Text Mining

Emanuele Guidotti

2019/2020

Topic Modeling

Topic modeling is an unsupervised machine learning technique that analyzes text data to determine **word clusters** for a set of documents. The clusters of similar words are called **topics**.

Given the text collection and the number of topics, we want to infer the actual topics and the topic distribution for each document.

Topics are just word distributions. Interpreting topics, making sense of words and generating labels is subjective.

Latent Dirichlet Allocation (LDA) is one of the most popular topic modeling methods.

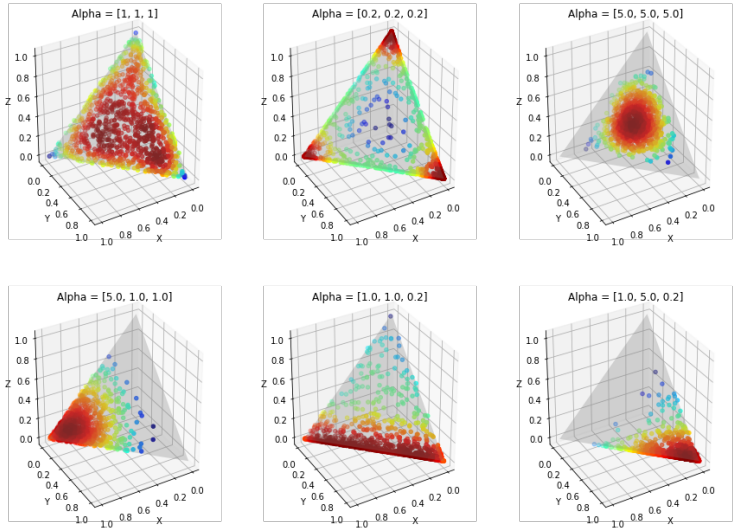
Dirichlet Distribution

Dirichlet Distribution

The Dirichlet distribution is a family of continuous multivariate probability distributions parameterized by a vector $\boldsymbol{\alpha}$ of positive reals.

The support of the Dirichlet distribution is the set of K-dimensional vectors $\boldsymbol{x}$ whose entries are real numbers in the interval $(0, 1)$; furthermore, $\sum_i x_i = 1$, i.e. the sum of the coordinates is 1. These can be viewed as the probabilities of a K-way categorical event.

Another way to express this is that **the domain of the Dirichlet distribution is itself a set of probability distributions**, specifically the set of K-dimensional discrete distributions.



Large α values push the distribution to the middle of the triangle, where smaller α values push the distribution to the corners.

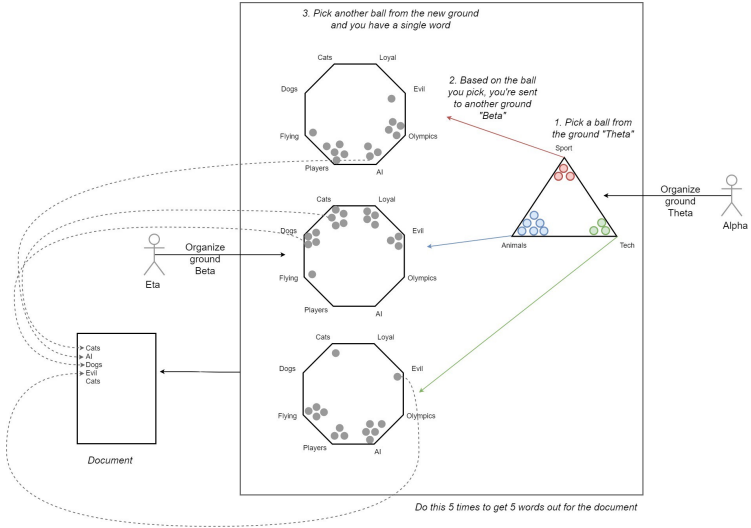
Latent Dirichlet Allocation

Latent Dirichlet Allocation

LDA assumes the following generative process for a corpus consisting of K topics and M documents each of length N_i:

- Choose $\boldsymbol{\theta_i} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where $i \in \{1, \ldots, M\}$
- Choose $\boldsymbol{\beta_k} \sim \text{Dirichlet}(\boldsymbol{\eta})$, where $k \in \{1, \ldots, K\}$
- For each of the word positions i, j, where $i \in \{1, \ldots, M\}$, and $j \in \{1, \ldots, N_i\}$
  - Choose a topic $z_{i,j} \sim \text{Multinomial}(\boldsymbol{\theta_i})$
  - Choose a word $w_{i,j} \sim \text{Multinomial}(\boldsymbol{\beta_{z_{i,j}}})$

Each document can be described by a distribution of topics and each topic can be described by a distribution of words.



How a document is generated. First α (Alpha) organise the ground θ (Theta). Then you pick a ball from θ. Based on what you pick, you are sent to ground β (Beta). β is organised by η (Eta). Now you pick a word from β and put it into the document.

What do we need to learn?

We need to learn:

- $\boldsymbol{\theta_i} = \theta_{i,j}$, probability of document i containing topic j
- $\boldsymbol{\beta_i} = \beta_{i,j}$, probability of topic i containing word j

Given the parameters:

- α, controlling the document-topic distribution
- η, controlling the topic-word distribution

```
# class
gensim.models.ldamodel.LdaModel
```

Take Home Concepts

Take Home Concepts

- Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections.
- Topics are just word distributions. Interpreting topics is subjective.
- Latent Dirichlet Allocation is a generative model used extensively for modeling large text corpora.
- LDA can also be used as a feature selection technique for text classification and other tasks.

References

References

- https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158
  - read (11 mins)