LDA Steps

Reference : https://stackoverflow.com/questions/10624760/latent-dirichlet-allocation-solution-example

Step1: Go through each document and randomly assign each word in the document to one of K topics (K is chosen beforehand)

Step2: This random assignment gives topic representations of all documents and word distributions of all the topics, albeit not very good ones

So, to improve upon them: For each document d, go through each word w and compute:

- p(topic t | document d): proportion of words in document d that are assigned to topic t
- p(word w| topic t): proportion of assignments to topic t, over all documents d, that come from word w

Step3: Reassign word w a new topic t', where we choose topic t' with probability

- p(topic t' | document d) * p(word w | topic t')

This generative model predicts the probability that topic t' generated word w. we will iterate this last step multiple times for each document in the corpus to get steady-state.

Solved calculation

Let's say you have two documents.

Doc i: “T**he bank called about the money.**”

Doc ii: “T**he bank said the money was approved.**”

After removing the stop words, capitalization, and punctuation.

Unique words in corpus: **bank called about money boat approved**

Randomly assign topics

	1	2	2	1	K=2 (two topics) in our case
Doc i	bank	called	about	money	

Similarly, done to each document in the corpus

	2	1	1	2
Doc ii	bank	said	money	approved

Next then,

Maintain the global statistics

word\topic	Topic 1	Topic 2
bank	1	1
called	0	1
about	0	1
money	2	0
said	1	0
approved	0	1

Doc i

1	2	2	1
bank	called	about	money

	Topic 1	Topic 2
Doc i	2	2

Total counts from all **docs**

After then, we will randomly select a word from doc i (word **bank** with topic assignment **1**) and we will remove its assigned topic and we will calculate the probability for its new assignment.

Randomly re-assign topics

word\topic	Topic 1	Topic 2
bank	0 ~~1~~	1
called	0	1
about	0	1
money	2	0
said	1	0
approved	0	1

~~1~~	2	2	1
bank	called	about	money

	Topic 1	Topic 2
Doc i	1 ~~2~~	2

decrement the count

Probability of new assignment

$$P(\text{topic } t \mid \text{doc } d) = \frac{\text{Word in doc(n) with topic(K) + alpha}}{\text{Word in doc(n) with topics + Number of topics * alpha}}$$

How much doc likes topic

Smoothing parameter from Dirichlet

$$P(\text{word } w \mid \text{topic } t) = \frac{\text{Word (bank) with topic (K) + beta}}{\text{Word with topic(K) + Unique word(V) * beta}}$$

Size of vocab

How much topic likes word

For the topic **k=1**

Probability of new assignment

word\topic	Topic 1	Topic 2
bank	0	1
called	0	1
about	0	1
money	2	0
said	1	0
approved	0	1

Doc i

?	2	2	1
bank	called	about	money

	Topic 1	Topic 2
Doc i	1	2

Our hyperparameters are:

- alpha = 0.5
- beta = 0.01
- 'topics' = 2 (i,e K= 1,2)
- 'iterations' = 1.

$$P11 = \frac{\text{Word in doc(n=i) with topic(K=1) + alpha}}{\text{Word in doc(n=i) with topics + Number of topics * alpha}} = \frac{1+0.5}{3+2*0.5} = 0.375$$

$$P12 = \frac{\text{Word (bank) with topic (K=1) + beta}}{\text{Word with topic(K=1) + Unique word(V) * beta}} = \frac{0+0.01}{3+6*0.01} = 0.003268$$

For the topic **k=2**

Probability of new assignment

word\topic	Topic 1	Topic 2
bank	0	1
called	0	1
about	0	1
money	2	0
said	1	0
approved	0	1

?	2	2	1
bank	called	about	money

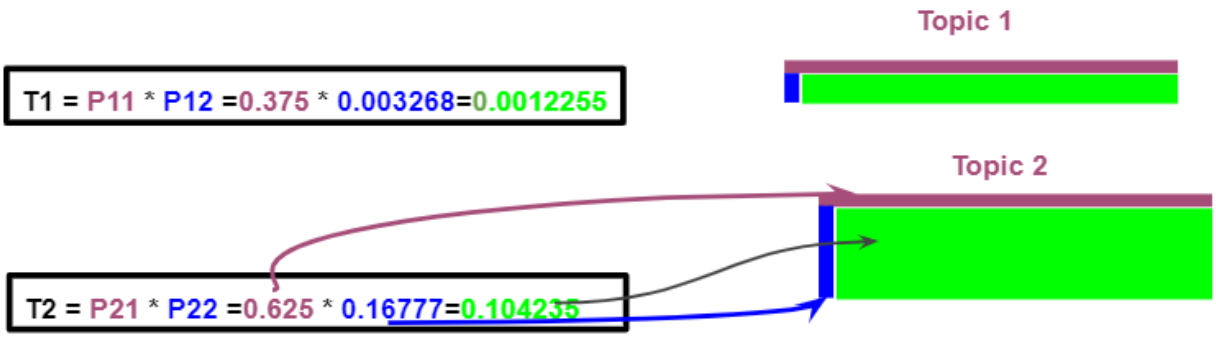
	Topic 1	Topic 2
Doc i	1	2

$$P21 = \frac{\text{Word in doc(n=i) with topic(K=2) + alpha}}{\text{Word in doc(n=i) with topics + Number of topics * alpha}} = \frac{2+0.5}{3+2*0.5} = 0.625$$

$$P22 = \frac{\text{Word (bank) with topic (K=2) + beta}}{\text{Word with topic(K=2) + Unique word(V) * beta}} = \frac{1+0.01}{4+6*0.01} = 0.16777$$

Now we will calculate the product of those two probabilities as given below:

Probability of new assignment



Good fit for both **document** and **word** for topic 2 (area is **greater**) than topic 1. So, our new assignment for word **bank** will be topic 2.

Now, we will update the count due to new assignment.

Update counts

word\topic	Topic 1	Topic 2
bank	0	~~1~~ 2
called	0	1
about	0	1
money	2	0
said	1	0
approved	0	1

2	2	2	1	1
bank	called	about	money	boat

	Topic 1	Topic 2
Doc i	1	3 ~~2~~

Increment the count based on new assignment

Now we will repeat the same step of reassignment. and iterate through each word of the whole corpus.

Iterate through all words/docs

2	?	2	1	1
bank	called	about	money	boat

word\topic	Topic 1	Topic 2
bank	1	1
called	0	~~1~~ 0
about	0	1
money	2	0
said	1	0
approved	0	1

	Topic 1	Topic 2
Doc i	1	2 ~~3~~

decrement the count