Department of Computer Science and Engineering
Spring 2012

Thesis Report on
Application of data mining identifying topics at the document level

---

Supervisor: Abu Mohammad Hammad Ali

Group Members

Marifa Farzin Reza – 08101013
Rizwana Matin – 08101012

Submission date - 12/04/12

Supervisor's signature: _____

Table of content:

Application of Data Mining Identifying Topics at the Document Level

Rizwana Matin (08101012)
BRAC University, Dhaka
rizwana.matin@gmail.com

Marifa Farzin Reza (08101013)
BRAC University, Dhaka
marifa_farzin@hotmail.com

Supervised by, Abu Mohammad Hammad Ali

Abstract

Data mining techniques are very popular in modern days and are used in NLP (Natural Language Processing). One of the techniques like clustering items to groups has been used way back. This technique is applied to find different topics for natural documents. In our thesis we aim to replicate some of these results and empirically verify this measure to identify hypothetical topic boundaries.

1. Introduction:

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [14].

Besides analysis of large databases, data mining algorithms have also been used in Natural Language Processing (NLP) to deal with large collections of text and learn predictive patterns. Such patterns have been used to solve NLP problems like part-of-speech tagging, opinion identification and topic clustering, among others [15]. It is this last problem that we aimed to study in our work. To be more specific, we wanted to explore different data mining

algorithms, find ones that can be used for the task of finding the different topic clusters from raw text corpora, and compare the performance of multiple algorithms such as unigram, bi-gram, and named entity extraction (proper noun extraction, both proper noun and common noun extraction) on this task.

2. Problem:

We started our research by reading extensively on previous work that has been done in applying data mining techniques to the task of finding topic clusters, or performing topic segmentation. Our original aim was to identify topic boundaries in the document level. We tried to do this using unigram, bigram and named entity extractions.

3. Previous work:

The first paper we read was [1]. This work aims to identify recurring topics from articles in a text corpus. They propose to do this by first identifying key entities in the articles using NLP methods like Named Entity Extraction, and then using data mining techniques to identify groups of related items. Their basic idea is to treat each document as a collection of named entities, which then allows them to model the topic clustering task as a market basket problem, which is a well-known data mining approach to group similar items. They conduct evaluation of their work in terms of whether the system is able to find topic clusters that would make sense to a human reader. The test corpus consisted of around 60,000 articles. They use a single metric that combines several probability measures to measure their performance. Alembic tagged the entire 144MB TDT2 corpus in less than 21 hours by named entity tagging. The co-reference mapping procedure required 6 hours and 49 minutes. Computing frequent item sets took 76 minutes. Hyper-graph clustering the TDT2 data took just under 5 minutes. TFIDF based merging of clusters took 67 minutes. While the quantitative scores reveal a very high miss probability, the authors claim that the system output is comprehensible to, and assists, human users.

In [2], the authors attempted to use statistical information to group speech data based on similarity. They used n-gram language modeling techniques and a tree-based clustering algorithm to generate a hierarchical structure of the corpus, and then used this to search for similar material in other corpora. Their evaluation results showed substantially good performance. Clustering for text was pretty good comparing to speech clustering. The average cluster purity for text was 97.3%, where for speech clustering average purity score was 61.4%. Though there were many well clustered topic groups. The automatic clustering on text was found to be very good for all tokenization, with a figure-of-merit (FOM) of 88.7% for word labels and 84.1% for 4-grams of phones.

Token	N-gram size	Clustering on text	Clustering on speech
Word	1	88.7%	70.2%

Word	2	79.0%	61.8%
Phones-word	2	86.4%	68.5%
Phones-word	4	84.1%	69.1%
Phone	2	-	54.4%
Phone	4	-	53.3%

Table: nearest neighbor classification %FOM scores for different tokenization

For a tokenization from a straight phone recognizer, performance was basically chance with 54.4% FOM for 2-grams monophone and 53.3% for 4-grams monophone labels. On that experiment the performance was very robust to insertion errors. Even with 0% accuracy, the FOM was over 70%.

The authors in [3] focused on dividing multiparty dialogue to different segments and then automatically label these segments. They apply algorithms that have been used to identify topic boundaries and then use conditional models to assign labels. Lexical cohesion means when a group of words is lexically cohesive, all of the words are semantically related; for example, when they all concern the same topic [10]. Cohesion is the grammatical and lexical relationship within a text or sentence [11]. Cohesion can be defined as the links that hold a text together and give it meaning. It is related to the broader concept of coherence. The authors compared two segmentation approaches. First, an unsupervised lexical cohesion-based algorithms using solely lexical cohesion information and then second, a supervised classification approach that trains decision trees on a combination of lexical cohesion and conversational features. Lexical cohesion-based algorithms showed that a major shift is likely to occur where there are strong term repetitions start and end. It worked with two adjacent analysis windows which had a fixed size that was empirically determined. For each boundary, lexical cohesion-based algorithms calculated a lexical cohesion score by computing the cosine similarity at the transition between the two windows. In supervised approach each potential topic boundary is potential topic boundary is labeled as either boundary (POS) or non-boundary (NEG). They wanted to train decision trees to learn the most predictive combinations of features that can characterize topic boundaries.

Lastly, in [4] we look at another work on finding topic boundaries from a document using a set of new features and novel algorithms. They also devote some attention to the possible application of knowing such boundaries, something we are not interested in just yet. They treat the segmentation task as labeling problem – given a document and set of potential boundary locations, they label each as either being a topic boundary or not. The basic idea of their work is to use statistical model to weigh evidence of a diverse set of clues, instead of using these to form hard

and fast rules. They use some common features but also propose the use of some new ones, including a few domain-specific features. The features they used are given below:

a. Domain-specific Cue Phrase
b. Word Bigram Frequency
c. Repetition of Named Entities
d. Pronoun usage

In addition to new features, they also evaluate two different algorithms and compare these. Evaluation results show that both their models perform significantly better than both the baselines and previously existing algorithms.

4. Our chosen approach:

4.1 Paragraph level:

At first we tried to find out topic boundaries in sentence level. For that we used unigram and bigram. But we didn't get satisfactory results using those algorithms in sentence level and sentence level proved to be trickier. Because, for unigram and bigram, we noticed that often human agreed the given texts as same topic, when our threshold value was pretty low and it happened very frequently. That's why we couldn't get fine tuning. So we revised our plan and next we decided to move to paragraph level. We used "proper noun extraction" and "proper noun and common noun extraction" to find out topic boundaries. At first we have found out proper noun from each paragraph and then found out overlapping percentage of those proper nouns. For example, suppose there are five proper nouns from first paragraph and ten proper nouns from second paragraph. If the first five proper nouns do not match at all with the ten proper nouns of second paragraph, then the overlapping rate is 0% and if the first five proper noun of first paragraph matches with the second paragraphs then overlapping rate is (5/10*100=50%) Proper nouns of first paragraph are matched with second paragraph to see how many words overlapped & calculating that, we got our result. For example, if our overlapping percentage is high, the topics are similar but if it is below the threshold value which we aim to find empirically, we get a topic boundary. Threshold values are not fixed values. We take a value by examining that works the best for the experiment.

4.2 Unsupervised learning:

In machine learning**,** unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning and reinforcement learning. Many methods employed in unsupervised learning are based on data mining methods used to preprocess data [8].

Considering machine (or living organism) which receives some sequence of inputs x1, x2, x3, . . ., where xt is the sensory input at time t. This input, which we will often call the data, could correspond to an image on the retina, the pixels in a camera, or a sound waveform. It could also correspond to less obviously sensory data, for example the words in a news story, or the list of items in a supermarket shopping basket. In unsupervised learning the machine simply receives inputs x1, x2, . . ., but obtains neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it doesn't get any feedback from its environment. However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise [9].

For our thesis, we followed unsupervised learning. We didn't have enough time to tag data. So we were interested to use unsupervised learning. Because, tagging a large set of data are really difficult and time consuming. Also to use supervised approach, it requires annotating the data and analysis it. Then everyone needs to agree with it and need to find out the ways for agreement, which is really very complicated approach. This is why we used unsupervised approach.

4.3 Similarity measurement: Measuring similarity between two entities is a key step for several data mining and knowledge discovery tasks. The notion of similarity for continuous data is relatively well-understood, but for categorical data, the similarity computation is not straightforward. Comparing strings and assessing their similarity is not a trivial task and there exists several contrasting approaches for defining similarity measures over sequential data. Several data-driven similarity measures have been proposed in the literature to compute the similarity between two categorical data instances but their relative performance has not been evaluated [6,7]. The key characteristic of categorical data is that the different values that a categorical attribute takes are not inherently ordered. Thus, it is not possible to directly compare two different categorical values. The simplest way to find similarity between two categorical attributes is to assign a similarity of 1 if the values are identical and a similarity of 0 if the values are not identical. For two multivariate categorical data points, the similarity between them will be directly proportional to the number of attributes in which they match. This simple measure is also known as the overlap measure in the literature. [5]

At first, we have taken a word & have matched it with the next entity to see if that has overlapped or not. If they overlaps we can say, they are similar and if they don't then, they are different. We did similarity measurement for

sentence level and paragraph level. At first we used 2 algorithms named unigram and bigram in sentence level to find out topic boundaries. For unigram we took the first word of the sentence and tested if it matched with the words of second sentence and then tried each word of the first sentence and matched with the second one. We compared our result with human annotators & fixed a threshold value. We have seen when our program gives a value larger than threshold value, both human and our program agrees that the topic is similar and when the value is less than threshold value, we get a topic boundary there. For bigram, we had to take two words of a sentence in each set and then we compared if each set consisting of two words matched with the sets of second sentences. If they matched, we could say the topic were similar and if they didn't then the topics were not. Then we used named entity extraction for proper noun extraction and proper & common noun extraction in paragraph level. For that we found out the proper nouns of all the paragraphs & checked if the proper nouns of first paragraph overlapped with the next one. When the words of first paragraph overlap with the second one, we can say both the paragraphs are talking about same topic. When there is few overlap, we get a topic boundary. We did that for proper and common both the nouns together as well. We also checked our result with human annotators. Sometimes human annotators got confused to say if the topics were similar or not. So for that we took the opinion of majority of the annotators and compared our result with that. Our experiment has some noise though.

5. Corpus:

To test our program, we were in need of large data sets. So we have used NLTK text corpus. As we have used unsupervised approaches, we have used Brown corpus from NLTK. The Brown Corpus was the first computer-readable general corpus of texts prepared for linguistic research on modern English. It was compiled by W. Nelson Francis and Henry Kučera at Brown University in the 1960s and contains of over 1 million words (500 samples of 2000+ words each) of running text of edited English prose printed in the United States during the calendar year 1961. There are six versions of the corpus available: the original Form A, Form B from which punctuation codes have been omitted, the tagged Form C, Bergen Forms I & II and the Brown MARC Form [12]. We used the tagged Form C. The corpus originally (1961) contained 1,014,312 words sampled from 15 text categories:

6. Experiment:

The basic idea of our work is to measure the similarity of every adjacent pair of sentences, and hypothesizing a topic boundary whenever this similarity falls below some threshold, which we would like to set based on empirical observation. Since two of the studies we came across placed an emphasis on the use of repetition of words as an important feature, we began by conducting an experiment on whether this gives a good estimation of how similar two sentences are. Later we moved to paragraph level due to inconclusive results at the sentence level.

6.1 Unigram:

For our first experiment we used an algorithm based on *n* -gram models to find out word overlap between two sentences. An *n*-gram is a contiguous sequence of *n* items from a given sequence of text or speech. The items in question can be phonemes, syllables, letters, words or base pairs according to the application. *N*-grams are collected from a text or speech corpus.

At first we tried "unigram" which takes each words of a sentence and matches if it overlaps with the words of the next sentences. By doing it we have noticed that the threshold value of unigram overlapping is 14%, shown in Figure-1. So when we get higher values, human agrees that the 2 sentences are of similar topic & when threshold value is below 14% they say that the topic is different. However when the overlapping percentage is marginal which is close to 14% human tends to get confused.

6.2 Bigram:

Then we used the bigram algorithm. A bigram is every sequence of two adjacent elements in a string of tokens, which are typically letters, syllables, or words; they are *n*-grams for *n*=2. Bigrams help provide the conditional probability of a token given the preceding token, when the relation of the conditional probability is applied:

$$P(W_n|W_{n-1}) = \frac{P(W_{n-1}, W_n)}{P(W_{n-1})}$$

That is, the probability $P()$ of a token W_n given the preceding token W_{n-1} is equal to the probability of their bigram, or the co-occurrence of the two tokens $P(W_{n-1}, W_n)$, divided by the probability of the preceding token. For example, if the first word is "beauty" and the second word is "beast", then the probability can be noted as the number of times both words appear divided by number of times only the first word "beauty" appears

So, P (beauty|beast) = P(beast,beauty)/ P (beast)

From Figure-1, we have noticed that the threshold value is relatively low comparing to unigram and it is about 4%. This is because one word of a sentence can be common in two different sentences. But it's unlikely to overlap two consecutive words to two different sentences. This also means the overlap is more likely to be meaningful rather than just small noise.

Sentence pair	Unigram Overlap Size	Bigram overlap size	Human 1	Human2	human3	Human4	Human5
1st	21.43%	7.143%	yes	yes	yes	confused	yes
2nd	33.33	2.381%	yes	yes	yes	yes	yes
3rd	10.20%	0.0%	no	confused	confused	yes	yes
4th	0.0%	0.0%	no	confused	no	confused	yes
5th	15.385%	0.0%	no	no	no	no	yes

Figure-1: unigram and bigram result

6.3 Overlap of proper noun (sentence level):

As we were not getting fine tuning, we decided to use another features called “Named Entity Extraction” for proper noun at sentence level.

From the following table we can see, when threshold value is 0%, for first pair of sentence, most of the human agreed that the topic were different but for second pair of sentence, they agreed the topics were same type. Though for both of the cases, threshold value was 0%. So we can see a noise here. But when threshold value was 25% most of the human agreed that the topic were different. So for this experiment, we got no threshold value. We could not distinguish topic boundary for this data.

Sentence pair	Proper noun overlap	Human 1	Human 2	Human 3	Human 4	Human 5	Majority
1st	**0.0**	No	No	No	No	Yes	No
2nd	**0.0**	No	No	Yes	Yes	Yes	Yes
3rd	**25**	No	Confused	Yes	No	No	No

Figure-2: Result of named entity extraction (proper noun for sentence level)

From Figure-3 we can notice, for all the pair of sentences threshold value is 0%. Though our human annotators agreed for few pairs the topics were similar and for few pairs, the topics were different. So we can see, we could not get any threshold value here to find the topic boundary. So named entity extraction for proper noun at sentence level could not give us any topic boundary.

Sentence pair	Proper noun overlap	Human 1	Human 2	Human 3	Human 4	Human 5	Majority
1st	0.0	Yes	Yes	Yes	Yes	Yes	Yes
2nd	0.0	Yes	Yes	Yes	Yes	Confused	Yes
3rd	0.0	No	No	Yes	No	No	No
4th	0.0	Yes	Yes	Yes	No	Yes	Yes

Figure-3: Result of named entity extraction (proper noun for sentence level)

So we revised our plan and went to experiment in paragraph level.

6.4 Overlap of proper noun (paragraph level):

For better performance we tried to extract the proper nouns & measure the overlap rate at the paragraph level. We checked our values with human annotators & found the threshold values.

From Figure-4, we can see that, when threshold value is 10% all the users agree that the topics are similar. At level 20% users gets confused to determine if the topics are same or not. Three users agree that the topics are same but two people disagree. If we take the opinion of the majority of the people we get the topic is similar. At 50% people got confused whether they are similar topics or not. Two people said they are similar, two disagreed and one got confused. So we can say for this table with some noise, threshold value is 10% because at this value, all the users always agree that the topics are similar.

Paragraphs	Threshold %	Human 1	Human 2	Human3	Human4	Human5	Majority
1st and 2nd	10	Yes	Yes	Yes	Yes	Yes	Same topic

2nd and 3rd	20	Yes	No	No	Yes	Yes	Same topic
3rd and 4th	50	No	Yes	Confused	Yes	No	Tie
4th and 5th	20	Yes	Confused	No	yes	Yes	Same topic

Figure-4: Result of named entity extraction (proper noun)

From Figure-5 we can see, for 2nd pair when threshold value is same 0%, everybody agreed that the topic were different but for the 1st ,4th and 5th pair, when threshold value is 0%, most of the human annotators agreed the topic were same even if there was no overlap. So we found some noise here. When threshold value is 33.33%, most of the human agreed that the topic were same & same goes for 40% as well. So for this experiment threshold value is 33.33% with some noise.

Paragraphs	Threshold %	Human 1	Human2	Human3	Human4	Human5	Majority
1st and 2nd	0	No	Yes	Yes	Yes	Yes	same topic
2nd and 3rd	0	No	No	No	No	No	Different topic
3rd and 4th	33.33	Yes	Yes	Confused	Yes	Yes	Same topic
4th and 5th	0	Yes	Yes	Yes	Yes	Yes	Same topic
5th and 6th	0	No	Yes	No	Yes	Yes	Same topic
6th and 7th	40	No	Yes	Yes	Yes	Yes	Same topic

Figure-5: Result of named entity extraction (proper noun)

From the next table Figure-6 we can see, when overlapping rate is 100% meaning, all the proper nouns of the first paragraph overlap with the second one, all human agrees to the fact that the paragraphs are of same topic & when over lapping percentage is 0%, meaning none of the proper noun of the first paragraph overlap with the second one, all human agree that it's of different topic. Also we have noticed for the third and fourth pair, three users agreed that the topics are different and two users agreed that the topics are not. As a majority of users agreed the topics are not similar, we considered their opinion and found a topic boundary. Again, when the threshold value is 33.33%, one got confused to figure out whether the topic are similar or not and the rest agreed that the topics are of similar. So for this table, our threshold value is 33.33%. So when we get higher values, humans agreed that the two sentences are of similar topic & when threshold value is below 33.33% they say that the topic is different.

Paragraphs	Threshold %	Human 1	Human2	Human3	Human4	Human4	Majority
1st and 2nd	100	Yes	Yes	Yes	Yes	Yes	same topic
2nd and 3rd	0	No	No	No	No	No	Different topic
3rd and 4th	0	Yes	No	Yes	No	No	Different topic
4th and 5th	50	Yes	Yes	No	Yes	No	Same topic
5th and 6th	0	No	Yes	No	No	No	Different topic
7th and 8th	0	No	No	No	No	No	Different topic
8th and 9th	0	Yes	No	Confused	Yes	Yes	Same topic
9th and10th	33.33	Yes	Confused	Yes	Yes	Yes	Same topic

Figure-6: Result of named entity extraction (proper noun)

From the Figure-7, we can see, when threshold value is 11.11%, two annotators agreed that the topic are different and three agreed they are similar. When threshold value is 25%, all human agreed that the topics are similar. When the value is 33.33% all the human except one agreed that the topic are similar but, when threshold value is 0% meaning when there is no overlapping of proper nouns between the two paragraphs, most of the human agreed the topic are similar. So for this experiment we got some noise. Taking the opinions of the majority of the people, we can say here the threshold value is 11.11% with some noise.

Paragraphs	Threshold %%	Human 1	Human2	Human3	Human4	Human5	Majority
1st and 2nd	11.11	Yes	Yes	Yes	No	No	Same topic
2nd and 3rd	9.09	Yes	No	No	Yes	No	Different topic
3rd and 4th	0	Yes	No	Yes	Yes	Yes	Same topic
4th and 5th	25	Yes	Yes	Yes	Yes	Yes	Same topic
5th and 6th	33.33	Yes	Yes	Yes	Yes	No	Same topic

Figure-7: Result of named entity extraction (proper noun)

Average threshold values of all the tables are 21.94%

6.5 Both proper noun and common noun extraction:

For our last experiment, we tried to extract both proper noun and common noun together to measure the overlap rate at the paragraph level. From our Figure-6 we can notice, when threshold value is 11.11% all the human annotators agreed that the topics of the paragraphs are of same. When the value is 5.88% three users agreed that the topics are same, one got confused to figure out whether it was of same topic or not and the other thought it was a different topic. But as majority of the people agreed that the topics were same, we took their opinion. But we can also see when threshold value is 9.09% (which is pretty high), there's a tie between the human annotators. 50% said they were same topic and 50% didn't agree. So we can say that here is some noise. For this table threshold value is 5.88

Paragraphs	Threshold %	Human 1	Human 2	Human3	Human4	Human5	Majority
1st and 2nd	11.11	Yes	Yes	Yes	Yes	Yes	Same topic
2nd and 3rd	9.52	Yes	No	No	Yes	Yes	Same topic
3rd and 4th	9.09	No	Yes	Confused	Yes	No	Tie
4th and 5th	5.88	Yes	Confused	No	Yes	Yes	Same topic

Figure-8: Result of named entity extraction (proper noun and common noun)

From Figure-9, we can see when threshold value is 8.82 most of the human annotators agreed the topics were same. But, when threshold value is 2.85% most of the human annotators agreed that they were different topic but for 3rd pair we can see the human said the topics were similar and threshold value was 6.45%. So we can see for this data topic boundary lies at this threshold.

Paragraphs	Threshold %	Human 1	Human2	Human3	Human4	Human5	Majority
1st and 2nd	8.82	Yes	Yes	No	Yes	No	Same topic
2nd and 3rd	2.85	No	No	No	No	Yes	Different topic
3rd and 4th	6.45	Yes	Confused	Yes	Yes	Yes	Same topic
4th and 5th	9.37	Yes	Yes	Yes	Yes	Yes	Same topic

Figure-9: Result of named entity extraction (proper noun and common noun)

From Figure-10, we can took at the threshold value 5.88% we can see that most of the human annotators agreed that the topic is different. For the threshold values 7.69%, 14.29% and 13.46% the majority agreed the topic to be same. So, by taking the lowest value of the agreement threshold we have decided the topic boundary of this data to be 7.69%. But for the 7th pair of paragraphs even though the threshold value is 0% the human reader agreed that the topic was similar so we can consider this as an experimental noise.

Paragraphs	Threshold %	Human 1	Human2	Human3	Human4	Human5	Majority
1st and 2nd	7.69	Yes	Yes	Yes	Yes	Yes	same topic
2nd and 3rd	0	No	No	No	No	No	Different topic
3rd and 4th	0	Yes	No	Yes	No	No	Different topic
4th and 5th	14.29	Yes	Yes	No	Yes	No	Same topic
5th and 6 th	5.88	No	Yes	No	No	No	Different topic
7th and 8 th	0	No	No	No	No	No	Different topic
8th and 9th	0	Yes	No	Confused	Yes	Yes	Same topic
9th and10th	13.64	Yes	Confused	Yes	Yes	Yes	Same topic

Figure-10: Result of named entity extraction (proper noun and common noun)

From Figure-11 we can see when threshold value is 6.67% four of our annotators agreed that the topics were same. For the 2nd pair of paragraphs the topic boundary is 0% and all the human annotators could easily distinguish the topic differences. When the value is 7.69% and 10.52% most of the human agreed that the topics were same. But for the 5th pair of paragraph even if the threshold value was 0%, human agreed that to be same topic. This is experimental noise. So extracting both proper and common noun does not guarantee a noise free result. From this table we can say that the performance of proper noun and common noun extraction is better than only proper noun extraction. Because for the first pair where most of the human agreed the topics were same, we got threshold value 0% for proper noun extraction. But for proper and common noun together, threshold value was 6.67%. So we can come to the conclusion that, proper noun and common noun extraction gives more fine tuning to the threshold value and it has less frequent noise.

Paragraphs	Threshold %	Human 1	Human2	Human3	Human4	Human5	Majority
1st and 2nd	6.67	No	Yes	Yes	Yes	Yes	same topic
2nd and 3rd	0	No	No	No	No	No	Different topic
3rd and 4th	7.69	Yes	Yes	Confused	Yes	Yes	Same topic
4th and 5th	7.69	Yes	Yes	Yes	Yes	Yes	Same topic
5th and 6th	0	No	Yes	No	Yes	Yes	Same topic
6th and 7th	10.52	No	Yes	Yes	Yes	Yes	Same topic

Figure-11: Result of named entity extraction (proper noun and common noun)

Average of threshold values of all the tables are: 6.6725%

7. Future work: In this paper, we have quantitatively assessed the effectiveness of finding topic boundary in document level using *n*-gram and named entity extraction. In future, if we can introduce it with synonyms, the performance will be even better. For example: for proper noun extraction, if there's a common noun in the first paragraph & synonym of that in the second paragraph our program cannot detect, it's similar. So in the future it can be done so that it understands the synonyms. For our thesis, we did not train human annotators about how to annotate the corpus. So in future we can train them to check if that helps to get even more accurate and better result.

8. Conclusion: We have noticed that named entity extraction as proper noun extraction and proper and common both noun extraction together works better than unigram and bigram. Even proper and common both the noun together extraction gives better result than only proper noun extraction, though we got some noises. One of the reasons could be we did not train our annotators about reading the corpus. We took their opinions randomly. If we could train them, we might have obtained better performance.

References:

[1] TopCat: Data Mining for Topic Identification in a Text Corpus – Chris Clifton, Robert Cooley

[2] Unsupervised Topic Clustering of Switchboard Speech Messages – Beth A. Carlson

[3] Automatic Topic Segmentation and Labeling in Multiparty Dialogue - Pei-hun Hsueh, Johanna D. Moore

[4] Statistical Model for Topic Segmentation – Jeffrey C. Reynar

[5] C. Stanfill and D. Waltz. Toward memory-based reasoning. Commun. ACM, 29(12):1213{1228,

1986.

[6] Similarity Measures for Categorical Data: A Comparative Evaluation - Shyam Boriah, Varun Chandola, Vipin Kumar

[7] Similarity measures for sequential data- Konrad Rieck

[8] http://en.wikipedia.org/wiki/Unsupervised_learning

[9] Unsupervised Learning* - Zoubin Ghahramani†

[10] http://www.apperceptual.com/ml_text_cohesion_apps.html

[11] http://en.wikipedia.org/wiki/Cohesion_%28linguistics%29

[12] Francis, W. Nelson 1964. *Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers.* Providence, R.I.: Department of Linguistics, Brown University.

[13] MANUAL OF INFORMATION to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. By W. N. Francis H. Kucera Brown University

[14] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[15] Adam Wilcox, M.A., George Hripcsak, M.D. - Knowledge Discovery and Data Mining to Assist Natural Language Understanding
[16] NLTK download kit http://www.nltk.org/download