Topic Modeling

15.10.2020

Text classification, text tagging, text categorization, rubrication

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

Techniques:

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

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 - topic modeling techniques based on SVD decomposition.

- Easy to understand
- Easy to specify
- Fast

Pipeline:

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

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

Decomposition notation:

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

LDA

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

PLSA

- Fast
- More "natural"coefficients

Links

- Introduction to Topic Modeling
- LDA
- How to Compare LDA Models