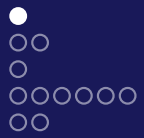
Классификация текстов

topic modeling

«Высшая школа экономики»

8.10



Outline

Классификация текстов

LSI

Другие методы

Классификация текстов

Text classification, text tagging, text categorization, rubrication

- Sentiment Analysis
- Topic Detection (modeling)
- Language Detection
- Exploratory Data Analysis

Классификация текстов

Techniques:

- distances
- KNN
- Kmeans
- PCA
- regression
- trees
- etc (many other variations)

Topic modeling

Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents.

Topic - in fact several important words.

- *LSI*, LDA
- PLSA, HDP

LSI

LSI - topic modeling techniques based on SVD decomposition.

- Easy to understand
- Easy to specify
- Fast

LSI

Pipeline:

input: corpus of documents, number of topics (n).

- Normalization, preprocessing
- Matrix (M) doc-term via BOW
- SVD decomposition
- get 3 matrices $M = U \times \Sigma \times V^{\mathsf{T}}$

LSI

Decomposition notation:

- M - initial matrix document×terms .
- U - docs×topics.
- Σ - topics×topics.
- V^T - topics×terms.

LSI

1. полиция основателя WikiLeaks
2. суде США против
3. Церемонию вручения Нобелевской премии стран
4. Великобритании арестован основатель Wikileaks
5. церемонию вручения Нобелевской премии
6. суд основателя Wikileaks
7. США стран против
8. Полиция Великобритании основателя WikiLeaks арестовала
9. вручение Нобелевских премий

	T1	T2	T3	T4	T5	T6	T7	T8	T9
wikileaks	1	0	0	1	0	1	0	1	0
арестовать	0	0	0	1	0	0	0	1	0
великобритания	0	0	0	1	0	0	0	1	0
вручение	0	0	1	0	1	0	0	0	1
нобелевский	0	0	1	0	1	0	0	0	1
основатель	1	0	0	1	0	1	0	1	0
полиция	1	0	0	0	0	0	0	1	0
премия	0	0	1	0	1	0	0	0	1
против	0	1	0	0	0	0	1	0	0
страна	0	0	1	0	0	0	1	0	0
суд	0	1	0	0	0	1	0	0	0
сша	0	1	0	0	0	0	1	0	0
церемония	0	0	1	0	1	0	0	0	0

LSI

wikileaks	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	-0.64
арестовать	0.34	0	0.07	0.41	-0.42	-0.02	0.1	0.17	0.01
великобритания	0.34	0	0.07	0.41	-0.42	-0.02	0.1	0.17	-0.01
вручение	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.07
нобелевский	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	0.32
основатель	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	0.64
полиция	0.31	0	0.05	0.07	0.57	-0.6	0.29	0.37	0
премия	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.25
против	0.02	0.03	-0.61	0.13	-0.05	-0.22	1	-0.25	0
страна	0.01	0.22	-0.31	0.39	0.41	0.56	-0.22	0.4	0
суд	0.12	0.01	-0.38	-0.62	-0.3	0.12	0.21	0.55	0
сша	0.02	0.03	-0.61	0.13	-0.05	-0.22	0	-0.25	0
церемония	0	0.38	0.03	0.02	0.08	0.31	0.82	-0.29	0

3.14	0	0	0	0	0	0	0	0
0	3.3	0	0	0	0	0	0	0
0	0	2.27	0	0	0	0	0	0
0	0	0	1.49	0	0	0	0	0
0	0	0	0	1.19	0	0	0	0
0	0	0	0	0	0.98	0	0	0
0	0	0	0	0	0	0.71	0	0
0	0	0	0	0	0	0	0.43	0
0	0	0	0	0	0	0	0	0

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.43	0.05	0.01	0.54	0	0.37	0.01	0.63	0
0	0.02	0.65	-0.01	0.59	0	0.09	-0.01	0.47
0.03	-0.7	-0.04	0.06	0.1	-0.16	-0.67	0.09	0.09
-0.22	-0.24	0.15	0.28	-0.11	-0.68	0.44	0.33	-0.13
0.69	-0.32	0.22	-0.49	-0.12	-0.03	0.27	-0.02	-0.19
-0.27	-0.34	0.44	0.29	-0.13	0.45	0.12	-0.31	-0.45
-0.03	0.3	0.14	-0.17	0.44	-0.15	-0.3	0.24	-0.71
-0.3	0.12	0.4	-0.39	-0.53	0.12	-0.23	0.46	0.13
0.35	0.35	0.35	0.35	-0.35	-0.35	-0.35	-0.35	0

LSI

wikileaks	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	-0.64
арестовать	0.34	0	0.07	0.41	-0.42	-0.02	0.1	0.17	0.01
великобритания	0.34	0	0.07	0.41	-0.42	-0.02	0.1	0.17	-0.01
вручение	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.07
нобелевский	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	0.32
основатель	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	0.64
полиция	0.31	0	0.05	0.07	0.57	-0.6	0.29	0.37	0
премия	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.25
против	0.02	0.03	-0.61	0.13	-0.05	-0.22	1	-0.25	0
страна	0.01	0.22	-0.31	0.39	0.41	0.56	-0.22	0.4	0
суд	0.12	0.01	-0.38	-0.62	-0.3	0.12	0.21	0.55	0
сша	0.02	0.03	-0.61	0.13	-0.05	-0.22	0	-0.25	0
церемония	0	0.38	0.03	0.02	0.08	0.31	0.82	-0.29	0

3.14	0	0	0	0	0	0	0	0
0	3.3	0	0	0	0	0	0	0
0	0	2.27	0	0	0	0	0	0
0	0	0	1.49	0	0	0	0	0
0	0	0	0	1.19	0	0	0	0
0	0	0	0	0	0.98	0	0	0
0	0	0	0	0	0	0.71	0	0
0	0	0	0	0	0	0	0.43	0
0	0	0	0	0	0	0	0	0

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.43	0.05	0.01	0.54	0	0.37	0.01	0.63	0
0	0.02	0.65	-0.01	0.59	0	0.09	-0.01	0.47
0.03	-0.7	-0.04	0.06	0.1	-0.16	-0.67	0.09	0.09
-0.22	-0.24	0.15	0.28	-0.11	-0.68	0.44	0.33	-0.13
0.69	-0.32	0.22	-0.49	-0.12	-0.03	0.27	-0.02	-0.19
-0.27	-0.34	0.44	0.29	-0.13	0.45	0.12	-0.31	-0.45
-0.03	0.3	0.14	-0.17	0.44	-0.15	-0.3	0.24	-0.71
-0.3	0.12	0.4	-0.39	-0.53	0.12	-0.23	0.46	0.13
0.35	0.35	0.35	0.35	-0.35	-0.35	-0.35	-0.35	0

Other methods

LDA

- Slower
- More popular
- A prior knowledge about topic distribution

Other methods

PLSA

- Fast
- More "natural"coefficients