database class

Document frequencies:

Database	System	Textbook	Oracle	Sells	Year	class
3	1	2	2	1	1	1

3.1.3 TF–IDF

normalized term-frequency (tf_{ij})

	Database	System	Textbook	Oracle	Sells	Year	class
D1	1	1	1	0	0	0	0
D2	1	0	0	1	1	1	0
D3	1	0	0.5	0.5	0	0	0.5

Document frequency(df_j)

Database	3	**Sells**	1
System	1	**Year**	1
Textbook	2	**Class**	1
Oracle	2		

$w_{ij} = tf_{ij} * \log(d/df_j)$

	Database	System	Textbook	Oracle	Sells	Year	class
D1	1 * log3/3 = 0	1 * log 3/1 =1.548	1 * log(3/2) =0.548	0	0	0	0
D2	0	0	0	0.548	1.548	1.548	0
D3	0	0	0.5 * log(3/2) = 0.274	0.274	0	0	0.774

3.1.3 TF–IDF

Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log\frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0,\log\frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

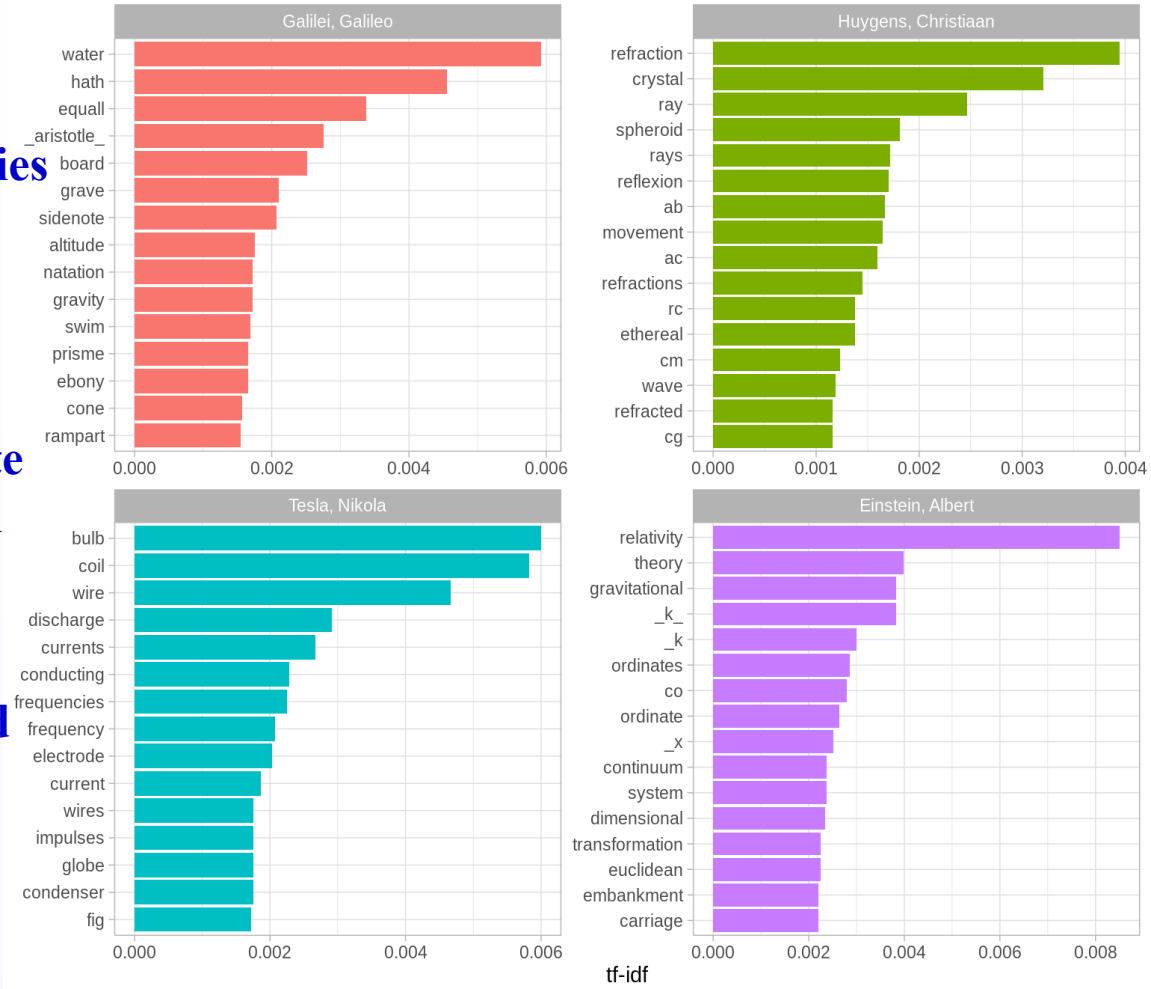
Raw term frequency is not what we want:

- **A document with 10 occurrences of the term is more relevant than a document with 1 occurrence of the term.**
- **But not 10 times more relevant.**
- **0 → 0, 1 → 1, 2 → 1.3, 10 → 2, 1000 → 4, etc.**

http://www.stanford.edu/class/cs276/handouts/lecture6-tfidf.ppt

3.1.3 TF–IDF

- ✓ **Discourse on Floating Bodies by Galileo Galilei**
- ✓ **Treatise on Light by Christiaan Huygens**
- ✓ **Experiments with Alternate Currents of High Potential and High Frequency by Nikola Tesla**
- ✓ **Relativity: The Special and General Theory by Albert Einstein.**



https://www.tidytextmining.com/tfidf.html

3.1.3 TF-IDF

关于TF-IDF的一组实验

https://www.tidytextmining.com/tfidf.html

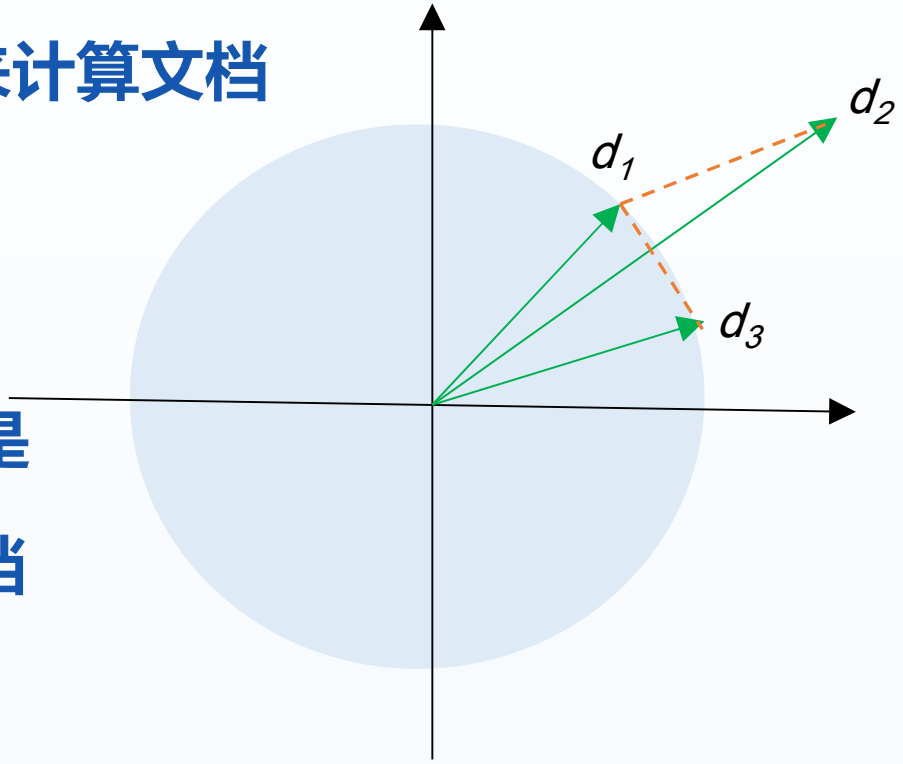
3.1.3 TF-IDF

■ 相似度计算（文档之间、文档与查询之间）

- ✓ 利用距离（欧氏距离）来计算文档之间的相似性

  $d(d_1, d_2) > d(d_1, d_3)$

- ✓ 如果$d_2 = 2*d_1$呢？（d_2是d_1复制自身后附加到文档尾部形成的文档）



3.1.3 TF-IDF

■ 相似度计算 $\cos(d_1, d_2)$

$$d_1 = w_{d_{11}}, w_{d_{12}}, \ldots, w_{d_{1t}}$$

$$d_2 = w_{d_{21}}, w_{d_{22}}, \ldots, w_{d_{2t}}$$

如果w被归一化：
$$sim(d_1, d_2) = \sum_{j=1}^{t} w_{1j} * w_{2j}$$

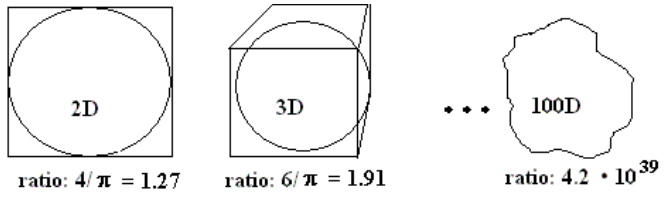
如果w未被归一化：
$$sim(d_1, d_2) = \frac{\sum_{j=1}^{t} w_{1j} * w_{2j}}{\sqrt{\sum_{j=1}^{t} (w_{1j})^2 * \sum_{j=1}^{t} (w_{2j})^2}}$$

3.1.4 NLU实例：情感分析、文本聚类

■ 文本聚类中的距离问题

✓ 高维数据时，传统距离方法不适用

$$\frac{V_{hypersphere}}{V_{hypercube}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0 \text{ as } d \to \infty.$$



$$\lim_{d\to\infty} E\left(\frac{\text{dist}_{\max}(d) - \text{dist}_{\min}(d)}{\text{dist}_{\min}(d)}\right) \to 0.$$

一个点与其他点的距离倾向于相等

数据倾向于分布在超立方的角上

— Committee of NRC of USA: Frontiers in Massive Data Analysis (2013)

3.1.4 NLU实例：情感分析、文本聚类

右图的实验结果表明：如果不控制稀疏度(蓝线)，结果很快收敛到0，在10,000与1,000,000处的纵坐标分别为0.0425与0.0045。但随着稀疏度的提高，收敛速度越来越慢。

由实验结果可推断出：**对于高度稀疏的文本向量，空间维度增加对距离计算的影响相对较小**。真实情况下，每个维度上的TF-IDF数值并不是符合实验设定中的均匀分布，因此，上述实验只是对该结论的一个粗略解释。

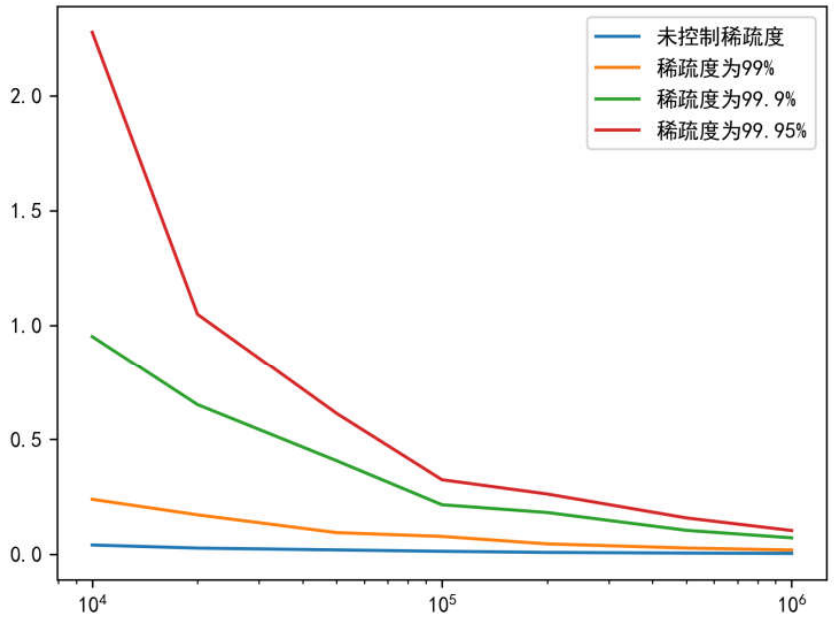
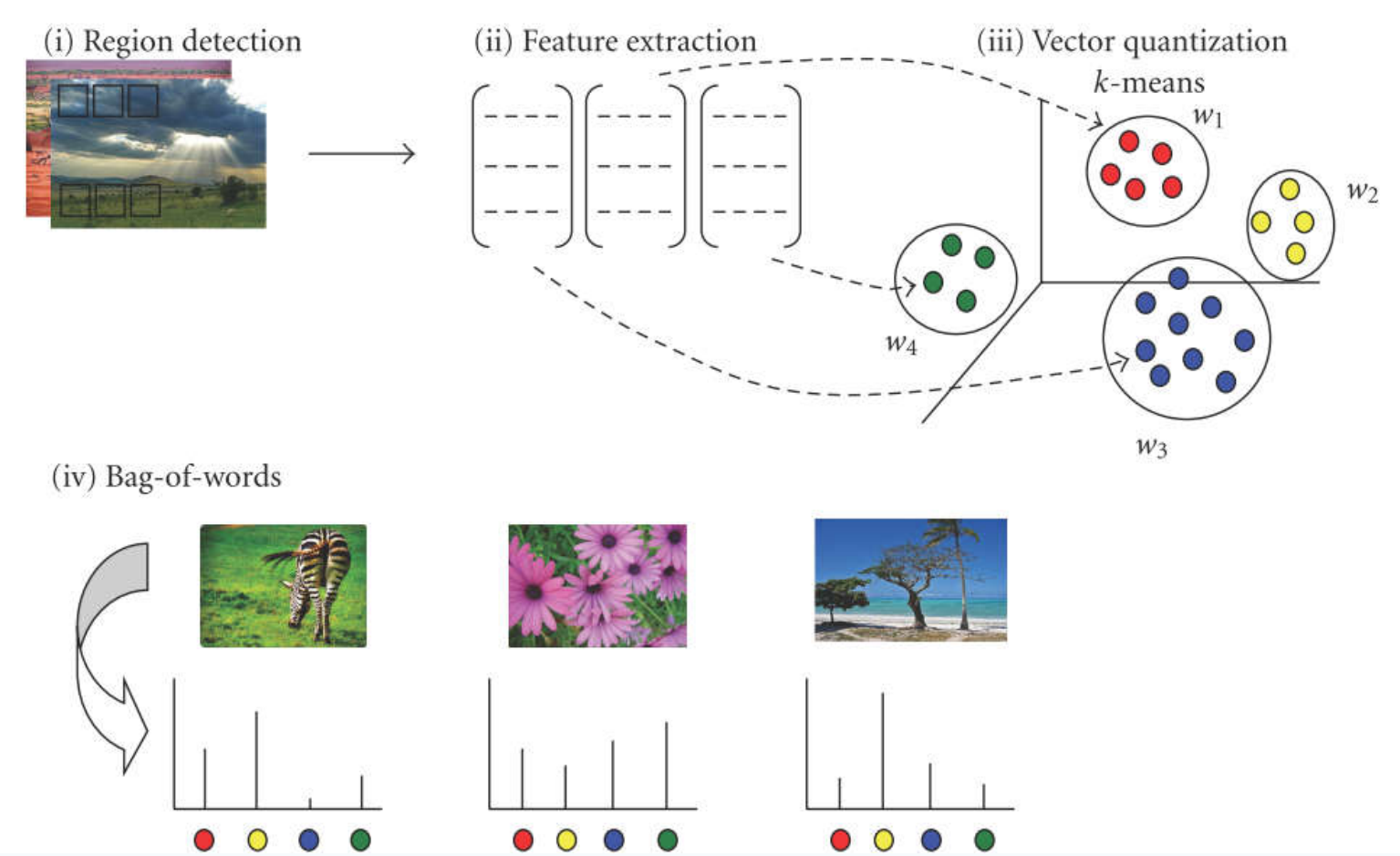


图 2.6: $(\text{dist}_{\max}(n) - \text{dist}_{\min}(n))/\text{dist}_{\min}(n)$随维度与稀疏度的变化趋势

图像中的BOW



Tsai C. Bag-of-Words Representation in Image Annotation: A Review. International Scholarly Research Notices, 2012: 1-19