Topic Modeling using Latent Dirichlet Allocation

...

By :
Shivam Sarkhar (20172019)
Rajesh Kumar Dansena (20163005)
Vineet (20172015)
Harshad (20173081)

Mentor
Krishna Chaitanya
Prof.
Naresh Manwani

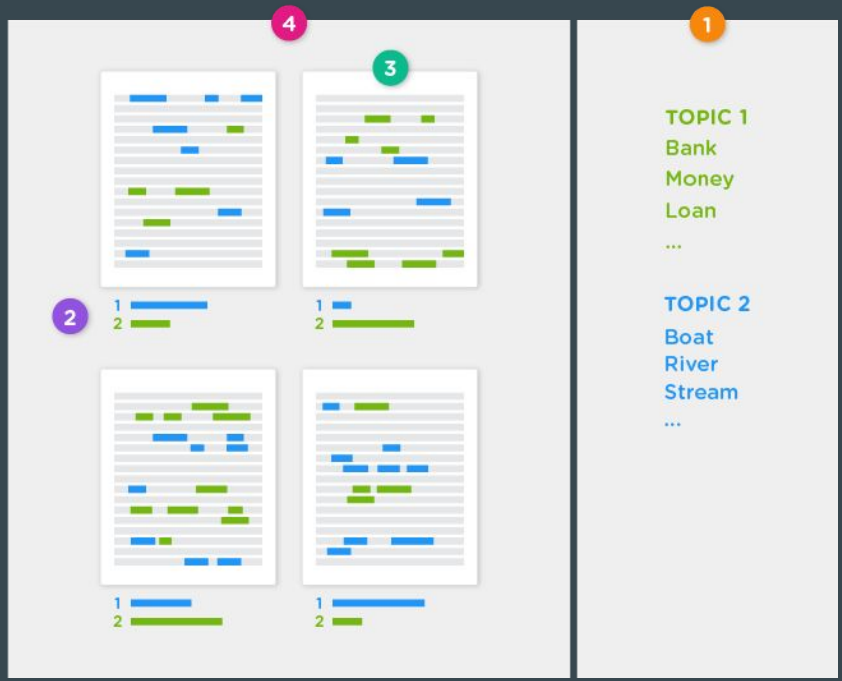
Project about(?)

Topic Modeling in a Corpus

- Context detection, sentiment analysis, news clustering

Assumptions :

- Document : Distribution of latent Topics
- Topics : Distribution of related words



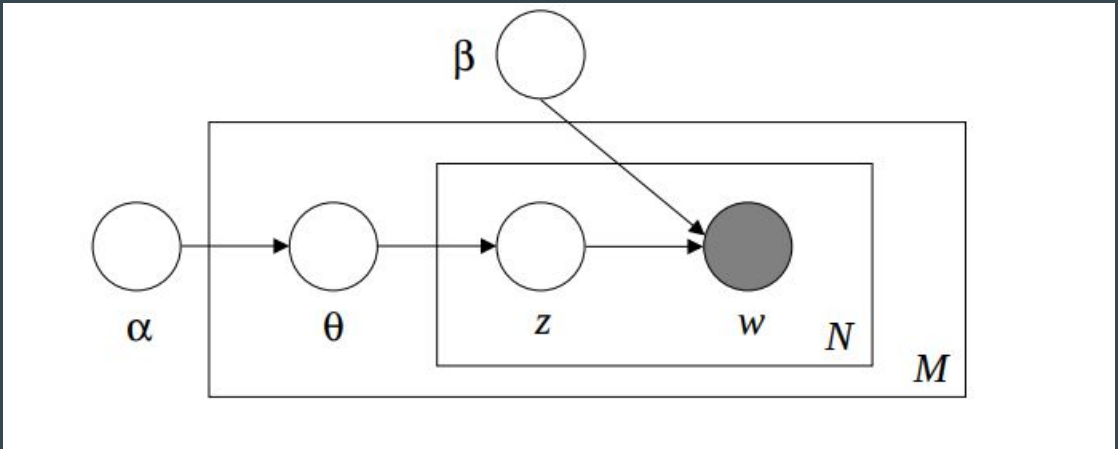
Generative Model

Story behind Document Generation :

LDA assumes the following generative process for each document **w** in a corpus D:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Plate Notation



Corpus Level Parameters : Alpha, Beta
Document Level Parameters : Theta
Word level Parameters : z, w
M : Number of Documents in corpus
N : Number of words in a document

Algorithm

1) Randomly assign each word in all the documents one of the K topics. This is the initial word-topic assignment, and will be updated later.

2) Iterate over each document d

3) For every word w in document d we assign it a topic according to the initial word-topic assignment

4) Now we reassign word w in document d to topic k according to the probability above

$$\Pr(z_{(d,n)} = k \mid z_{-(d,n)}, V, \alpha, \beta)$$

5) Use the word-topic assignment in 4) as the new initial word-topic assignment (updating step 1)

6) Repeat steps 2-5 several times (say 100 times).

Implementation

- Preprocessing Dataset
  - Punctuation Removal
  - Removing numbers and symbols
  - Removing Stop words
  - Stemming
- Generating vocabulary
- Assigning IDs to words to improve comparison
- Initialization:
  - Number of topics
  - Random topics to all words in all the documents
  - Word-Topic matrix
  - Document-Topic matrix
  - Hyperparameters

Implementation

- Algorithm
- Gibbs sampling
- Updating Theta and Phi
- Output the most probable words in each topic

Results

Gibbs Sampling

$$p(z_{a,b} \mid z_{-(a,b)}, w, \alpha, \beta) = \frac{p(z_{a,b}, z_{-(a,b)}, w \mid \alpha, \beta)}{p(z_{-(a,b)}, w \mid \alpha, \beta)}$$

$$p(z_{a,b} \mid z_{-(a,b)}, w, \alpha, \beta) = \frac{n_{d,k} + \alpha_k}{\sum_{i=1}^{k} (n_{d,i} + \alpha_i)} \times \frac{n_{k,w_{d,n}} + \beta_k}{\sum_{i=1}^{k} (n_{k,i} + \beta_i)}$$

Parameter Estimation

- Dirichlet Prior Parameters α and $\boldsymbol{\beta}$
- Number of Topics $\mathbf{K}$

Applications

- Collaborative filtering
- Spam detection
- Music industry.
- Image

Limitations

- Fixed K (the number of topics is fixed and must be known ahead of time)
- Uncorrelated topics (Dirichlet topic distribution cannot capture correlations)
- Non-hierarchical (in data-limited regimes hierarchical models allow sharing of data)
- Static (no evolution of topics over time)
- Bag of words (assumes words are exchangeable, sentence structure is not modeled)
- Unsupervised (sometimes weak supervision is desirable, e.g. in sentiment analysis)

Improvements

- We have tagged parts of speech to model.
- We should stop common stop words (the, a, it, etc),should also make sure that you don't allow very high frequency words to overpower the rest of the corpus & very infrequent words either.
- LDA Limitation : Fixed K (the number of topics is fixed and must be known ahead of time) ; Solution : HDP-LDA

HDP also uses a Dirichlet process to capture the uncertainty in the number of topics. So a common base distribution is selected which represents the countably-infinite set of possible topics for the corpus, and then the finite distribution of topics for each document is sampled from this base distribution.

Improvements

- LDA Limitation : Uncorrelated topics ; Solution : (CTM)

A topic model for text or other discrete data that models correlation between the occurrence of different topics in a document. Rather than use a Dirichlet, the CTM draws a real valued random vector from a multivariate Gaussian and then maps it to the simplex to obtain a multinomial parameter.

This is the defining characteristic of the logistic Normal distribution.The covariance of the Gaussian induces dependencies between the components of the transformed random simplicial vector, allowing for a general pattern of variability between its components.

Improvements

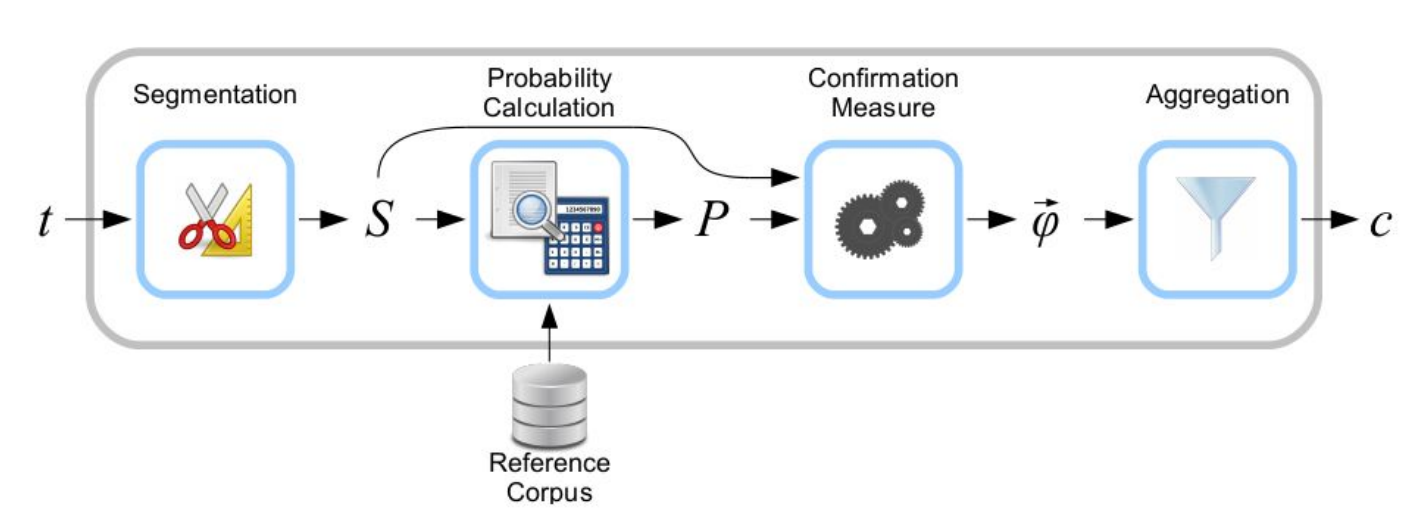
- LDA Limitation : Bag of words (assumes words are exchangeable, sentence structure is not modeled) ; Solution : TF-IDF

In TF-IDF, it is the term weight which is represented in Vector space model.Thus entire document is a feature vector. which points to a point in vector space such that there is an axis for every term in our bag.

Improvements

K-Value Analysis using topic coherence

Reference : http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf



References

- Research Paper : http://ai.stanford.edu/~ang/papers/jair03-lda.pdf
- Understanding of the Paper : https://www.youtube.com/watch?v=DWJYZq_fQ2A
- Generative process: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- Gibbs sampling: https://lingpipe.files.wordpress.com/2010/07/lda3.pdf
- https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/