Speech and Natural Language Processing
CS60057

Mid Term Report

Submitted by:
Amit Behera
Apoorv Gawande
Shyam Swaroop

Guide - Ankan

Data Description

Dataset has been collected from the website www.theguardian.com. Features like Beautiful Soup and Selenium were implemented for scraping the website and collecting articles for 8 different categories.

The following table i.e. Table 1 shows description of articles. The type, token and sentences have been counted after stemming has been performed on the articles.

Table 1.

Category	Types	Tokens	Sentences
Business	2461	10504	728
Environment	3098	13558	902
Opinion	3757	13444	1209
Politics	2885	15250	966
Sports	3361	12833	1128
Technology	2880	12852	1039
World	3146	10695	841

These articles along with their comments dataset were preprocessed to remove unwanted new line characters and spaces and empty comments.

All the articles were then clustered together to form a single document so that we could apply topic modeling. LDA was implemented on this single document in each category to form around 20 topics. These topics consisted of words along with a certain weight.

Overall Approach

All documents have been clustered together. Let's call this cluster **C**.

LDA will be run on **C** to get **N** topics. Set of Topics : **T**. Each element of **T** i.e. **t** is a vector of dimension of Vocabulary.

$$\mathbf{t : [\alpha_1, \alpha_2, \alpha_3, \ldots\ldots \alpha_v]}$$

$$\mathbf{\alpha_1, \alpha_2, \alpha_3, \ldots\ldots \alpha_v \epsilon (0,1)}$$

$$\mathbf{\Sigma \alpha = 1}$$

$$\mathbf{v=dim(V)}$$

1. Cosine Similarity can be calculated for each sentence in **C**.
2. $\mathbf{w_1, w_2, w_3, \ldots\ldots w_v \epsilon [0,1]}$

Code Layout

τ <- threshold for similarity match
for i in length(Sentences):
    $S_i = [w_1, w_2, w_3 \ldots\ldots, w_v]$
    [Sim_value , Topic_Num] = max $_{tj}$ (CosSim (S_i , t_j))
    while (Sim_value < τ) :
        S_i = Club(S_i , S_{i+1})
        [Sim_value , Topic_Num] = max $_{tj}$ (CosSim (S_i , t_j))
return (Si, Sim_value, Topic_Num)

Note : Above code has an inductive bias of producing smallest possible paragraphs that can be labelled to a topic.

3. Doc2vec is to be applied on the set of **S** and Comments **U.**
4. For each element in **U** a pair (**u** , **s**) is formed where **s** is the element of **S** closest to **u**.
5. For each element in **U** a pair **(u , t)** is formed where **t** is topic of **s** in the pair (**u** , **s**).

Thus we will have pairs (**u** , **s** ,**t**) for every element **u** in **U**.

Note:

1. In this documentation a group of sentences that can labeled to a topic has been referred to as **paragraph**.

Topic Extraction using LDA

Topic Number	Top 10 words and their weight

Topic 1	('bank', 0.024896913315671509) ('said', 0.016474019914249311) ('market', 0.015500265463135167) ('year', 0.011806360375281049) ('account', 0.010044370267534911) ('need', 0.0095202761910126842) ('fine', 0.0092904907359189614) ('financi', 0.0091598530932882546) ('custom', 0.0087312801173483143) ('peopl', 0.0076499862536429592)
Topic 2	('industri', 0.026082786977670122) ('fall', 0.018301642026033849) ('loan', 0.018029141228842051) ('10', 0.017029001819354208) ('ecb', 0.016194265622394868) ('manufactur', 0.014280286661599341) ('seen', 0.013895960213474836) ('product', 0.013827340474048122) ('declin', 0.012972332779967609) ('relat', 0.012519346546732943)
Topic 3	('would', 0.020150604874483106) ('said', 0.017053884787808143) ('greec', 0.016090087080465023) ('greek', 0.01239440119436978) ('govern', 0.012104470073018169) ('debt', 0.011697054847073354) ('vote', 0.01033895660866742) ('econom', 0.0093163467596133684) ('minist', 0.008402908360267608) ('uk', 0.0081929170521552724)
Topic 4	('bank', 0.050115399849883994) ('overdraft', 0.018437097900949419) ('cma', 0.014367902913676588) ('trader', 0.011847599856770433) ('publish', 0.010898065625137196) ('investig', 0.010869288354940174 ('currenc', 0.010554810029753838)

	('switch', 0.01029870606970117 ('treasuri', 0.009647127032079237) ('1bn', 0.0089476620372932122)
Topic 5	('greec', 0.023064489878254327) ('euro', 0.022765209927016673) ('risk', 0.015223832619917362) ('could', 0.01454636440265371) ('merkel', 0.01301743635189586) ('eurozon', 0.012652057701098214) ('referendum', 0.012403349737871447) ('week', 0.011408057283243317) ('tsipra', 0.010996543081327503) ('effect', 0.010193236271124312)
Topic 6	('push', 0.02635050081241436) ('outcom', 0.018490872544948359) ('uncertainti',0.013748313091625334) ('hope', 0.010140732989822539) ('produc', 0.0094988694693485237) ('drop', 0.0094009780587256883), ('quarter', 0.0087023814650861436) ('suffer', 0.0077282620360164724) ('march', 0.0073667265085934162) ('analyst', 0.0073023142403612845)
Topic 7	('meanwhil', 0.00040633890294043321) ('produc', 0.00040633889284791405) ('realiti', 0.00040633889257322715) ('live', 0.0004063388917524365) ('effort', 0.00040633889094406149) ('hope', 0.0004063388907875849 3) ('outcom', 0.00040633889058453766) ('appear', 0.00040633888931878583) ('receipt', 0.00040633888926829703) ('plummet', 0.00040633888926829703)
Topic 8	('chang', 0.022671967301661506) ('world', 0.01909085701990066) ('countri', 0.016582328688333835) ('global', 0.015934818046669862)

	('bank', 0.015200880212409056) ('said', 0.011913527424052616) ('import', 0.011713815210514095) ('climat', 0.011048584812020248) ('citi', 0.010436019394560852) ('disast', 0.01005300369841761)
Topic 9	('mcdonnel', 0.018121096870731494), ('strategi', 0.015131702855912242), ('debat', 0.014743889547410259), ('asic', 0.0098089117683224803), ('cut', 0.0086590279345162783), ('away', 0.0080846693462426518), ('morrison', 0.0078840894908849615), ('turnbul', 0.0078840894908849615), ('wide', 0.0075621188341404508), ('john', 0.0072469505138068914)
Topic 10	('retail', 0.031163029762717745), ('labour', 0.02494080807948466), ('austin', 0.018853544640201142), ('reed', 0.018853544640201142), ('administr', 0.01723399059347501), ('bhs', 0.01610843064630367), ('compani', 0.014282333699535077), ('green', 0.012869996803957416), ('stori', 0.012418129447765646), ('protect', 0.0088924062732756316)

How well the model is?

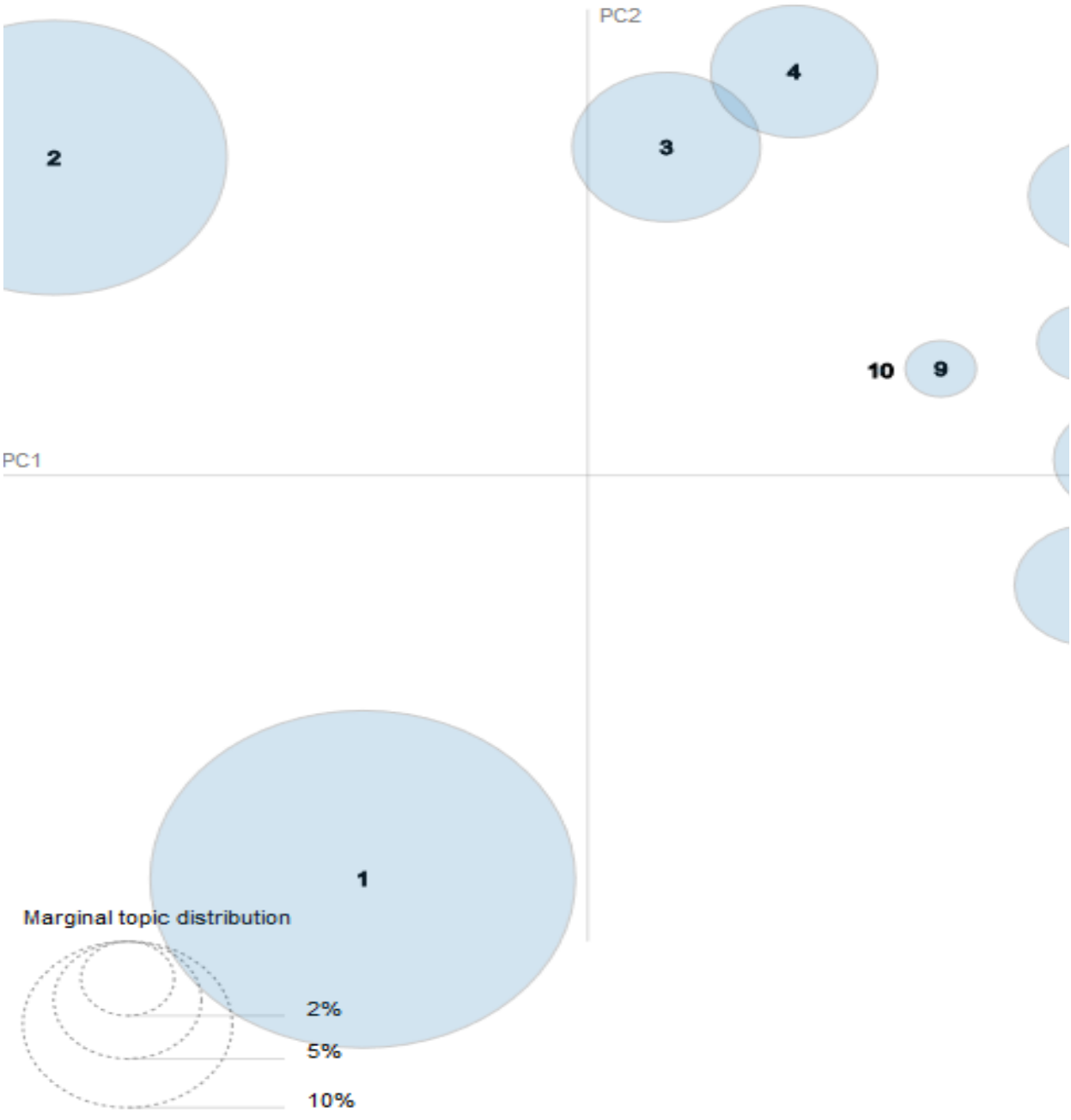


Fig 1.

Fig. 1 shows the result of Linear Decomposition Analysis. Multiple LDA models with different parameters such as chunk size etc. has been run. The best model has been selected on the basis of least overlap of topics. Least overlap in principal component directions assures that there is least overlap in higher dimensional too. The circles become small very rapidly due to small size of data which leads to less variation in different topics. Topic 10 has reduced to a point.

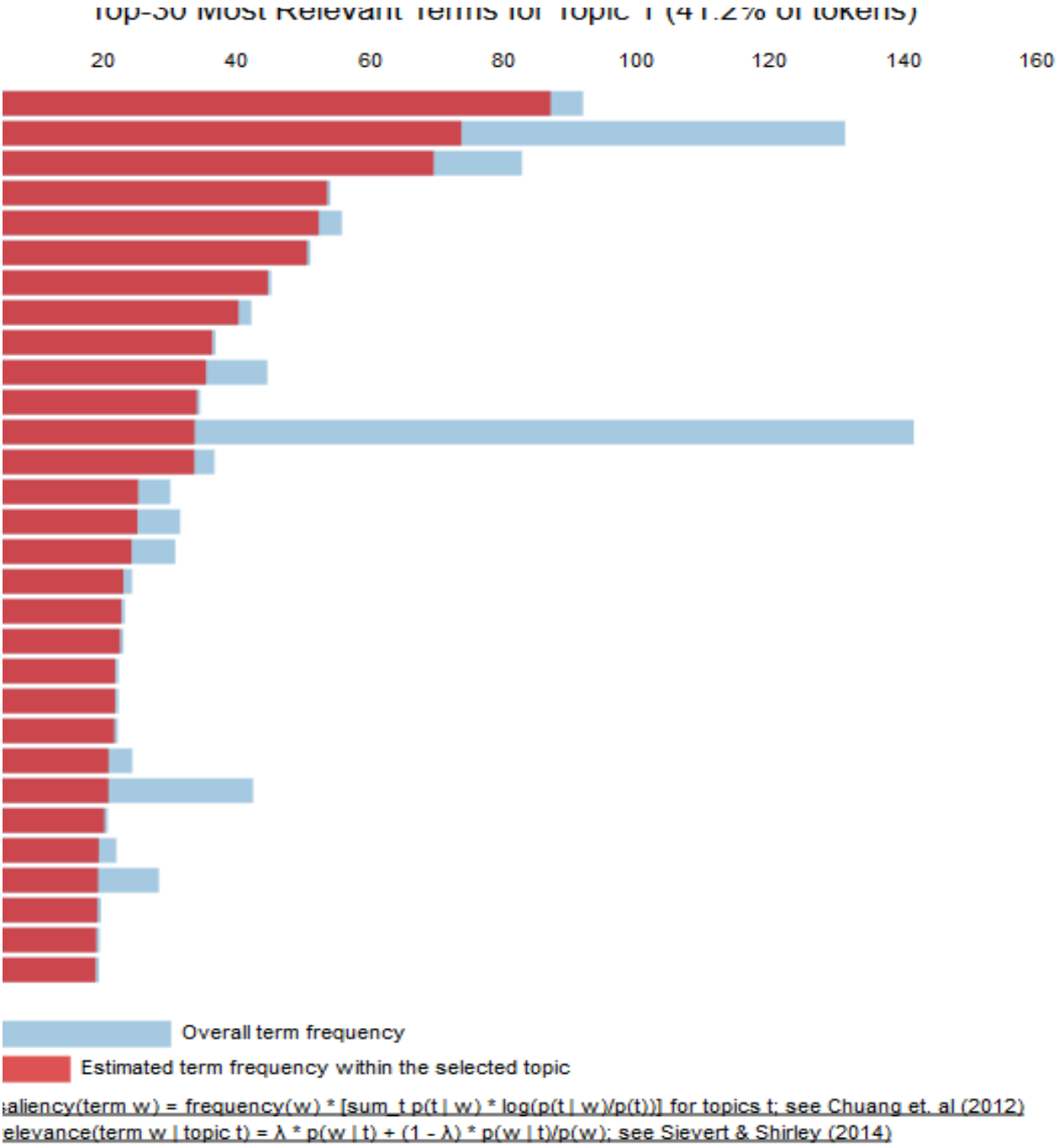
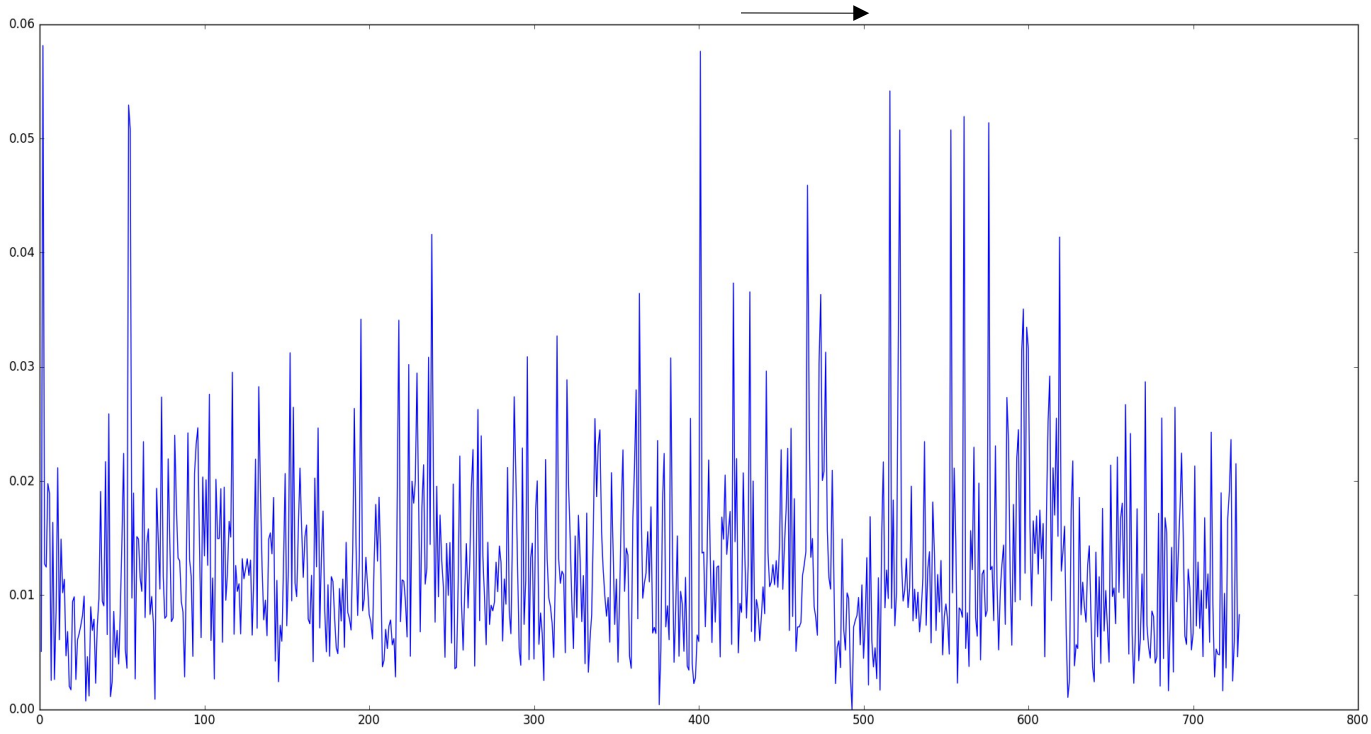


Fig. 2.

Fig. 2. shows top relevant words for topic 1. Relevance of a word is different from its weight in the topic.

How well sentences get labeled WITHOUT clubbing them into paragraph?



Sentence

Graph 1.

The y-axis shows max. similarity score of a sentence with any of the 10 topics an x-axis shows the sentence-id.

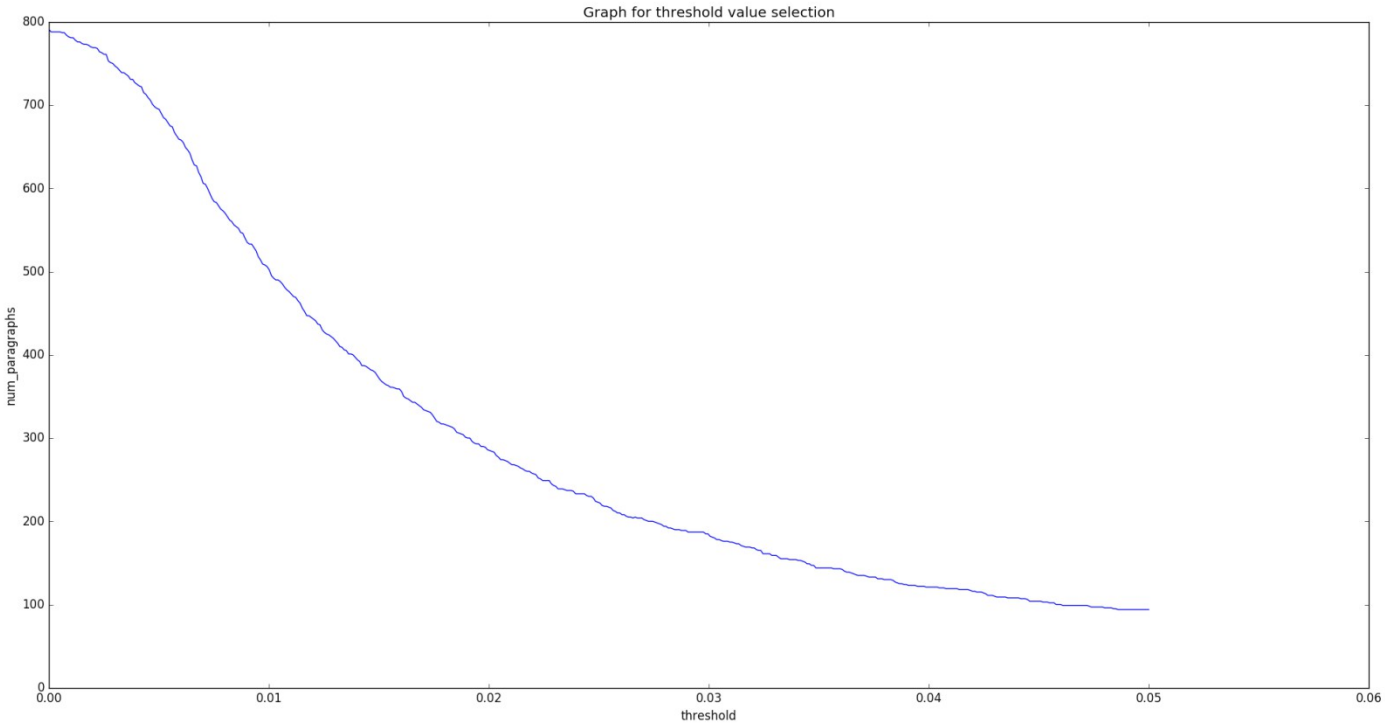
Average Similarity = 0.0128

Std. deviation = 0.0086

We have used cosine similarity as a metric to formulate similarity between sentence and topics.

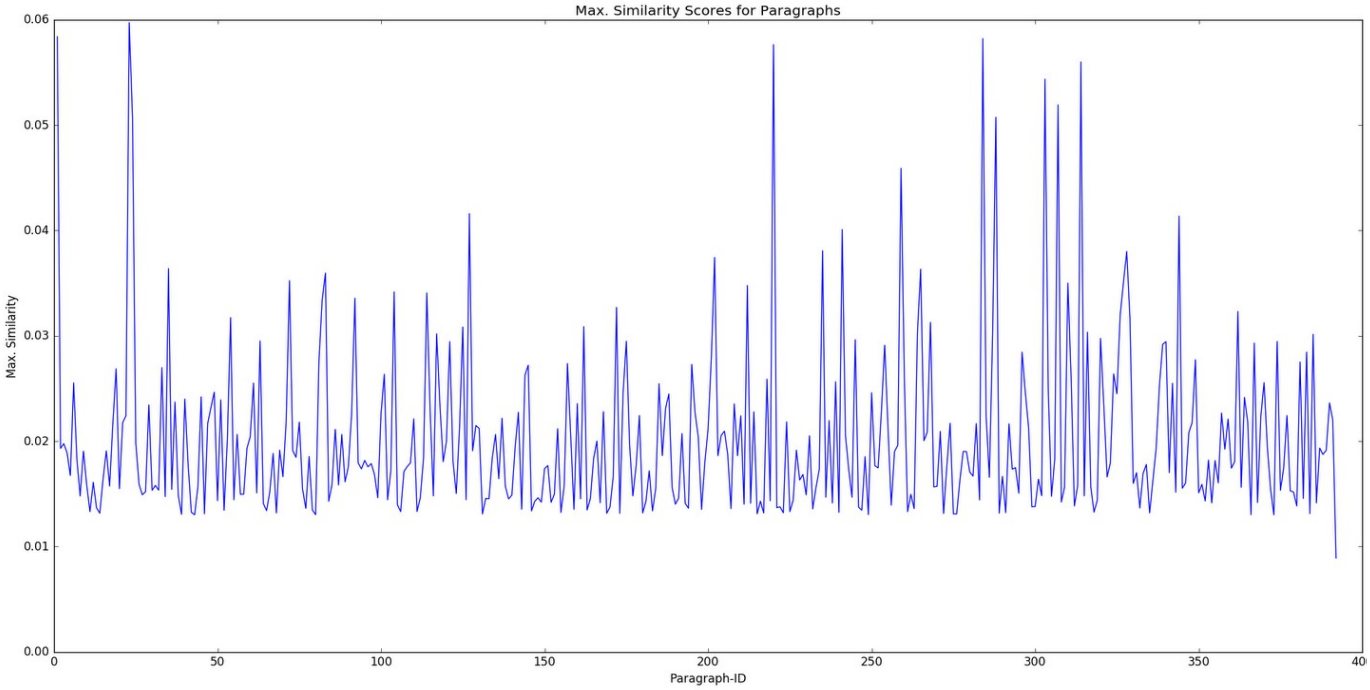
Threshold (τ) Selection

Threshold has been selected depending upon the average size of a paragraph we want. The Graph 2. below shows variation of number of paragraphs as a function of threshold for Business Category. For our project, we kept an average size of 2 sentences per paragraph, which gives threshold value of about 0.013 .



Graph 2.

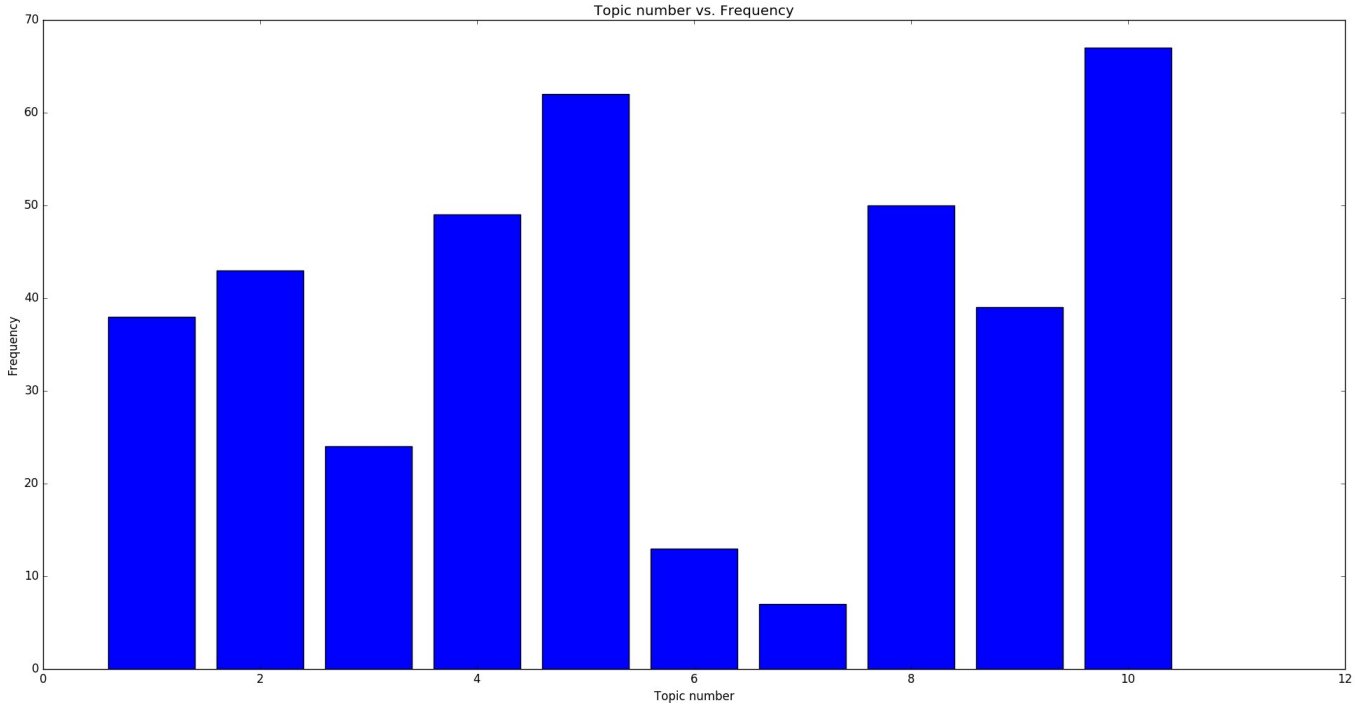
How well sentences get labeled AFTER clubbing them into paragraph?



Graph 3.

Average Similarity = 0.0204

Std. deviation = 0.0082



Graph 4. Topic number vs. Frequency