Topic Modelling

Princeton AI4ALL

Produced by Princeton University for AI4ALL

Topic Modelling

Statistical model for discovering abstract “topics” that occur in a collection of documents

Topic Modelling

Statistical model for discovering abstract “topics” that occur in a collection of documents

Intuitively, given that a document is about a particular topic, one would expect certain words to appear more frequently

- Example: “dog” and “bone” appear more often in documents about dogs

Topic Modelling

Statistical model for discovering abstract “topics” that occur in a collection of documents

Intuitively, given that a document is about a particular topic, one would expect certain words to appear more frequently

- Example: “dog” and “bone” appear more often in documents about dogs

Topic Models allow us to capture this intuition in a mathematical framework

Topic Modelling

Statistical model for discovering abstract "topics" that occur in a collection of documents

Intuitively, given that a document is about a particular topic, one would expect certain words to appear more frequently

- Example: "dog" and "bone" appear more often in documents about dogs

Topic Models allow us to capture this intuition in a mathematical framework

Examples include:

- Latent Dirichlet Allocation (LDA)
- Latent Semantic Indexing (LSI)
- Pachinko Allocation

Latent Dirichlet Allocation

Topics do not need to be specified a priori

Number of topics **<u>does</u>** need to be specified a priori

A single word can belong to several topics

Output of LDA is two matrices

- Topics vs Words: For each topic, a probability distribution over words is given
- Topics vs Documents: For each document, a probability distribution over documents is given

Latent Dirichlet Allocation

In this example number of topics is 3

Document 0	🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞
Document 1	🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑
Document 2	🥓 🥓 🥓 🥓 🥓 🥓 🥓 🥓 🥓 🥓
Document 3	🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑

Topics vs word

	Topic 0	Topic 1	Topic 2
🍞	0.000	0.000	0.999
🥑	0.000	0.999	0.000
🥓	0.999	0.000	0.000

Topics vs document

	Topic 0	Topic 1	Topic 2
Document 0	0.030	0.030	0.939
Document 1	0.030	0.939	0.030
Document 2	0.939	0.030	0.030
Document 3	0.333	0.333	0.333

Latent Dirichlet Allocation

In this example number of topics is 2

Document 0	🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞
Document 1	🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑
Document 2	🥓 🥓 🥓 🥓 🥓 🥓 🥓 🥓 🥓 🥓
Document 3	🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🍞 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑 🥑

Topics vs word

	Topic 0	Topic 1
🍞	0.000	0.500
🥑	0.999	0.000
🥓	0.000	0.500

Topics vs document

	Topic 0	Topic 1
Document 0	0.031	0.969
Document 1	0.969	0.031
Document 2	0.031	0.969
Document 3	0.337	0.663

Latent Dirichlet Allocation

At a high level, LDA works as follows:

1. Initially, assign each word to a random topic.

2. Iteratively, for each word W re-assign W to new topic T based on
   - the topics of nearby words
   - the probability of seeing W in T

Latent Dirichlet Allocation

At a high level, LDA works as follows:

1. Initially, assign each word to a random topic.

2. Iteratively, for each word W re-assign W to new topic T based on
    - the topics of nearby words
    - the probability of seeing W in T

Why is LDA useful for the FNC challenge?

- Can provide more information for our SVM
- Used by FNC high scorers (in addition to other features)
    - FNC high scorers used 300 topics

Latent Dirichlet Allocation

To apply LDA:

- Use gensim package to obtain topic vectors for a document and a headline.
- Then apply cosine similarity to the topic vectors
- Feed result as input into SVM (along with other features)

Topics vs document

	Topic 0	Topic 1
Document 0	0.031	0.969
Document 1	0.969	0.031
Document 2	0.031	0.969
Document 3	0.337	0.663

Calculate cosine similarity

Resources

LDA simulator
https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

LDA walkthroughs
https://www.kaggle.com/ktattan/lda-and-document-similarity

https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

Detailed explanation
https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd

Original paper
http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Using LDA to calculate similarity
https://stats.stackexchange.com/questions/271359/using-lda-to-calculate-similarity/271368