1 Implementation

In this section we present our implementation of LDA in python and discuss our results.

1.1 Library Used

- numpy
- scikit learn for pre-processing
- scipy probability distributions

1.2 Dataset Description

The dataset used is the same one used in the authors implementation. it can be found at the following link: `https://nlp.stanford.edu/software/tmt/tmt-0.4/` It consist of 2246 small documents.

1.3 Pre-processing

Tokenization of the documents: LDA use the bag of word assumption (the order of words apparitions in a document is ignored). This allows us to use a compact representation of every documents in a matrix of M x V. Each line in the matrix represent a document and the values are integer that represent the number of occurrence of a word in the vocabulary. This is similar to the one hot encoding trick that we used in the class.

In addition to tokenization, we included the following filters to remove words that have either a too low or too high frequency:

- Removed words with only 1 occurrence across the corpus.
- Removed words that occur in 95 % of the document or more.

1.4 Parameter Initialization

As stated in section xxxx, LDA with with variational EM has 4 set of parameters to estimate, each of the parameters are randomly initialized to the following value:

Parameter	Dimension	Initialization
α	[K]	np.random.gamma(shape=np.ones((K)), scale=1/K)
ϕ	[M x N[d] x K]	1 / K * np.ones((N[d], K)) (per doc)
β	[K x V]	np.random.dirichlet(np.ones(V), K)
γ	[M x K]	alpha + np.max((N[d] / K, 0.2))

We added a second term to the innitial γ to ensure that we dont encounter overflow in the first iterations of the variational inference algorithm.

1.5 Stopping Criterion

In the paper, the author doesn't state very clearly the stopping criterion used in its implementation. So we chose to stop the model once the parameter L2 norm of γ from one iteration to the next is is below a given threshold. stopping metric: in the paper

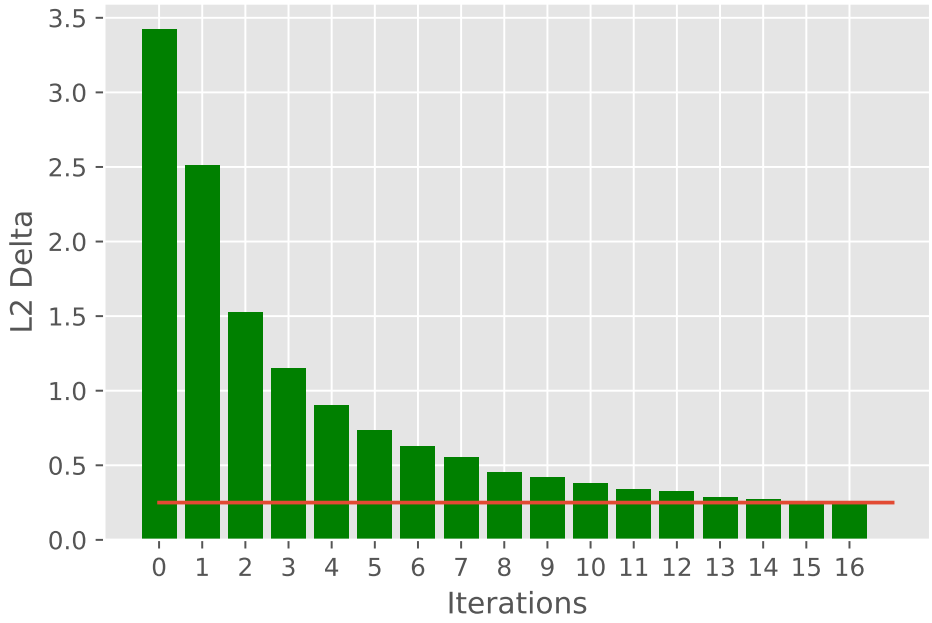

Figure 1: Change in γ parameter across EM iterations. The green bars represent the L2 change of γ parameters. The red line represent the stopping criterion used (0.25 in our case).

1.6 Results

Table 1: Topic Sample from LDA model: each column represent the top 10 words from the topic

Topic Sample				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
club	new	child	ago	american
year	percent	people	africa	states
york	yen	having	leaders	union
old	economy	report	france	told
building	rate	education	people	political
business	rates	aids	african	leaders
years	said	children	french	conference
new	prices	said	south	president
city	market	care	police	said
said	dollar	health	said	soviet