Векторная модель документа

Квантитативный анализ текста

Кирилл Александрович Маслинский

14.03.2022 / 04

ЕУ СПб

Настройка зрения

> However, I'll argue against the commonly held notion that counting words reduces complexity, and I'll suggest instead that semantic models embed textual objects in highly complex structures that, when constructed using relevant corpora, are extremely sensitive to historical context and subtle nuances in meaning.

Michael Gavin *Is there a text in my data? (Part 1): On Counting Words* // Journal of Cultural Analytics. 09.17.19

Векторная модель документа

План

Векторная модель документа

- Bag-of-words: мешок слов
- Вектора в многомерном пространстве
- От частотного списка к матрице

Модель документа: мешок слов

Для статистического анализа необходимо преобразование:

Текст $\longrightarrow$ Набор чисел/свойств

- Каждое слово текста [type] — свойство
- Частотность слова — значение свойства

Векторное представление документа

- Мне, пожалуйста, двойной виски. - Девочка! Это школьная столовая! - Ой, извините, я задумалась. Компот, пожалуйста...

виски	двойной	девочка	задумалась	извините	
1	1	1	1	1	
компот	мне	пожалуйста	столовая	школьн	
1	1	2	1	1	
это					
1					

План

Векторная модель документа

- Bag-of-words: мешок слов
- Вектора в многомерном пространстве
- От частотного списка к матрице

Вектор

9 6 8

Вектор

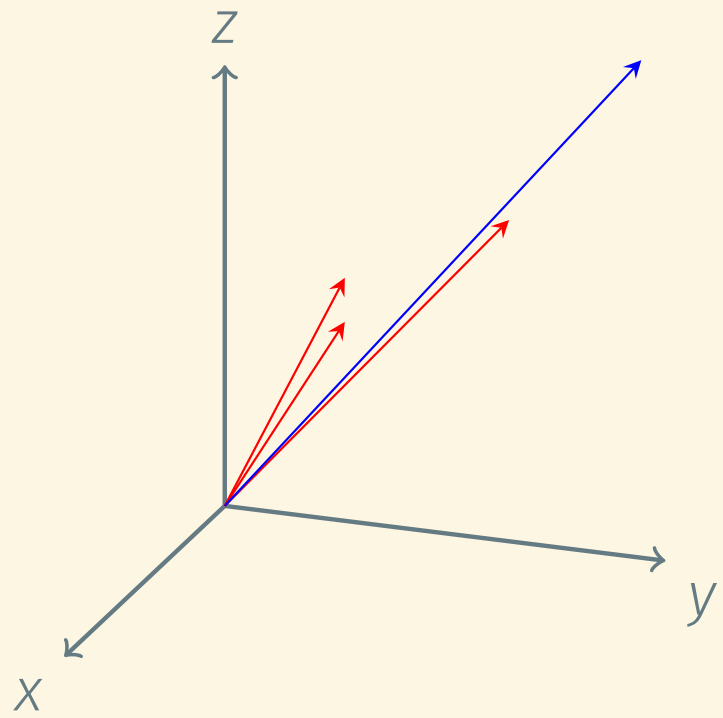
я	ты	мы
9	6	8

Вектор

я	ты	мы
9	6	8
9	6	9
7	9	10

Вектор

я	ты	мы
9	6	8
9	6	9
7	9	10



Меры расстояния

Сходство документов = близость векторов (в N-мерном пространстве)

- Евклидова мера $L_2(x, y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$
- $L_1(x, y) = \sum_{i=1}^{m} |x_i - y_i|$
- Косинусная мера $1 - \frac{x \cdot y}{|x| \cdot |y|} = 1 - \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}}$

План

Векторная модель документа

- Bag-of-words: мешок слов
- Вектора в многомерном пространстве
- От частотного списка к матрице

Матрица терминов-документов

Термины с наибольшим DF (не менее 15%)

Docs	вовочк	дет	класс	урок	учител	учительниц	школ
1	2	0	0	1	0	1	0
2	2	0	0	0	0	1	0
3	1	1	0	0	0	1	2
4	3	1	0	1	0	1	0
5	3	0	0	0	0	1	0
6	1	0	0	0	0	1	0

Матрица частотностей описывает наблюдаемые пересечения двух множеств:

- множество допустимых слов (types)
- множество допустимых текстовых объектов

Задавая фиксированный словарь, мы определяем поле возможностей, в рамках которого могут быть описаны наши документы.

Например, у Шекспира:

ipad	такого слова еще не было (нет смысла его включать)
theology	мог бы использовать (как некоторые современники), но не использовал
christ	никогда не использовал это слово в комедиях, но зато использовал в нескольких исторических пьесах

Лексическое пространство × Пространство документов

- **каждый документ** — система из слов, сделанных из документов
- **каждое слово** — система из документов, сделанных из слов

Текст в матрице — структурированное множество исторически релевантных связей с исторически релевантными контекстами.

Лексическое пространство × Пространство документов

- **каждый документ** — система из слов, сделанных из документов
- **каждое слово** — система из документов, сделанных из слов

Текст в матрице —
структурированное множество исторически релевантных связей с исторически релевантными контекстами.

Анализ корпуса на уровне документов

План

Анализ корпуса на уровне документов

Лексическая дисперсия

Взвешенная частотность: TF-IDF

Распределение лексики по документам

- Частотность отражает значимость слова в языке/коллекции документов
- Но сильно зависит от состава коллекции, ср. частотность слова хоббит
- В дополнение к частотности нужно учитывать дисперсию — насколько равномерно слово распределено по документам

IDF: обратная документная частота

$$IDF = \log_2 \frac{D}{df} \tag{1}$$

где

D — количество документов в корпусе

df — количество документов, содержащих термин (x раз)

IDF: примеры

распределение	D	df	IDF
везде	10000	10000	0
часто	10000	1000	3,32
достаточно	10000	100	6,64
несколько	10000	10	9,96
в одном документе	10000	1	13,29

План

Анализ корпуса на уровне документов

Лексическая дисперсия

Взвешенная частотность: TF-IDF

Поправка на дисперсию

Идея: откорректировать ранжирование в частотном списке в соответствии с дисперсией. Задачи:

информационный поиск понизить ранг слов, распределенных равномерно
Цель — повысить ранг слов, различающих отдельные документы.

частотные словари понизить ранг слов, распределенных неравномерно
Цель — понизить ранг слов, получивших неоправданно высокую частотность в силу особенностей состава корпуса.

TF-IDF

Мера, предложенная для информационного поиска: выделение релевантных слов документа:

$$TF \times IDF = tf \log_2 \frac{D}{df} \tag{2}$$

tf частота термина (в документе или в коллекции), term frequency

df число документов, в которых встречается термин, document frequency

D общее число документов в коллекции

Взвешивание терминов

Слова должны различать документы:

- не слишком частотные (неинформативны, не позволяют разделять различные документы)
- не слишком редкие (не позволяют объединять сходные документы)

Нормализация по длине текста

Нормализация: разделить каждое значение на количество слов в документе

	Terms						
Docs	вовочк	дет	класс	урок	учител	учительниц	шко
1	0.5000000	0.0000000	0	0.2500000	0	0.2500000	0.
2	0.6666667	0.0000000	0	0.0000000	0	0.3333333	0.
3	0.2000000	0.2000000	0	0.0000000	0	0.2000000	0.
4	0.5000000	0.1666667	0	0.1666667	0	0.1666667	0.
5	0.7500000	0.0000000	0	0.0000000	0	0.2500000	0.
6	0.5000000	0.0000000	0	0.0000000	0	0.5000000	0.

Взвешивание терминов: TF-IDF

Нормализация: вместо частоты слова (TF) его взвешенная частота (IDF)

Docs	Terms вовочк	дет	класс	урок	учител	учительниц	
1	0.6897996	0.0000000	0	0.4192463	0	0.4022703	0.0
2	0.9197328	0.0000000	0	0.0000000	0	0.5363604	0.0
3	0.2759198	0.4781918	0	0.0000000	0	0.3218162	0.4
4	0.6897996	0.3984932	0	0.2794975	0	0.2681802	0.0
5	1.0346994	0.0000000	0	0.0000000	0	0.4022703	0.0
6	0.6897996	0.0000000	0	0.0000000	0	0.8045405	0.0

Косинусное расстояние

```
> dissimilarity(sp, method="cosine")
          1          2         3         4          5
1 0.0000000 0.11461112 0.5542388 0.1226977 0.12552228
2 0.1146111 0.00000000 0.4965362 0.1747931 0.01232358
3 0.5542388 0.49653624 0.0000000 0.3369433 0.53009027
4 0.1226977 0.17479311 0.3369433 0.0000000 0.16449672
5 0.1255223 0.01232358 0.5300903 0.1644967 0.00000000
```

Docs	вовочк	дет	класс	урок	учител	учительниц	школ
1	2	0	0	1	0	1	0
2	2	0	0	0	0	1	0
3	1	1	0	0	0	1	2
4	3	1	0	1	0	1	0
5	3	0	0	0	0	1	0