Vector space model

Computational Methods for Text Analysis

Pestova Alena

HSE Saint-Petersburg

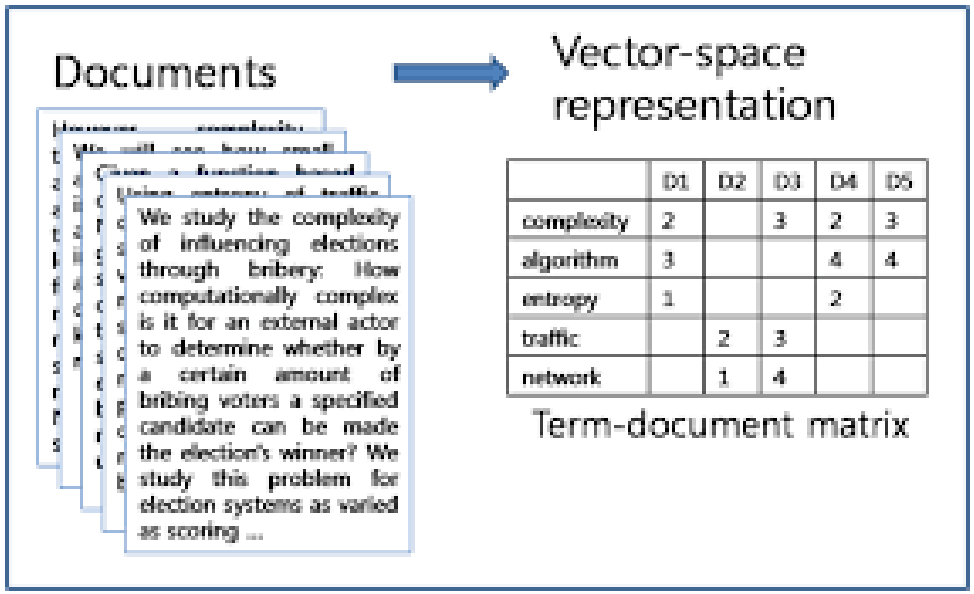
04.10.2021

Vector space model: bag of words

For statistical analysis we need a transformation:

Text $\longrightarrow$ Set of features

- Each word of the text [type] — свойство
- Word Frequency - a value of the feature



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Why do we need to transoform documents to vectors?

- to compare them, to separate them into groups
- for classification
- many other tasks

Measures of distance

Similarity of documents = vectors similarity (in N-dimensional space)

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

Document-term matrix

Terms with the highest DF (document frequency)

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

The frequency matrix describes the observed intersections of two sets:

- set of allowed words (types)
- set of valid text objects

Vocabulary

By defining a fixed dictionary, we define the features , within which our documents can be described.
For example, in Shakespeare:

ipad	there were no such a word (there is no sense in including it)
theology	could have used (as some him contemporaries do), but did not use
christ	never used this word in comedies, but used in several historical plays

Distribution of vocabulary across documents

- Frequency reflects the importance of the word in the language/document collection
- But it strongly depends on the composition of the collection, (for ex: mean word frequency hobbit)
- In addition to frequency, dispersion needs to be taken into account — how evenly the word is distributed across the documents

Document frequency

- We can just take the document frequency of each word and put to the matrix
- The problem - the length of documents varies greatly
- The second problem - not all words characterize the document (stop-words and just common words)

IDF: inverse document frequency

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

где

- D — the number of documents in the corpus
- df — the frequency of the word in the collection of documents

IDF: examples

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

Correction for variance

Idea: adjust the ranking in the frequency list in according to dispersion. Tasks:

information search lower the rank of words distributed evenly
: The goal is to increase the rank of words that distinguish individual documents.

frequency dictionaries lower the rank of words allocated unevenly
: The goal is to lower the rank of words that received an unreasonably high frequency due to the composition of the case.

TF-IDF

Term-frequency-inversed document frequency - is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

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

- tf term frequency - is the relative frequency of term t within document d (raw count of a term in a document / the total number of terms in document)
- df document frequency - the frequency of the word in the collection of documents
- $\frac{D}{df}$ inverse document frequency - a measure of how much information the word provides, i.e., if it is common or rare across all documents.
- D the number of all documents in the corpus

Weighing Terms

Words should distinguish documents:

- not too frequent (uninformative, do not allow to separate various documents)
- are not too rare (do not allow similar documents to be combined)

TF-IDF

Normalization: instead of simple word count (DF) we can weighted frequency (TF-IDF)

	Terms						
Docs	вовочк	дет	класс	урок	учител	учительниц	школ
1	0.6897996	0.0000000	0	0.4192463	0	0.4022703	0.0000000
2	0.9197328	0.0000000	0	0.0000000	0	0.5363604	0.0000000
3	0.2759198	0.4781918	0	0.0000000	0	0.3218162	0.4736512
4	0.6897996	0.3984932	0	0.2794975	0	0.2681802	0.0000000
5	1.0346994	0.0000000	0	0.0000000	0	0.4022703	0.0000000
6	0.6897996	0.0000000	0	0.0000000	0	0.8045405	0.0000000

Cosine distance

```
> 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
```