Automatic summarization of cricket text commentaries using NLP methods [*]

Abhinav Ravi
IIT Guwahati
r.abhinav@iitg.ernet.in

Debendra Kumar Naik
IIT Guwahati
n.debendra@iitg.ernet.in

Subrata Tikadar
IIT Guwahati
t.subrat@iitg.ernet.in

Pintu Kumar
IIT Guwahati
pintu.kumar@iitg.ernet.in

ABSTRACT

These days huge amount of information is stored online. Area of automatic text summarization explores the Natural Language Processing (NLP) methods to generate summaries of this huge data. Automatic summarization of large text is one of the most challenging areas of research in NLP domain. This reduces a large content of text, retaining the important information of the text. We aim at analyzing the game of cricket by generating the summaries of text commentaries. Different text summarization techniques will be used to design the model. The proposed model summarizes the cricket text commentary. This could be used to produce correlations among the different entities of the game viz. players, venues, innings etc.

Keywords

Automatic summarization, NLP, Topic modeling, TF-IDF, K-mean clustering

1. INTRODUCTION

One of the NLP's trending applications these days is automatic document summarization. Extracting import information from the large texts is one of the challenging areas of research. Texts are created from multiple number of text documents in which each document conveys the important information in an abstract manner. From the large text it is difficult to extract the exact information [18]. Automatic summarization is one of the solution for this problem. Summarization captures the information in three aspects.

1. summarization may be captured from single document or multiple documents,
2. Summarization may be in short form.
3. Summarization may capture exact information.

A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. Summarization methods can be classified into extractive and abstractive summarization.
An extractive summarization is the process of selecting important information from the sentence and paragraphs etc [15]. of the original documents. The important information extracted based on the different features of the sentences.
Abstractive summarization develops the main concepts in documents and express in the natural language. This method uses the linguistic approach to examine the sentences [10].

Statistical and linguistic features of sentences are emphasized to decide the importance of the sentences [14]. On the other hand, abstractive summarization method understands the original text and rewrites that in fewer words [4]. The method uses the linguistic methods to examine the text but interpret it with new concepts and expressions for best describing it by generating new shorter text that carries the most important information of the original text document [6].

Although the extractive summarization is easier than the abstractive one, there are some problems. First of all, some unnecessary part of the segments also get included; at the same time it sometimes cannot capture the relevant and important information unless the summary is long enough to hold all those sentences. Secondly, conflicting information may not be presented accurately always.

We will use both the techniques in our application. Extractive summarization will be used for clustering and abstractive method will be used for the final summarization form clusters; because finally we are going to find out the match summery as well as different correlations among different entities like the correlation between player and venue of the match to see the performance of a particular player in different venues worldwide and so on. For clustering we will use K-mean clustering method; and for final summarization we will use Topic Modeling Method. There are some other methods like machine learning technique, graph based method, TF-IDF (Term Frequency-Inverse Document Frequency) method, etc.

2. MOTIVATION / BRIEF REVIEW

It is very difficult for humen beings to manually go through the large records of players to extract the relevant information and draw co-relation among different events and entities.

1. The work done in [13] uses score cards and coverage articles to acquire details about a particular match, while in our case we use cricket text commentaries. The reason being that text commentary captures each and every minute details of the game at each ball.
2. A very limited use of cricket text for video segmentation

[*] This template is adapted from http://www.acm.org/publications/article-templates/SIG%20Proceedings%20Template-May2015%20Zip.zip

was attempted by [19]. This does not fall under the domain of natural language processing.
So under the view of above circumstances we present the summarization based techniques to analyze cricket text commentary [16].

3. PROBLEM STATEMENT

Multi Document Automatic summarization of text commentary to obtain the match summaries as well as to draw correlations among different entities of the game viz. players, venues, innings etc. Clustering and Summarization techniques are used for designing our model.

4. CHALLENGES

4.1 Peculiar nature of cricket text commentaries:

4.1.1 Short text and Data Sparsity

Each text commentary is at most 20 words long. It poses a challenge of handling data sparsity of words in them. As cricket text commentary has a definite structure in which both bowler and batsman actions are described. Commentators at times focus either only on bowler's action or only on batsmen action. This poses challenge in extracting all the relevant features across every delivery commentary.

4.1.2 Stop Words

Most of the words which are treated as stop words in English dictionary are important words in the cricket context. The words like on(on stumps), off(off stumps), into(into the pads), from, up, behind etc. are stop words in English but are important in cricket glossary. Table1 enumerates such words.

4.1.3 Round about speech, periphrasis, or ambage

Roundabout speech refers to using many words to describe something for which a concise and commonly known expression exists. In linguistics, periphrasis is a device by which grammatical meaning is expressed by one or more free morphemes (typically one or more function words accompany a content word), instead of by inflectional affixes or derivation.

For example: As most of the time commentators explain any event indirectly like ”out event” can be explained by:

”Hoggard strike a beauty! swing in Tharanga play forward beat him completely goes between the bat and pad and knock the stumps”.

The inference that player got out can only be drawn if we derive the semantic meaning of above commentary.

5. PROPOSED DIRECTION

5.1 About the Data set

Text commentaries were obtained from the sites using the web crawlers written in python. We have gathered 5 years of test match text commentaries [1]. In the field of cricket huge amount of data is documented in the form of scorecards, video broadcasts, text commentaries and coverage articles. We focused on analyzing text commentaries because it expresses each and every minute details about the game.Each text commentary details the set of events following each delivery of ball, while other sources fails to do so. Each text commentary is treated as a single document.

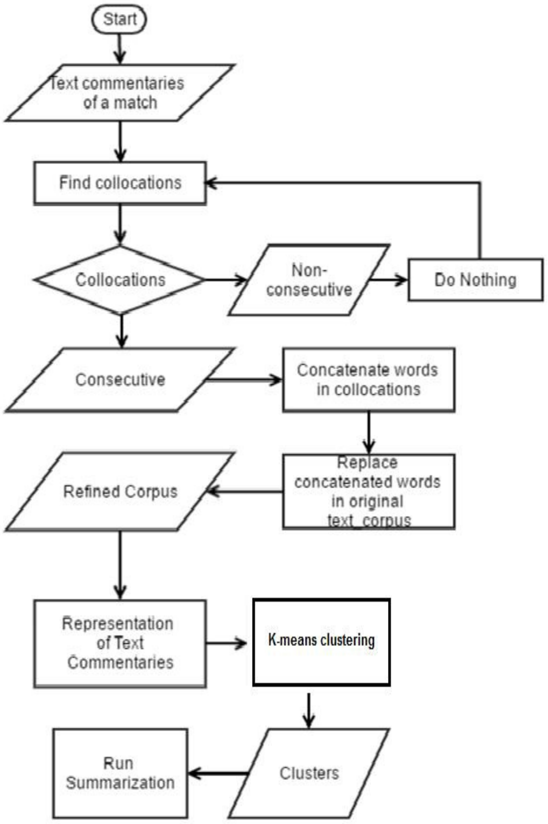

Start
Text commentaries of a match
Find collocations
Collocations
Non-consecutive
Do Nothing
Consecutive
Concatenate words in collocations
Replace concatenated words in original text corpus
Refined Corpus
Representation of Text Commentaries
K-means clustering
Clusters
Run Summarization

Figure 1: Flow diagram for the proposed model

Text commentary could be explaining about different topics viz. bowler, batsman, fielding, qualities of ball, type of shot played etc.. These topics are the semantic structure running across the text commentaries. They represent the theme of the text commentaries. Extracting these themes would help us in summarizing the text commentaries.

5.2 Preprocessing the Dataset

5.2.1 Stop word removal and stemming

After obtaining the data set next important thing is to represent it by removing stop words and stemming it. In the context of cricket text commentaries, there is a need to focus on the fact that most of the stop words into English dictionary are treated as important informative word in cricket world. We obtained such a list from [2]. The cricket glos- sary was obtained from [3] were removed which are not present in the cricket glossary.Such words are ”for,a,of,the and, to” etc.

Now all such words are ignored which have intersection with the cricket glossary from being considered as stop words. These were not removed.

5.2.2 Concatenation of Consecutive Collocations

There is a need to focus on the fact that collocations consisting of two or more words together convey a semantic meaning.

Stop words							
off	on	room	across	behind	back	out	place
good	great	into	away	up	down	long	below
turn	point	from	further	under	full	open	high

Table 1: List of stop words which are significant for cricket text commentaries

In the context of collocations we need to focus on how consecutive and non-consecutive collocations can contribute differently in specifying a thematic pattern [9].

On one hand consecutive collocations can be merged together to be represented as a single token in the final sentence representation, while on other hand non-consecutive collocations are left as it for further processing.

The consecutive collocations capture tiny themes within the sentence. There could be multiple such tiny themes present in a particular text commentary.

For example:

1.good ball first up defend well to the off side.
2. good ball out side the off stump play at it away from his body beat the bat.
3. good ball defend well to short leg.
4. good ball on target Sangakkara forward defend well down the track.
5. good ball get the edge wide of Strauss at third slip ball run away to the third man boundary.

Here consecutive collocations ‘good ball’ capture the quality of ball and can be merged together as a single token ’good-ball’ for the final sentence representation.

Non- consecutive collocations will play a momentous role while filtering the sentences for finding topics among them. They bind the different tokens in the sentence together to finally represent an aggregate theme for the whole sentence.

Non- consecutive collocations also captures the correlations among different entities of the game play.

For example consider below commentaries:
1. excel deliveri outside the off stump Sangakkara play at it and get beaten.
2.excel deliveri outside the off stump Sangakkara play forward beat the outside edge through to Jones.

Here ‘excel deliveri’ and ‘Beat’ are the non- consecutive collocations but it helps in deriving the correlation between ‘Ball type’ (excel deliveri outside the off stump) and ‘players response’(Beaten).

5.2.3 *Processing for Roundabout speech, periphrasis, or ambage*

It will be explained in the final paper once discussion on ‘Event detection’ is done in the class.

5.2.4 *Text commentary representation*

In the context of summarization of the text commentaries and draw the inferences out of it, we need to first focus on the way we represent the text commentaries to facilitate so.This is a very standard problem and comes under the domain of short text representation.
Text commentaries(documents) are represented as a set of feature vectors from the words present in the sentence. Here a document represents a data point in d-dimensional space where d is the size of the corpus vocabulary. Document doc_i is represented as $(v_1, v_2, ...v_d)$ where v_{ij}s the tf_{idf} score for of jth word in doc_i [11].

F_{id}/S_{id}	F_1	F_2	F_3	.	F_n
S_1	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}
S_2	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}
:	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}
:	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}
:	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}
S_n	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}	tf_{idf}

Table 2: Representation of text commentaries as feature vector

6. PROPOSED MODEL

This paper mainly based on two techniques first one is clustering of text document and second one is summarization of text document.

- From the architecture diagram as shown in the figure. 2 the first step is the clustering of the text commentaries for which the input will be set of text commentaries for which we want to obtain the summary.

- Output from the first step is the clusters of similar text commentaries.The LDA would be performed on the above obtained clusters.

- For mining the co-relation between entities like a player’s performance in home ground and abroad pitches, we have to run our model on the text commentaries of that particular player on home ground and abroad pitches.

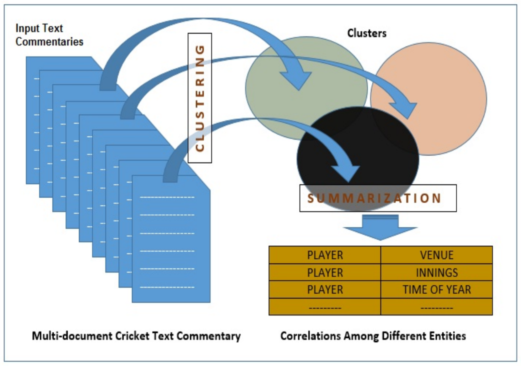


Figure 2: Architecture Diagram

6.1 Clustering Of Documents

The first process of our model is the clustering of text document. K-mean clustering technique is used for clustering the text document into k number of cluster [7]. This process can be classified in three different process which is describe below:-

6.1.1 Step-1

The Table.2 shows that each sentence or documents compared with each features; and if its corresponding features is present in this document then its gives the term frequency and weighting.
F_i: List of different features.
S_i: List of different documents with corresponding sentence.

$$W_{ij} = \begin{cases} \mathrm{tf}_{idf}, & if \quad F_i \in S_i \\ 0, & otherwise. \end{cases}$$

6.1.2 Step-2:Execution of the k-mean clustering on text commentaries.

Input:
Each input sentence is a text commentary.
A matrix whose row is identified by sentence-id(s_{id}) and column is identified by feature-id(f_{id}). Value in cell (i,j) is calculated by fraction of words(within the feature vector with feature-id j) found in the Sentence with Sentence-id i. Sentence set is $S_1, S_2, , S_N$each of the feature vector is represented by the corresponding row of the matrix. K is the parameter which signifies the number of clusters to be formed.

Algorithm:Lloyd's implementation
The Algorithm starts with initializing the mean-set μ with random K sentences [12]. Now for each sentence it calculates the distance from each μ_j ,1<=j<=K and puts the sentence in the cluster with minimum distance from this sentence. Thus K clusters are formed. New means $\mu - j$,1<=j<=K corresponding to new clusters are calculated and μ is updated with the new means.
The process of finding minimum distant mean from the sentence, allocating it to that cluster and updating new means is repeated until termination condition is met.
Termination condition:
Terminate when the decrease in RSS falls below a threshold theta. For small θ this indicates convergence has reached. A bound on number of iterations to prevent very long runtimes.
Optimal values of K: Sum of Squared Errors (SSE)
For finding the optimal clustering, we have to find the optimal no. of clusters required to segregate our text commentaries.
we have used Sum of Squared Error (SSE) to find the optimal value of k. It is defined as the sum of the square of the distances of the points in a cluster from the cluster mean value.
From the figure 3 we have got an elbow shape curve with the minimum SSE error value at k=8.
This value will be close to the optimal minimum, moreover finding optimum is a NP hard problem.

6.1.3 Dealing with sparse vectors: Singular Value Decomposition (SVD)

As we have already discussed that short text and sparse vector as one of the major problems for the texts commentaries. To deal with the sparseness of vector representation of the text commentaries, we have employed the dimensionality reduction using singular value decomposition [9][8]. Thus SSE error for the number of clusters has been reduced as enumerated in figure 4.
The improvement in the SSE error has been observed as the difference between the red and the blue lines.
The red line is the SSE error corresponding to the clustering using naive k-means on the sparse vectors with optimal no. of clusters =8, while the blue one is the clustering results for dimensionally reduced vectors with local minimum at k=10.
We obtain K clusters. Each cluster corresponds to collection of similar text commentaries. We run summarization on each of these clusters.

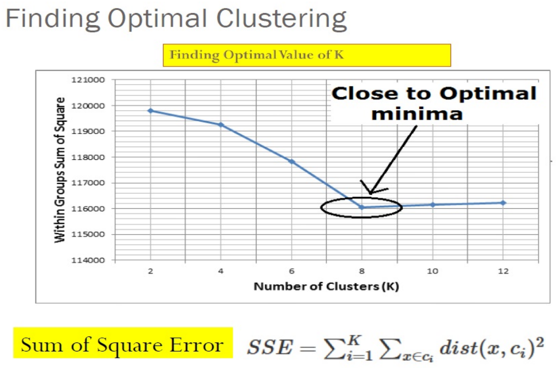


Figure 3: Finding Optimal Clustering

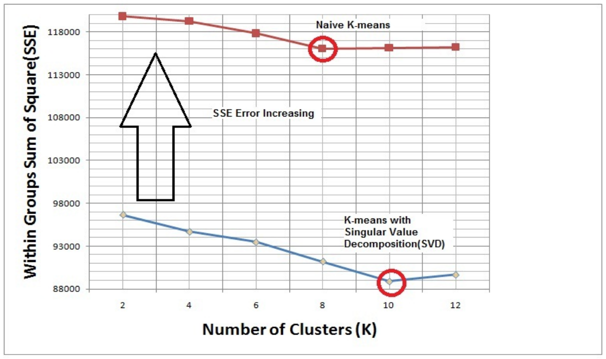
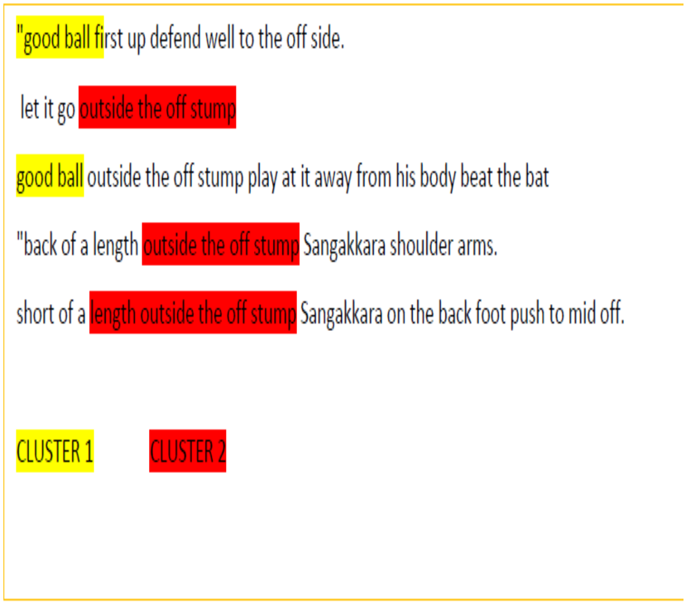


Figure 4: SVD, lowering the SSE error

6.2 Summarization technique:

The limitations of the TF-IDF reductions approach for summarization was that it really results in a very little reduction in the description [17]. It also fails to cover the intra document statistics.
The limitations of the TF-IDF reductions approach for summarization was that it really results in a very little reduction in the description. It also fails to cover the intra document statistics.
Hence to fulfilling the need for dimensionality reduction and overcome above limitations, idea of latent space was introduced. Latent space uncovers the hidden thematic/pattern. Singular Value Decomposition (SVD) was used by was Deerwester et al.[9] to project the documents in the lower dimension.
They ultimately obtain the features as a co-relational among the terms and label it as feature .The weight of these features are nothing but the linear combination of original score



"good ball first up defend well to the off side.

let it go outside the off stump

good ball outside the off stump play at it away from his body beat the bat

"back of a length outside the off stump Sangakkara shoulder arms.

short of a length outside the off stump Sangakkara on the back foot push to mid off.

CLUSTER 1 CLUSTER 2

Figure 5: Clustering of text commeantaries based on Features similarity

for TF-IDF features.
In addition it can address synonymy and polysemy aspect of language processing.
Though the notion of using LSI was not clear as we can take a more direct approach of fitting the model into the data by employing maximum likelihood and Bayesian methods.It paved the way for probabilistic LSI known as pLSI which was based on likelihood.Hence came Hofmann [11] with probabilistic LSI. It was an important step towards probabilistic representation of the text.It considered document as a collection of mixture models(topics).Each word in the document can come from any topic.The topics were multinomial random variables.Hence ultimately documents were represented as distribution over the topics.
Though pLSI was very helpful but the shortcoming in it was that it provided no probabilistic model at the document level that is there was no probabilistic model to generate the topics.Other demerit was the over fitting problem that is no. of parameters grow linerally with the corpus size.
Hence came the LDA [5] which is a 3 tier process.First at the corpus level, second at the document level and third at the words level as shown in 6 It is important to note that all the topic modeling approach are based on ”bag of word”assumption means the relative ordering of the words are immaterial for us within a document.This assumption is formally treated as ‘exchangeability’.

6.3 LDA workflow for text commentaries

- LDA is a set of Unsupervised machine learning algorithms which learns the underlying themes/patterns of a large collection of unorganized text commentaries over the years [20].
- It analyzes the words of the original texts commentaries to discover the themes that run through them to discover topics from whole text commentary corpus.

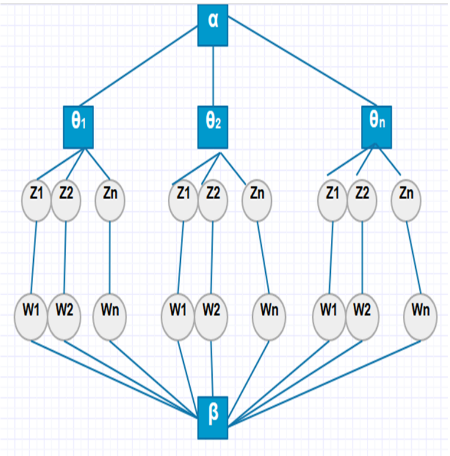


Figure 6: 3-tier expansion for LDA

- It models Connection among topics, how they are connected to each other.
- Can model evolution of topics over time. How the topics evolve over the time and each year/play quarters etc.
- They are hierarchical probabilistic models of text that uses multinomial/categorical distribution over words to capture the hidden thematic.

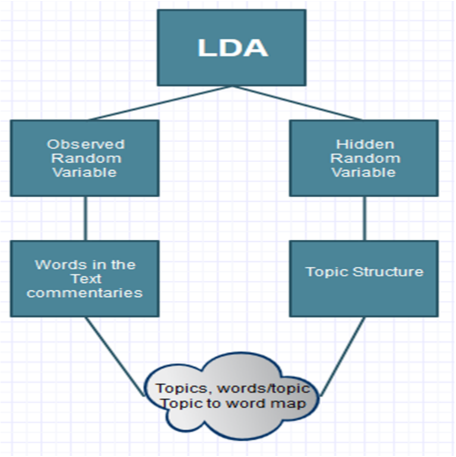

LDA
Observed Random Variable
Hidden Random Variable
Words in the Text commentaries
Topic Structure
Topics, words/topic Topic to word map

Figure 7: Workflow of LDA

- We define a joint probabilistic distribution over the both observed variable (words of the text commentaries) and hidden random variables(Topic Structure) as enumerated in figure 7.
- Then we perform a data analysis by using that joint distribution to compute conditional distributional of the hidden variables(uncovered topics) given the observed variables.
- This Conditional distribution is also called posterior distribution.

7. RESULTS AND ANALYSIS

The Results will be obtained as probability distribution of set of words in set of topics.
For clusters with more no. of text commentaries in it, we have included more no. of topics for that cluster.For clusters with lesser no. of text commentaries in it, we have included only lesser no. of topics for that cluster.
The results for Indian Innings are:
topic 0:
0.217*back + 0.165*foot + 0.113*played-off + 0.113*point + 0.113*towards + 0.062*single + 0.062*cover + 0.062*bowler-off + 0.062*played + 0.010*outside-off

topic 1:
0.112*back + 0.112*foot + 0.112*cut-off + 0.111*deep + 0.111*point + 0.111*two + 0.111*more + 0.110*single + 0.110*through

topic 2:
0.039*on + 0.025*rahane + 0.020*rohit + 0.019*forward + 0.018*wicket + 0.018*back + 0.017*towards + 0.017*where + 0.017*short-length + 0.017*ball

topic 3:
0.050*foot + 0.044*front + 0.031*on + 0.025*ball + 0.022*towards + 0.019*defended + 0.019*outside-off + 0.017*leg + 0.016*into + 0.013*rohit

topic 4:
0.038*on + 0.029*ball + 0.027*leg + 0.022*moves + 0.019*front + 0.018*square + 0.018*at + 0.017*foot + 0.015*that + 0.015*rahane

topic 5:
0.031*leg + 0.024*on + 0.022*into + 0.021*rohit + 0.020*square + 0.019*from + 0.017*through + 0.016*ball + 0.015*tries + 0.015*front

topic 6:
0.025*leg + 0.023*square + 0.019*behind + 0.018*off-pads + 0.018*rahane + 0.018*as + 0.017*8 + 0.016*angling + 0.016*ball + 0.016*rohit

topic 7:
0.022*ball + 0.021*kohli + 0.017*that + 0.017*deep + 0.014*short + 0.014*midwicket + 0.012*at + 0.011*back + 0.011*up + 0.011*in

topic 8:
0.036*ball + 0.023*in + 0.020*from + 0.014*kohli + 0.012*wide + 0.011*outside-off + 0.011*rohit + 0.009*boundary + 0.009*that + 0.008*on

topic 9:
0.031*in + 0.026*kohli + 0.023*on + 0.019*ball + 0.017*from + 0.015*long + 0.014*crease + 0.013*full + 0.012*out + 0.011*fielder

topic 10:
0.025*on + 0.024*ball + 0.023*kohli + 0.016*midwicket + 0.013*through + 0.013*in + 0.013*towards + 0.012*that + 0.012*they + 0.011*at

topic 11:
0.013*leg + 0.013*in + 0.013*ball + 0.012*dhoni + 0.011*not + 0.011*bravo + 0.010*fine + 0.010*there + 0.010*full + 0.009*boundary

Each topic for each cluster has to be analyzed to obtain the analysis.

Analysis for Indian Batting Innings

1.Rahane and Rohit Sharma(Sharma) played defensive back foot shots on good length and outside off deliveries of west indies bawlers to the points and cover points, Outcome: single runs.
2. Kohli has shown attacking play while playing at backfoot contrast to Rahane and Sharma to the points and deep point, Outcome: single run.
3. Rahane and Rohit play short length balls by coming forward.
4. Rohit Shrama defends ball outside off and leg side by coming at front foot.
5. Rahane and Rohit plays leg side balls towards square direction.
6.Kohli has played short balls towards midwicket to four runs.
7.Kohli and Rohit hits wide and outside off balls to boundary.
The results for West Indies Innings are:

topic 0
0.040*ball + 0.015*back + 0.014*full + 0.012*at + 0.012*gayle + 0.011*leg + 0.011*on + 0.011*from + 0.010*misses + 0.010*short-length

topic 1
0.033*ball + 0.018*on + 0.016*in + 0.015*charles + 0.014*towards + 0.014*outside-off + 0.012*that + 0.012*away + 0.010*cover + 0.010*plays

topic 2
0.029*on + 0.024*ball + 0.021*in + 0.017*leg + 0.016*foot + 0.016*outside-off + 0.013*back + 0.012*charles + 0.012*simmons + 0.011*at
topic 3
0.025*simmons + 0.018*boundary + 0.017*back + 0.015*ball + 0.015*in + 0.015*on + 0.013*foot + 0.012*that + 0.012*short + 0.012*leg

topic 4
0.022*in + 0.013*on + 0.012*ball + 0.011*simmons + 0.010*back + 0.010*west + 0.009*midwicket + 0.009*russell + 0.009*good-length + 0.009*outside-off

Analysis for West Indies Batting Innings
1.Gayle misses ball of short length and full length, plays from back.
2.Charles plays outside off balls towards cover.
3.Simmons played boundary for the short balls at back foot positions.
4.Simmons and Russell plays good length and outside off balls to the mid wicket region.

8. CONCLUSION AND FUTURE WORK

We have explored various methods for summarization of short text commentaries. Out of these Topic modelling ought to be best for our case. We have proposed feature vector model for text commentary representation.
For clustering, K-means method has been used. We have improved it's performance using dimensionality reduction techniques.
Optimal value of k is used using the sum of square error analysis technique. Implementation of the proposed model provides the abstract summaries of the match.

Moreover we are able to draw the correlation among different entities by running the above model for text commentaries specific to those entities. For example: To draw the inference about the performance of the players we have run our model for text commentaries for the T-20 match played between India and West Indies for T-20 world cup 2016. Summaries obtained for different cases are helpful in deriving the player's performance on different ball types.
Still we require some work on how to deal with the round about speeches. Future work will try to achieve on handling round about speeches.

9. ADDITIONAL AUTHORS

10. REFERENCES

[1] http://www.espncricinfo.com.

[2] http://xpo6.com/list-of-english-stop-words/.

[3] http://www.cricker.com/glossary/.

[4] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine learning*, 34(1-3):177–210, 1999.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] G. Carenini and J. C. K. Cheung. Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41. Association for Computational Linguistics, 2008.

[7] X. Cui and T. E. Potok. Document clustering analysis based on hybrid pso+ k-means algorithm. *Journal of Computer Sciences (special issue)*, 27:33, 2005.

[8] L. De Lathauwer, B. De Moor, J. Vandewalle, and B. S. S. by Higher-Order. Singular value decomposition. In *Proc. EUSIPCO-94, Edinburgh, Scotland, UK*, volume 1, pages 175–178, 1994.

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[10] K. Ganesan, C. Zhai, and J. Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*, pages 340–348. Association for Computational Linguistics, 2010.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[12] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, 2002.

[13] C. Kelly, A. Copestake, and N. Karamanis. Investigating content selection for language generation using machine learning. pages 130–137, 2009.

[14] A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In *Advances in Information Retrieval*, pages 181–196. Springer, 2004.

[15] G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. 2005.

[16] N. Nagwani. Summarizing large text collection using topic modeling and clustering based on mapreduce framework. *Journal of Big Data*, 2(1):1–18, 2015.

[17] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004.

[18] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207, 1997.

[19] K. P. Sankar, S. Pandey, and C. Jawahar. Text driven temporal segmentation of cricket videos. In *Computer Vision, Graphics and Image Processing*, pages 433–444. Springer, 2006.

[20] H. M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.