Natural Language Processing I

9. 向量语义

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

2022/04/15

言者所以在意，得意而忘言

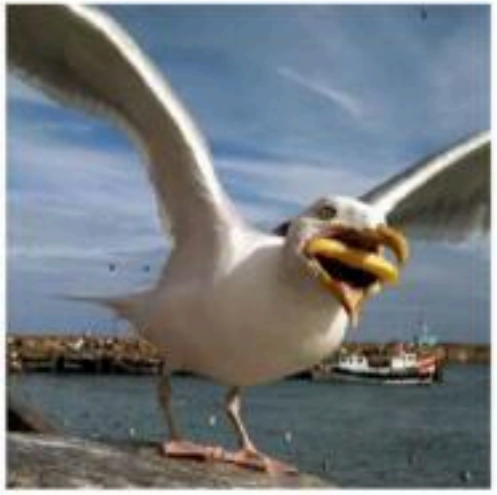
言与意

> *荃者所以在鱼，得鱼而忘荃*
> *言者所以在意，得意而忘言*
> *–《庄子·外物》*

待会儿去码头整点薯条

我们要飞向何方?
我打算待会儿去码头整点薯条。
你误会我了伙计，我说的是咱这一辈子的终极目标：活着是为了什么?
为了待会儿去码头整点薯条。

–《待会儿去码头整点薯条》



–《去码头整薯条的海鸥》

词义与上下文

出现在相似上下文的词通常有相似的词义

- [语言学家, 1950s] **分布假说 distributional hypothesis**
  - [Harris 1954] 词义的区别“大体上与上下文的区别正相关”
- 例如：说“视力测试”时，上下文出现“眼睛”的概率非常高

词义与上下文

出现在相似上下文的词通常有相似的词义

- [语言学家, 1950s] **分布假说 distributional hypothesis**
  - [Harris 1954] 词义的区别“大体上与上下文的区别正相关”
- 例如：说“视力测试”时，上下文出现“眼睛”的概率非常高

推论：文本**上下文中词的分布**对词义、语义理解非常重要

- 问题：如何将词转化为可以用来计算的数据表示?

向量语义与词嵌入

词嵌入 embedding：词义的向量表示

- 来自数学上的“嵌入”，本质上是（不同维数空间的）坐标系变换

向量语义与词嵌入

词嵌入 embedding：词义的向量表示

- 来自数学上的“嵌入”，本质上是（不同维数空间的）坐标系变换
- 词与词嵌入是**同一词义的不同表示**
  - 文本中的词：字符串表示
  - 语义空间中的词嵌入：向量表示

> *凡所有相皆是虚妄 – 《金刚经》*

向量语义与词嵌入

词嵌入 embedding：词义的向量表示

- 来自数学上的“嵌入”，本质上是（不同维数空间的）坐标系变换
- 词与词嵌入是**同一词义的不同表示**
  - 文本中的词：字符串表示
  - 语义空间中的词嵌入：向量表示

> *凡所有相皆是虚妄 – 《金刚经》*

向量语义 vector semantics：将词转换成语义空间中的等价向量表示

- 在向量空间中可以构造度量：比较词之间的差异
  - 能够**度量差异**，就能研究语义
- 向量表示可以直接用于存储、计算

向量表示：独热编码

独热编码 one-hot encoding：单有效位编码的向量

- $y = [y_1, y_2, .., y_C]$
- C个类；$y_c = 1$表示正确类别是c

例如数字3的编码：

- $[0, 0, 0, 1, 0, 0, 0, 0, 0, 0]$
- $C = 10, c = 3$

向量表示：独热编码

独热编码 one-hot encoding：单有效位编码的向量

- $y = [y_1, y_2, .., y_C]$
- C个类；$y_c = 1$表示正确类别是c

例如数字3的编码：

- $[0, 0, 0, 1, 0, 0, 0, 0, 0, 0]$
- $C = 10, c = 3$

为什么独热编码是很差的语义表示？相似度**差异绝对化**：非0即1

- 无法表示“有点相似”，如“优”、“良”

表示学习

回顾“学习”的两种观点：

- 数据观点：寻找对目标任务而言的最佳**数据表示**
- 函数观点：寻找逼近目标任务的最佳**显函数**

表示学习

回顾“学习”的两种观点：

- 数据观点：寻找对目标任务而言的最佳**数据表示**
- 函数观点：寻找逼近目标任务的最佳**显函数**

NLP强调自动化的**数据表示学习 representation learning**

- 而非人工特征工程，例如one-hot

词语义

如何表示词义?

到目前为止，讨论的都是字符串、词汇列表索引

- 本质上是**形式化替换**，类似于数理逻辑运算
  - “dog”的含义是DOG：$\forall x, DOG(x) \rightarrow MAMMAL(x)$

如何表示词义?

到目前为止，讨论的都是字符串、词汇列表索引

- 本质上是**形式化替换**，类似于数理逻辑运算
  - “dog”的含义是DOG：$\forall x, DOG(x) \rightarrow MAMMAL(x)$

[Carlson 1977] 语言学家的冷笑话

> *Q: What's the meaning of life?*
> *A: LIFE*

表示能力非常有限，几乎没有解决任何问题

对词义表示有何需求?

讨论表示方法之前，应该先考虑**需求**

词语义 Lexical Semantics：对词义的语言学研究

- 词义相近、相反，隐含意义
- 词形变化、多义词、异体字
- 同音字，多音字
- 可以用于推断任务

词目与词义

- mouse：老鼠；鼠标
- 量：名词；动词

词目与词义

- mouse：老鼠；鼠标
- 量：名词；动词

词目 lemma：词典中的不同条目

- 屈折语会涉及到不同**词形 wordform**，例如复数mice
- 中文**异体字**：“茴香豆的茴有几种写法？”
  - 茴，回，囘，囲，还有一种是“口”字里面有个“目”字

词目与词义

- mouse：老鼠；鼠标
- 量：名词；动词

词目 lemma：词典中的不同条目

- 屈折语会涉及到不同**词形 wordform**，例如复数mice
- 中文**异体字**："茴香豆的茴有几种写法？"
  - 茴，回，囘，囲，还有一种是"口"字里面有个"目"字

词义 word sense：同一词条的不同含义

- 一词多义非常普遍，对机器理解造成麻烦
  - **上下文**对词义理解至关重要，除非桌子上确实有老鼠

同义词

同义词 synonym：词义相同或非常接近

美丽、漂亮、帅、靓
顽强、顽固

- 有可能是性别、感情色彩的区分

同义词

同义词 synonym：词义相同或非常接近

美丽、漂亮、帅、靓
顽强、顽固

- 有可能是性别、感情色彩的区分

命题 propositional 词义：形式化定义，替换后不改变句子的真值条件

水、H_2O

- **形式区别**是语义的一部分
  - H_2O一般只用于科学上下文

同义？近义？

对比原理 principle of contrast：表现形式的改变一定带来含义区别

[Girard 1718] “我不认为任何语言里有同义词”

- 同种语言词义完全相同时：没必要区分词目
- 因此：我们讨论的其实是近义词


LA' JUSTESSE
DE LA
LANGUE FRANÇOISE,
OU
LES DIFFERENTES SIGNIFICATIONS
DES MOTS QUI PASSENT
POUR
SYNONIMES.
A PARIS,
M. DCC. XVIII.

间接讨论词义：相似度

直接讨论词义的悖论：**词义与表现形式绑定**

- 如果固定表现形式：给不同任务的计算带来困难
- 如果不确定表现形式：没法讨论问题

> *无名天地之始，有名万物之母。–《道德经》*

间接讨论词义：相似度

直接讨论词义的悖论：**词义与表现形式绑定**

- 如果固定表现形式：给不同任务的计算带来困难
- 如果不确定表现形式：没法讨论问题

> *无名天地之始，有名万物之母。–《道德经》*

间接讨论词义：只关注词（的表现形式）之间的**相似度**

> *Q: What is the universe?*
> *A: Not Nut, but very close.*
> *– Running like Clockwork*

- 为什么“very close”? – *Universe in a Nutshell*

间接讨论词义：相似度

直接讨论词义的悖论：**词义与表现形式绑定**

- 如果固定表现形式：给不同任务的计算带来困难
- 如果不确定表现形式：没法讨论问题

> *无名天地之始，有名万物之母。–《道德经》*

间接讨论词义：只关注词（的表现形式）之间的**相似度**

> *Q: What is the universe?*
> *A: Not Nut, but very close.*
> *– Running like Clockwork*

- 为什么“very close”? – *Universe in a Nutshell*
- 相似度可以用于构造**损失函数**
- 对后面**语义**理解任务同样有简化作用

如何计算相似度?

最直接的方法是人工标注

- 例如：[Hill 2015] SimLex-999 数据集

vanish	disappear	9.8
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

- 可以为机器学习任务提供训练集

如何计算相似度?

最直接的方法是人工标注

- 例如：[Hill 2015] SimLex-999 数据集

vanish	disappear	9.8
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

- 可以为机器学习任务提供训练集

缺点：**主观性**带来偏差

关联度：语义场

例如：

```
咖啡、茶：相似
茶、杯：相关联
```

关联度：经常在同一上下文出现

- 词义上几乎不共享任何特征

关联度：语义场

例如：

```
咖啡、茶：相似
茶、杯：相关联
```

关联度：经常在同一上下文出现

- 词义上几乎不共享任何特征

语义场 semantic field：针对特定领域的词集合，词间共享**结构关系**

- 住房：卧室、厨房、家电

关联度：语义框架、角色

语义框架 semantic frame：针对特定事件的词集合，词间共享**事件过程**

- 商业交易：买、卖、付款

关联度：语义框架、角色

语义框架 semantic frame：针对特定事件的词集合，词间共享**事件过程**

- 商业交易：买、卖、付款

同一语义框架有不同的**角色 role**

- 商业交易：买方、卖方、商品、金额

关联度：语义框架、角色

语义框架 semantic frame：针对特定事件的词集合，词间共享**事件过程**

- 商业交易：买、卖、付款

同一语义框架有不同的**角色 role**

- 商业交易：买方、卖方、商品、金额

鉴别语义框架有助于语法结构分析、机器理解：问答、翻译

张三向李四买了一本书。
李四卖给了张三一本书。

关联度：语义框架、角色

语义框架 semantic frame：针对特定事件的词集合，词间共享**事件过程**

- 商业交易：买、卖、付款

同一语义框架有不同的**角色 role**

- 商业交易：买方、卖方、商品、金额

鉴别语义框架有助于语法结构分析、机器理解：问答、翻译

```
张三向李四买了一本书。
李四卖给了张三一本书。
```

```
Ich esse Fisch.
Fisch esse Ich.

Fisch isst mich.
mich isst Fisch.
```

反义词

反义词 antonym：词义只在一个特征维度上相反

```
亮/暗，短/长
```

- 其他特征维度可以非常相似
  - 例如：通常可以充当相同的句子成分

反义词

反义词 antonym：词义只在一个特征维度上相反

> 亮/暗，短/长

- 其他特征维度可以非常相似
  - 例如：通常可以充当相同的句子成分

反向词 reversives：过程的逆向

> 上/下，涨/汐

同音字，多音字

中文语音识别任务的难点

同音字

这种食物可以zhì'ái。

- “致癌”还是“治癌”？

同音字，多音字

中文语音识别任务的难点

同音字

```
这种食物可以zhì'ái。
```

- “致癌”还是“治癌”?

多音字

```
张三还欠款100万。
```

- “hái”还是“huán”?

涵义

涵义 connotation：情感含义或内涵意义

- 限制讨论：情绪、情感、观点、评价等
- 正面、负面或积极、消极两个维度

涵义

涵义 connotation：情感含义或内涵意义

- 限制讨论：情绪、情感、观点、评价等
- 正面、负面或积极、消极两个维度

情感表达可以很细微

> 仿制、伪造，吸纳、抄袭

涵义

涵义 connotation：情感含义或内涵意义

- 限制讨论：情绪、情感、观点、评价等
- 正面、负面或积极、消极两个维度

情感表达可以很细微

> 仿制、伪造，吸纳、抄袭

应用：情感分析、立场识别，商业、政治特殊用语等

量化情感维度

[Osgood 1957] 情感词在三个主要（特征）维度上进行区分

- valence：带来的**愉悦程度**
- arousal：激起的**感情强度**
- dominance：施加的**控制级别**

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

量化情感维度

[Osgood 1957] 情感词在三个主要（特征）维度上进行区分

- valence：带来的**愉悦程度**
- arousal：激起的**感情强度**
- dominance：施加的**控制级别**

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

将词表示为空间中的点

- 开创了**向量语义**的建模方法

量化情感维度

[Osgood 1957] 情感词在三个主要（特征）维度上进行区分

- valence：带来的**愉悦程度**
- arousal：激起的**感情强度**
- dominance：施加的**控制级别**

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

将词表示为空间中的点

- 开创了**向量语义**的建模方法

本质上是**人为指定特征**，即特征工程

向量语义

语义的计算模型

在讨论了对语义的众多需求之后，我们能实现多少?

- 词义相近、相反，隐含意义
- 词形变化、多义词、异体字
- 同音字，多音字
- 可以用于推断任务

语义的计算模型

在讨论了对语义的众多需求之后，我们能实现多少?

- 词义相近、相反，隐含意义
- 词形变化、多义词、异体字
- 同音字，多音字
- 可以用于推断任务

向量语义 Vector Semantics：NLP中表示词义的标准方法

向量语义的起源

1950s两种观点的融合

- [Osgood 1957] 将词表示为**三维空间中的点**

> *[Wittgenstein 1953] "The meaning of a word lies in its use."*

- 词义即其用法的分布，即词或语法的**上下文环境**

> *[Harris 1954] "If A and B have almost identical environments we say that they are synonyms."*

什么是空心菜?

> *菜无心可活，人无心可不可活? —《封神演义》*

- 原产東南亞，但目前主要分布于中国长江以南亞热带地区。
  - 我作为北方人，非常好奇这个菜长什么样子

什么是空心菜?

> *菜无心可活，人无心可不可活? —《封神演义》*

- 原产東南亞，但目前主要分布于中国长江以南亞热带地区。
  - 我作为北方人，非常好奇这个菜长什么样子

网页搜索得到如下信息：

> ……将空心菜拌上大蒜酱
> 以绿叶和嫩茎供食用

同时，我知道如下信息：

> 蒜香菠菜盖浇饭
> 甜菜的茎和叶很好吃

什么是空心菜?

> *菜无心可活，人无心可不可活? —《封神演义》*

- 原产東南亞，但目前主要分布于中国长江以南亞热带地区。
  - 我作为北方人，非常好奇这个菜长什么样子

网页搜索得到如下信息：

> ……将空心菜拌上大蒜酱
> 以绿叶和嫩茎供食用

同时，我知道如下信息：

> 蒜香菠菜盖浇饭
> 甜菜的茎和叶很好吃

推断：空心菜是类似菠菜的绿叶蔬菜，茎也可以吃

向量语义与表示空间

两种思路的融合

- 将词表示为**多维空间中的点**
- 词义即其用法的分布，即词或语法的**上下文环境**

向量语义与表示空间

两种思路的融合

- 将词表示为**多维空间中的点**
- 词义即其用法的分布，即词或语法的**上下文环境**

向量语义：将词表示为多维语义空间中的点

- **语义空间**：由词的上下文分布特定决定
- **词嵌入**：语义空间中表示词义的向量

词嵌入的可视化

从情感分类任务训练得到的60维空间，投影（t-SNE）到二维平面

- 词嵌入：语义空间中的点

词嵌入的可视化

从情感分类任务训练得到的60维空间，投影（t-SNE）到二维平面

- 词嵌入：语义空间中的点
- 注意三个区域：正面（绿）、负面（红）、中性（蓝）
- 词义相近：在语义空间中相邻

向量化表示有何优点?

之前：词的**字符表示**

- 对应的特征：这个词的字符表示是否出现
- 需要**完全一致**的字符串：同时出现在训练集、测试集
  - 问题：很难处理未知词

向量化表示有何优点?

之前：词的**字符表示**

- 对应的特征：这个词的字符表示是否出现
- 需要**完全一致**的字符串：同时出现在训练集、测试集
  - 问题：很难处理未知词

现代NPL算法全都使用**词嵌入**作为词义的表示形式

- 对应的特征：词嵌入向量中的每个维度
- 可以推广到**相似但未知**的词
  - 比较（嵌入后）两点之间的距离

两种常用模型

tf-idf

- 信息检索的计算基础；常用作基准模型
- 稀疏表示：上下文中**词计数**的函数

两种常用模型

tf-idf

- 信息检索的计算基础；常用作基准模型
- 稀疏表示：上下文中**词计数**的函数

word2vec

- 稠密表示：训练**分类器预测**词义相近的可能性

词与向量

共生矩阵

共生 co-occurrence 矩阵：记录词同时出现的频率

- 词嵌入计数的基础工具

共生矩阵

共生 co-occurrence 矩阵：记录词同时出现的频率

- 词嵌入计数的基础工具

两种常用形式

- 词条-文档矩阵：文档内词的分布统计
- 词条-词条矩阵：语料库横向范围内词的分布统计

词条-文档矩阵

词条-文档矩阵 term-document matrix：每个文档由词的计数向量表示

例如 Shakespeare的四部戏剧

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- 行：词汇表中的词
- 列：文档集合
- 元素值：词在对应文档中的计数

文档的向量表示

文档：表示成列向量，每一列对应一个文档，即**文档向量**

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

文档的向量表示

文档：表示成列向量，每一列对应一个文档，即**文档向量**

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

- 文档向量的集合：在**语义向量空间**中
  - 维数：即行数，词汇表中词的数量

文档的特征

词：代表文档的一个**特征**

- 数值代表（以当前词度量的）**特征维度上的词义强度**

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

文档的特征

词：代表文档的一个**特征**

- 数值代表（以当前词度量的）**特征维度上的词义强度**

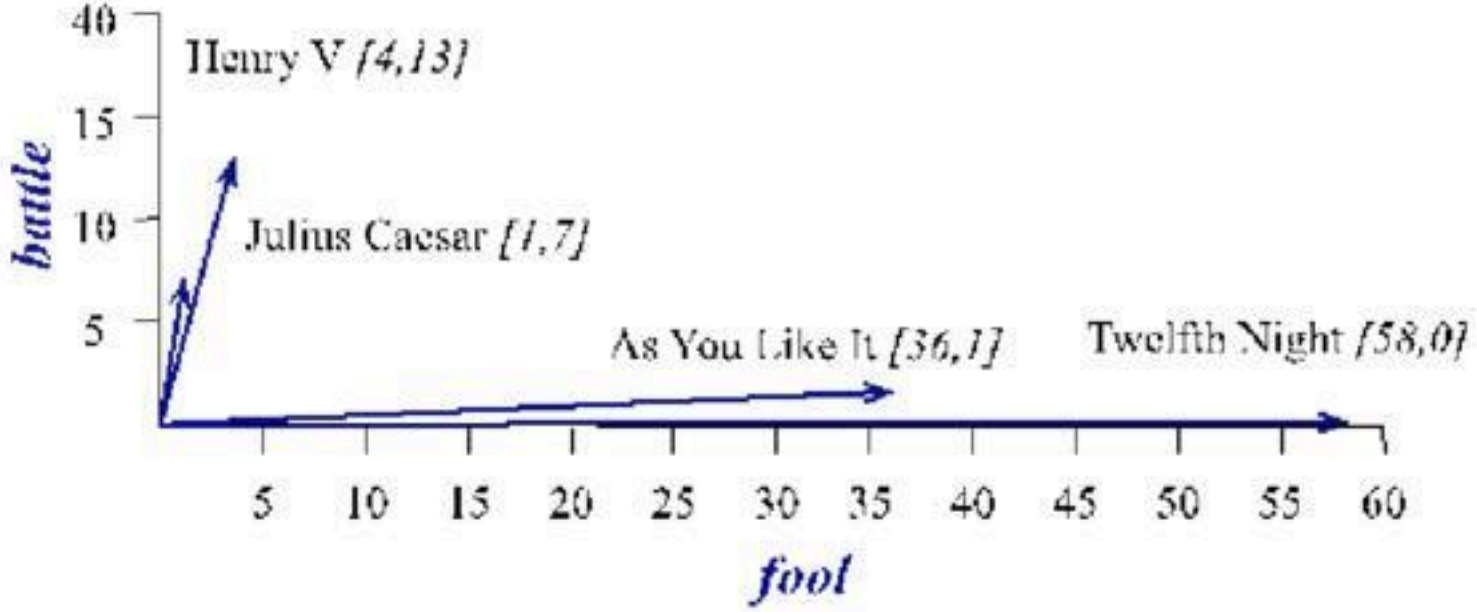
	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

文档向量通过（以词度量的）特征描述语义

- 不同特征间数值的相对大小：反映**文档风格**的不同
  - 应用：找出同作者、同类别的文章

文档向量的可视化

选取两个比较有代表性的特征



- 很清楚地将文档划分成两个类别：喜剧/悲剧

向量表示是信息检索的基础

[Salton 1971] 信息检索的向量空间模型：基于向量表示的**相似度**

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

一般情况：相似的列向量代表文档本身相似

- 例如：左边两部喜剧的数值分布较为相似

词的向量表示

词：表示成行向量，即**词向量**

- 每个元素代表（以当前文档度量的）**特征维度上的词义强度**

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

词的向量表示

词：表示成行向量，即**词向量**

- 每个元素代表（以当前文档度量的）**特征维度上的词义强度**

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

相似的词行向量代表**词义相似**

- 通常出现在相似的文档中
- 词义：表示成词在文档中的**分布统计**

词条-词条矩阵

词条-词条矩阵 term-term matrix：列也是词条，即**词向量**的另一形式

- 每个元素代表（以当前词度量的）**特征维度上的词义强度**

[Davies 2015] Wikipedia 语料集

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

- 列：代表**上下文**中的词
- 元素值：词条对**同时出现**的计数

上下文窗口

上下文可以是文档，但更常用**很小的上下文范围**

- 统计的是局部范围词之间的相关性
- 例如：前后各4个词，称为**上下文窗口**

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

上下文窗口

上下文可以是文档，但更常用**很小的上下文范围**

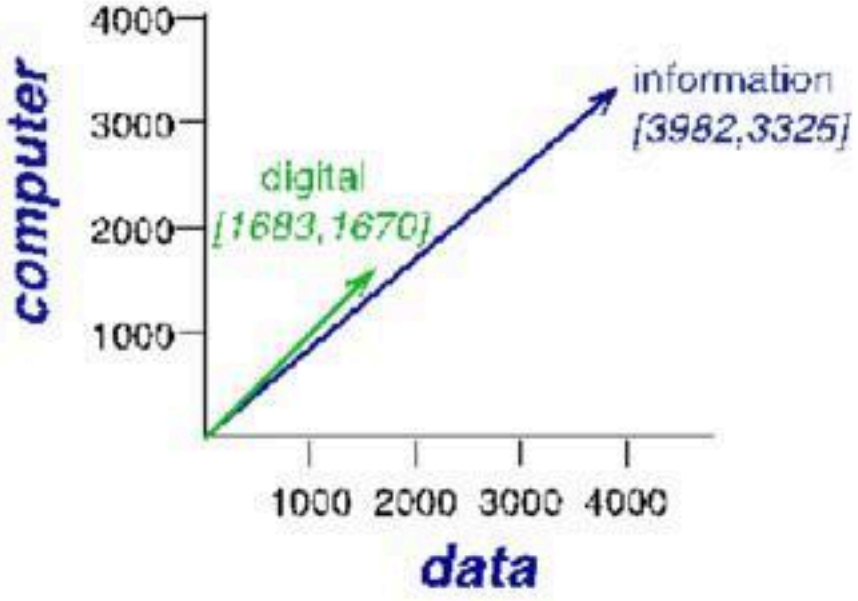
- 统计的是局部范围词之间的相关性
- 例如：前后各4个词，称为**上下文窗口**

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

- 一定是**对称**矩阵
- 相似的行/列向量：词义相似

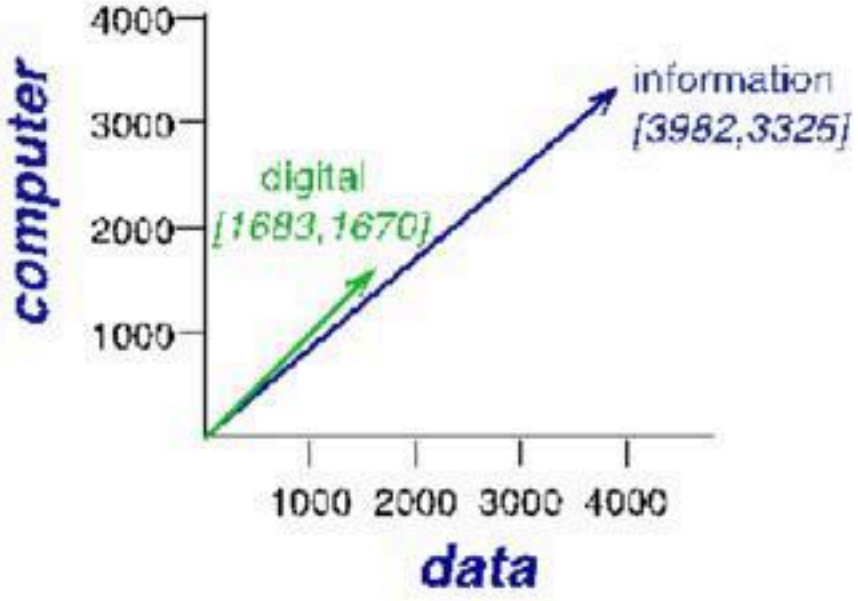
词向量的可视化

反过来也成立：选取两个关联度高的词，向量也比较相似



词向量的可视化

反过来也成立：选取两个关联度高的词，向量也比较相似



问题：如何度量向量间相似性?

余弦相似度

内积与度量

内积是**内积空间的标准度量**

$$\mathbf{v} \cdot \mathbf{w} = \sum_i v_i w_i = v_1 w_1 + v_2 w_2 + \dots$$

内积与度量

内积是**内积空间的标准度量**

$$\mathbf{v} \cdot \mathbf{w} = \sum_i v_i w_i = v_1 w_1 + v_2 w_2 + \dots$$

向量的内积是标量：

- 数值大时：向量在同方向上重合较大
- 最小值0：向量正交

绝对数值的问题

内积作为度量：更长的向量数值也更大

$$|\mathbf{v}| = \sqrt{\sum_i v_i^2}$$

绝对数值的问题

内积作为度量：更长的向量数值也更大

$$|\mathbf{v}| = \sqrt{\sum_i v_i^2}$$

频率更高的词：对应的向量更长

- 例如词条-词条计数：有些词（“的”等停用词）更有可能与其他词出现在一起

绝对数值的问题

内积作为度量：更长的向量数值也更大

$$|\mathbf{v}| = \sqrt{\sum_i v_i^2}$$

频率更高的词：对应的向量更长

- 例如词条-词条计数：有些词（“的”等停用词）更有可能与其他词出现在一起
- 这与理想度量相矛盾：度量应该反映对象的**内在特性**
  - 内在特性不随外在表现而变化，故度量应该是频率无关的

绝对数值的问题

内积作为度量：更长的向量数值也更大

$$|\mathbf{v}| = \sqrt{\sum_i v_i^2}$$

频率更高的词：对应的向量更长

- 例如词条-词条计数：有些词（“的”等停用词）更有可能与其他词出现在一起
- 这与理想度量相矛盾：度量应该反映对象的**内在特性**
  - 内在特性不随外在表现而变化，故度量应该是频率无关的
- 解决方法：应该对长度做**归一化**

cosine 相似度

对向量长度做归一化

$$\begin{aligned}\cos\theta &= \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} \\ &= \frac{\sum_i v_i w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}\end{aligned}$$

- 单位化向量：长度为1
  - 频率计数非负：取值范围是$[0,1]$

cosine 相似度

对向量长度做归一化

$$\begin{aligned}\cos\theta &= \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}||\mathbf{w}|} \\ &= \frac{\sum_i v_i w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}\end{aligned}$$

- 单位化向量：长度为1
  - 频率计数非负：取值范围是$[0,1]$

很明显，这就是**余弦定理**

$$\mathbf{v}\cdot\mathbf{w} = |\mathbf{v}||\mathbf{w}|\cos\theta$$

cosine 相似度：示例

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos\theta = \frac{\sum_i v_i w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}$$

- $\cos(\text{cherry}, \text{information}) = .017$
- $\cos(\text{digital}, \text{information}) = .996$

cosine 相似度：示例

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

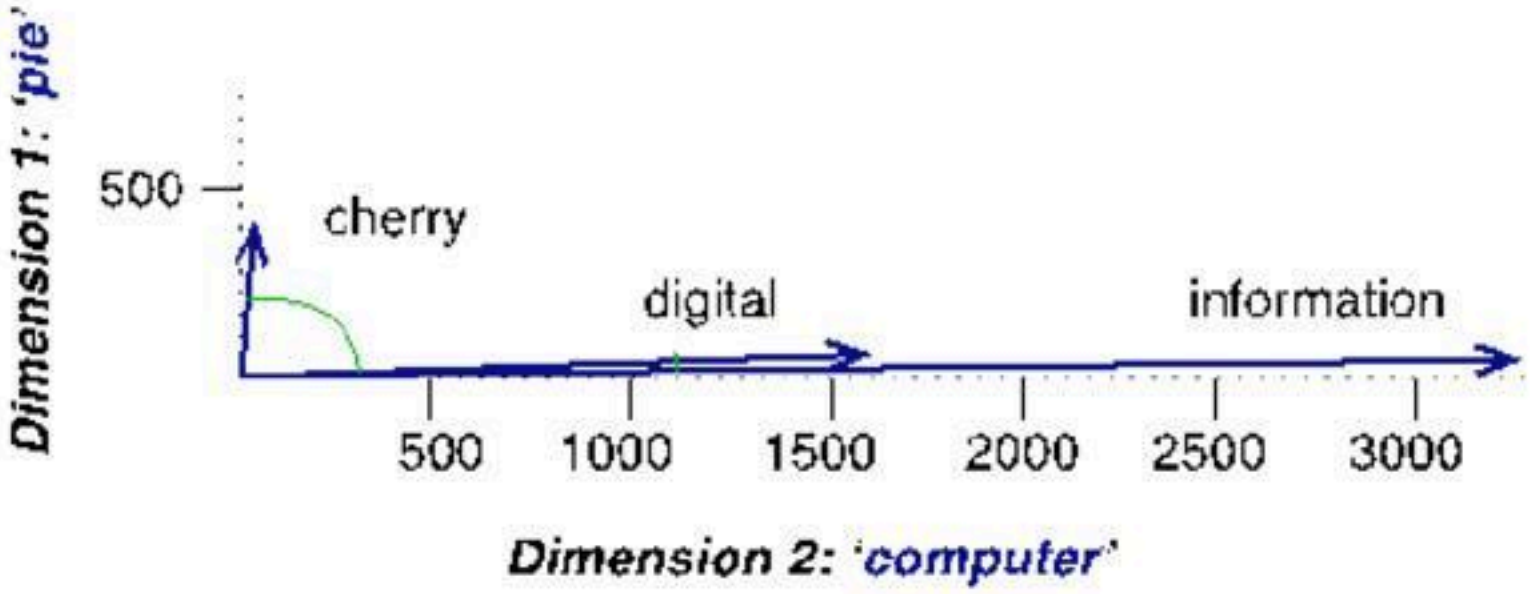
$$\cos\theta = \frac{\sum_i v_i w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}$$

- $\cos(\text{cherry, information}) = .017$
- $\cos(\text{digital, information}) = .996$

由相似度：很容易区分两组词间的**词义差距**

cosine 相似度：可视化

同样，已知两组词间的词义差距作为**先验知识**



- 由图中向量夹角：很容易验证相似度

TF-IDF

简单计数的问题

共生矩阵中的元素代表**词频**：相对文档、上下文

- 回忆Zipf定律：绝大多数高频词没有实意
  - `the`，`it`，了，的

简单计数的问题

共生矩阵中的元素代表**词频**：相对文档、上下文

- 回忆Zipf定律：绝大多数高频词没有实意
  - `the`，`it`，了，的
- 并且原始词频**辨识度**不强
  - 例如：`good`甚至可以涵盖中性、负面态度
    - 考核非常难的任务时，“好”可以认为是“勉强通过”

简单计数的问题

共生矩阵中的元素代表**词频**：相对文档、上下文

- 回忆Zipf定律：绝大多数高频词没有实意
  - `the`，`it`，了，的
- 并且原始词频**辨识度**不强
  - 例如：`good`甚至可以涵盖中性、负面态度
    - 考核非常难的任务时，“好”可以认为是“勉强通过”

词频悖论

- 上下文中高频词远比低频词重要
- 但**频率过高**的词反倒不重要：所有上下文都出现

简单计数的问题

共生矩阵中的元素代表**词频**：相对文档、上下文

- 回忆Zipf定律：绝大多数高频词没有实意
  - `the`，`it`，了，的
- 并且原始词频**辨识度**不强
  - 例如：`good`甚至可以涵盖中性、负面态度
    - 考核非常难的任务时，“好”可以认为是“勉强通过”

词频悖论

- 上下文中高频词远比低频词重要
- 但**频率过高**的词反倒不重要：所有上下文都出现

解决方法：需要**给词频加权**的合理方法

加权词频

TF-IDF：通常用于文档维度

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

- 减少**无辨识度**的高频词权重

加权词频

TF-IDF：通常用于文档维度

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

- 减少**无辨识度**的高频词权重

Pointwise Mutual Information (PMI)：通常用于词维度

$$\mathrm{PMI}(w, c) = \log_2 \frac{P(w, c)}{P(w)P(c)}$$

- 解决的问题：上下文中的词是不是**出现**的？例如“蝴蝶”
  - 计算思路：与**先验期望值**对比

词条频率 tf

之前使用文档中词条的简单计数

$$\mathrm{tf}_{t,d} \propto \Gamma(t,d)$$

词条频率 tf

之前使用文档中词条的简单计数

$$\mathrm{tf}_{t,d} \propto \Gamma(t,d)$$

[Luhn 1957] 修正后，按**对数“挤压”**数值

$$\mathrm{tf}_{t,d} = \log_{10}(\Gamma(t,d)+1)$$

- “挤压”：重要性随着词频增加而减小
  - 不能认为“好”x100次的文档比“好”x1次重要100倍

词条频率 tf

之前使用文档中词条的简单计数

$$\mathrm{tf}_{t,d} \propto \Gamma(t,d)$$

[Luhn 1957] 修正后，按**对数“挤压”**数值

$$\mathrm{tf}_{t,d} = \log_{10}(\Gamma(t,d)+1)$$

- “挤压”：重要性随着词频增加而减小
  - 不能认为“好”x100次的文档比“好”x1次重要100倍
- 加一：防止词频为零时计算对数

文档频率 df

df_t：词条t出现过的**文档数量**

- 区别：出现过的总数

例如 Shakespeare 的 37 部戏剧

	总数	df
Romeo	113	1
action	113	31

文档频率 df

df_t：词条t出现过的**文档数量**

- 区别：出现过的总数

例如 Shakespeare 的 37 部戏剧

	总数	df
Romeo	113	1
action	113	31

- “Romeo”是非常有**辨识度**的词：只有一部出现过

逆文档频率 idf

idf的意义：给df_t较低的词赋予更高的权重

$$\mathrm{idf}_t \propto \frac{1}{\mathrm{df}_t}$$

- df_t较低：**辨识度**高，能够帮助区分文档
- 反之，所有文档都出现的词并不重要

逆文档频率 idf

idf的意义：给df_t较低的词赋予更高的权重

$$\mathrm{idf}_t \propto \frac{1}{\mathrm{df}_t}$$

- df_t较低：**辨识度**高，能够帮助区分文档
- 反之，所有文档都出现的词并不重要

定义：文档总数N vs 文档频率df_t

$$\mathrm{idf}_t = \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

- 同样按对数“挤压”数值

逆文档频率 idf：示例

idf的意义：给df_t较低的词赋予更高的权重

$$\mathrm{idf}_t \propto \frac{1}{\mathrm{df}_t}$$

- df_t较低：**辨识度**高，能够帮助区分文档
- 反之，所有文档都出现的词并不重要

定义：文档总数N vs 文档频率df_t

$$\mathrm{idf}_t = \log_{10}(\frac{N}{\mathrm{df}_t})$$

- 同样按对数“挤压”数值

词	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

什么是文档?

通常意义的文档

- Shakespeare 的戏剧集
- 《资治通鉴》294卷

什么是文档?

通常意义的文档

- Shakespeare 的戏剧集
- 《资治通鉴》294卷

但文档可以是**任何语料**

- 跟任务有关，比如将一卷书按自然段拆分

tf-idf 定义

tf与idf的乘积

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

tf-idf 定义

tf与idf的乘积

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

- “good”出现在所有文档：不重要

逐点互信息

加权词频的两个维度

TF-IDF：通常用于文档维度

- 使用词条-文档矩阵计算

加权词频的两个维度

TF-IDF：通常用于文档维度

- 使用词条-文档矩阵计算

逐点互信息 **Pointwise mutual information (PMI)**：通常用于词维度

- 使用词条-词条矩阵计算

加权词频的两个维度

TF-IDF：通常用于文档维度

- 使用词条-文档矩阵计算

逐点互信息 **Pointwise mutual information (PMI)**：通常用于词维度

- 使用词条-词条矩阵计算
- 解决的问题：上下文中的词是不是**出现**的？例如“蝴蝶”
  - 计算思路：与**先验期望值**对比

PMI 定义

如何描述“偶然（同时）出现”？例如“上课”、“下雨”

- 数学上：独立事件、同时发生

PMI 定义

如何描述“偶然（同时）出现”？ 例如“上课”、“下雨”

- 数学上：独立事件、同时发生

[Fano 1961, Church 1989] 目标词w、上下文词c

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

- 分子：**同时观测**到两个词的可能性

PMI 定义

如何描述“偶然（同时）出现”？例如“上课”、“下雨”

- 数学上：独立事件、同时发生

[Fano 1961, Church 1989] 目标词w、上下文词c

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

- 分子：**同时观测**到两个词的可能性
- 分母：两个**独立事件同时发生**的概率
  - 等于两个事件概率的乘积

PMI （不）适用场景

趋向正无穷：用于找出**关联性强**的词

$$\mathrm{PMI}(w, c) = \log_2 \frac{P(w, c)}{P(w)P(c)}$$

PMI （不）适用场景

趋向正无穷：用于找出**关联性强**的词

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

趋向负无穷：极度不相关？通常**不可靠**

- 例如：单独出现的概率是10^{-6}
  - 只有共同出现的概率小于10^{-12}，才能确定两个词不相关
  - 这个数量级需要非常大的语料集才能计算、训练

PMI （不）适用场景

趋向正无穷：用于找出**关联性强**的词

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

趋向负无穷：极度不相关？通常**不可靠**

- 例如：单独出现的概率是10^{-6}
  - 只有共同出现的概率小于10^{-12}，才能确定两个词不相关
  - 这个数量级需要非常大的语料集才能计算、训练
- 更严重：一般来说**人也很难判断**两个词到底有多“不相关”

PMI （不）适用场景

趋向正无穷：用于找出**关联性强**的词

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

趋向负无穷：极度不相关？通常**不可靠**

- 例如：单独出现的概率是10^{-6}
  - 只有共同出现的概率小于10^{-12}，才能确定两个词不相关
  - 这个数量级需要非常大的语料集才能计算、训练
- 更严重：一般来说**人也很难判断**两个词到底有多“不相关”

因此，通常只讨论**非负区间**取值

非负PMI

非负 Positive PMI (PPMI)：将PMI数值截取在非负区间

$$\mathrm{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

非负PMI

非负 Positive PMI (PPMI)：将PMI数值截取在非负区间

$$\mathrm{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

给定词条-词条矩阵$F \in \mathbb{R}^{W \times C}$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}, p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}, p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

$$\mathrm{PPMI}_{ij} = \mathrm{PPMI}(w_i, c_j) = \max(\log_2 \frac{p_{ij}}{p_{i*}p_{*j}}, 0)$$

PPMI：示例 I

	computer	data	result	pie	sugar	$\Gamma(w)$
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
$\Gamma(c)$	4997	5673	473	512	61	11716

- $P(w = \text{information}, c = \text{data}) = \frac{3982}{11716} = .3399$
- $P(w = \text{information}) = \frac{7703}{11716} = .6575$
- $P(c = \text{data}) = \frac{5673}{11716} = .4842$
- $PPMI(\text{information}, \text{data}) = \log_2(.3399/(.6575 \times .4842)) = .0944$

PPMI：示例 II

	computer	data	result	pie	sugar
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

- cherry, strawberry 与 pie, sugar 非常相关
- data 与 information 轻度相关

PPMI 与罕见词

PPMI的问题：**低频词**的PMI值非常高

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

PPMI 与罕见词

PPMI的问题：**低频词**的PMI值非常高

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

解决办法：**幂函数修正**上下文词的概率

$$\mathrm{PPMI}_\alpha(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P_\alpha(c)}, 0)$$

$$P_\alpha(c) = \frac{\Gamma(c)^\alpha}{\sum_c \Gamma(c)^\alpha}$$

PPMI 与罕见词

PPMI的问题：**低频词**的PMI值非常高

$$\mathrm{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

解决办法：**幂函数修正**上下文词的概率

$$\mathrm{PPMI}_\alpha(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P_\alpha(c)}, 0)$$

$$P_\alpha(c) = \frac{\Gamma(c)^\alpha}{\sum_c \Gamma(c)^\alpha}$$

- [Levy 2015] $\alpha = 0.75$时，可以改善很多任务的计算性能
  - $P(c) \ll 1, \alpha < 1 \Rightarrow P_\alpha(c) > P(c)$
  - 例如：$P(a) = .99, P(b) = .01 \Rightarrow P_\alpha(a) = .97, P_\alpha(b) = .03$

向量模型的应用

回顾：向量语法

向量语法模型将词表示为向量，向量中的元素是**加权后的计数**

- 词条-文档：文档集；TF-IDF
- 词条-词条：上下文窗口；PPMI

向量通常是稀疏的

回顾：向量语法

向量语法模型将词表示为向量，向量中的元素是**加权后的计数**

- 词条-文档：文档集；TF-IDF
- 词条-词条：上下文窗口；PPMI

向量通常是稀疏的

TF-IDF模型、PPMI模型

- 相似度：余弦值

文档相似度

TF-IDF模型通常用于确定**文档之间**的相似度

- 将文档表示成所有词向量的**重心**：$d = \frac{\sum_i^k w_i}{k}$
- 使用余弦相似度比较文档：$\cos(d_1, d_2)$

文档相似度

TF-IDF模型通常用于确定**文档之间**的相似度

- 将文档表示成所有词向量的**重心**：$d = \frac{\sum_i^k w_i}{k}$
- 使用余弦相似度比较文档：$\cos(d_1, d_2)$

应用：信息检索，抄袭检测，推荐系统等

词相似度

TF-IDF模型和PPMI模型都可以用于比较**词之间**的相似度

应用：查找同义段落，跟踪词义变化，自动发现未知词义等

Review

本章内容

向量语义、词嵌入。词义需求。向量化词义。余弦相似度。TF-IDF模型、PPMI模型。

重点：向量语义、词嵌入；向量表示、上下文分布；共生矩阵（词条-文档矩阵、词条-词条矩阵）；余弦相似度；TF-IDF模型、PPMI模型。

难点：向量表示、上下文分布的优点。

学习目标

- 理解词、词嵌入都是词义的表现形式。
- 理解间接讨论语义的原因（形、义不绑定）、方法（相似度）。
- 理解语言学对语义的需求。
- 理解向量语义是两种观点的融合：向量表示、上下文分布。
- 了解向量表示的优点：便于计算、压缩信息、容易处理未知词。
- 了解上下文分布的优点：通过相似度间接讨论词义。
- 理解共生矩阵的两种常用形式（词条-文档矩阵、词条-词条矩阵），及其用法（行、列代表词义）。
- 理解余弦相似度。
- 理解TF-IDF模型的构造方法（统计词条-文档频数）、解决的问题（降低高频词的重要性）。
- 理解PPMI模型的构造方法（统计词条-词条频数）、解决的问题（判断是否偶然同时出现）。

问题

简述启发向量语义的两种观点，及其（间接讨论词义的）优点。

简述向量语义、词嵌入的含义。

简述共生矩阵的两种常用形式，及其用法。

简述TF-IDF模型、PPMI模型的构造方法、解决的问题。