Document Summarization

Submitted by :

Abhishek Gangwar	15CS60R24
Ajay Jaiswal	15CS60R10
Krunal Parmar	15CS60R13
Piyush Balwani	15CS60R25

Guided by :
Dr. Animesh Mukherjee

Contents

Introduction

The Document Summarization is defined as to find a representative subset of the data, which contains the *information* of the entire document.

"Summaries as short as 17% of the full text length speed up decision making twice, with no significant degradation in accuracy."

"Query-focused summaries enable users to find more relevant documents more accurately, with less need to consult the full text of the document."

[Mani et al., 2002]

Motivation

- Summarization technology include document summarization, image collection summarization, video summarization etc.
- Simulate the work of intelligence analyst.
- Extract non redundant more centralized data from text.
- Judge if a document is relevant to a topic off interest.
- Reduce search time.
- Finding more relevant and as much information quicker.
- Quick Overview: Allow user to overview the chapter quickly.
- Improve Productivity: It not only improves the productivity of the chapter, but also the whole document.
- Work instantly: Retrieve the less, but all important information, which helps user to make instant conclusions.
- Speed up in surfing: As it gives overview of the document. This improves productivity of surfing.
- Important Facts: As it summarizes the document, therefore all important facts will be easily referred by the user.

The development of numerous summarization applications for news, email threads, lay and professional medical information, scientific articles, spontaneous dialogues, voicemail, broadcast news and video, and meeting recordings.

Automatic text summarization can be used:

- To summarize news to SMS or WAP-format for mobile phones/PDA.
- To let a computer synthetically read the summarized text. Written text can be to long and boring to listen to.
- In search engines to present compressed descriptions of the search results (e.g. Internet search engine Google).
- In keyword directed subscription of news which are summarized and pushed to the user (e.g. Nyhetsguiden in Swedish)
- To search in foreign languages and obtain an automatically translated summary of the automatically summarized text.

Dataset used

- Dataset used: **DUC01** (Document Understanding conferences), which contains
  - Context text,
  - Abstract Summary of 303 documents

Previous work

- Express each sentence of the Document as a feature vector.
- Calculating Cosine Similarity between contexts and create the network .
- Calculate degree , average shortest path and locality index strategy for each node and rank accordingly and selected top ranked sentences.

Methods Implementations

- **KL Divergence**
- **Probsum Method**
- Clustering with Page Rank.
- LDA
- Doc2Vec with Louvain

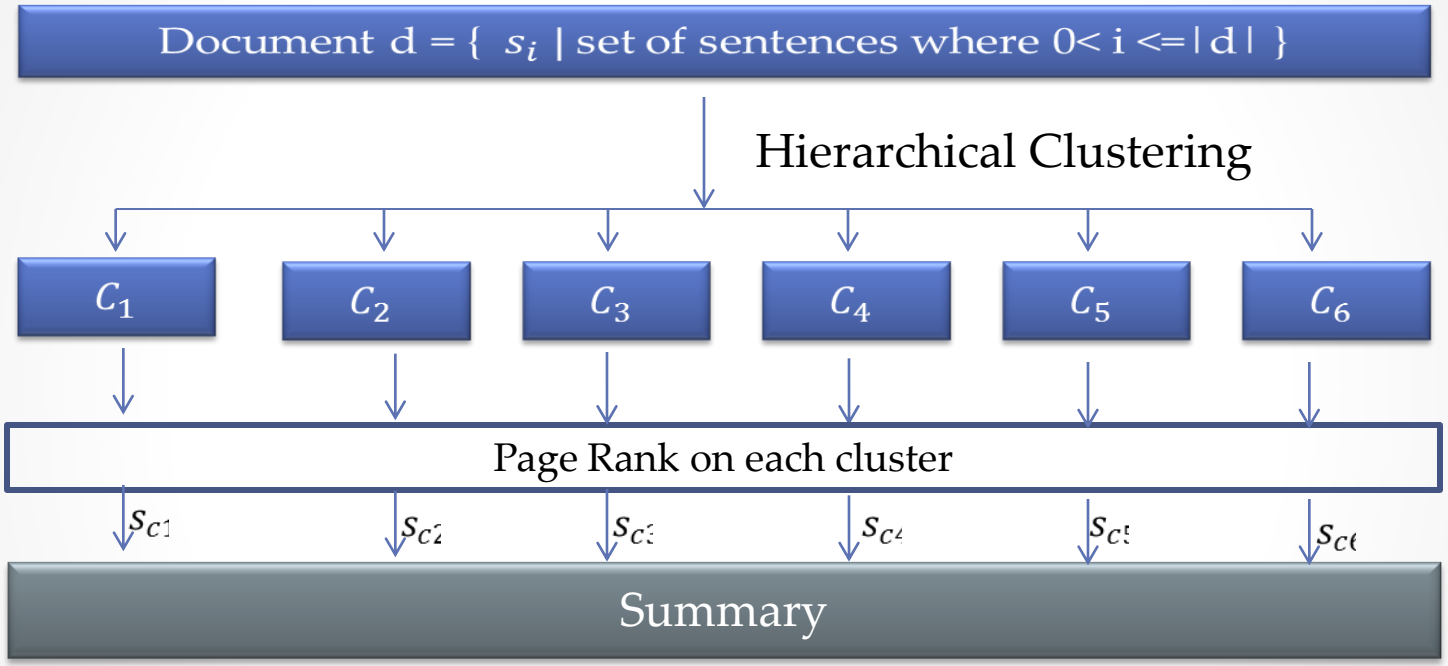
Observed Results

Method	Variations	Summary length	Recall	Precision	F1-Score
1st Approach			0.392245	0.403756	0.396781
KL		6 Sentences	0.471413	0.354657	0.398638
		100	0.385981	0.394678	0.389819
ProbSum					
	withNNP	100	0.305918	0.415645	0.305918
	Without NNP	100	0.323288	0.439959	0.363677
	withNNP(new normalisation)	100	0.330116	0.436304	0.36502

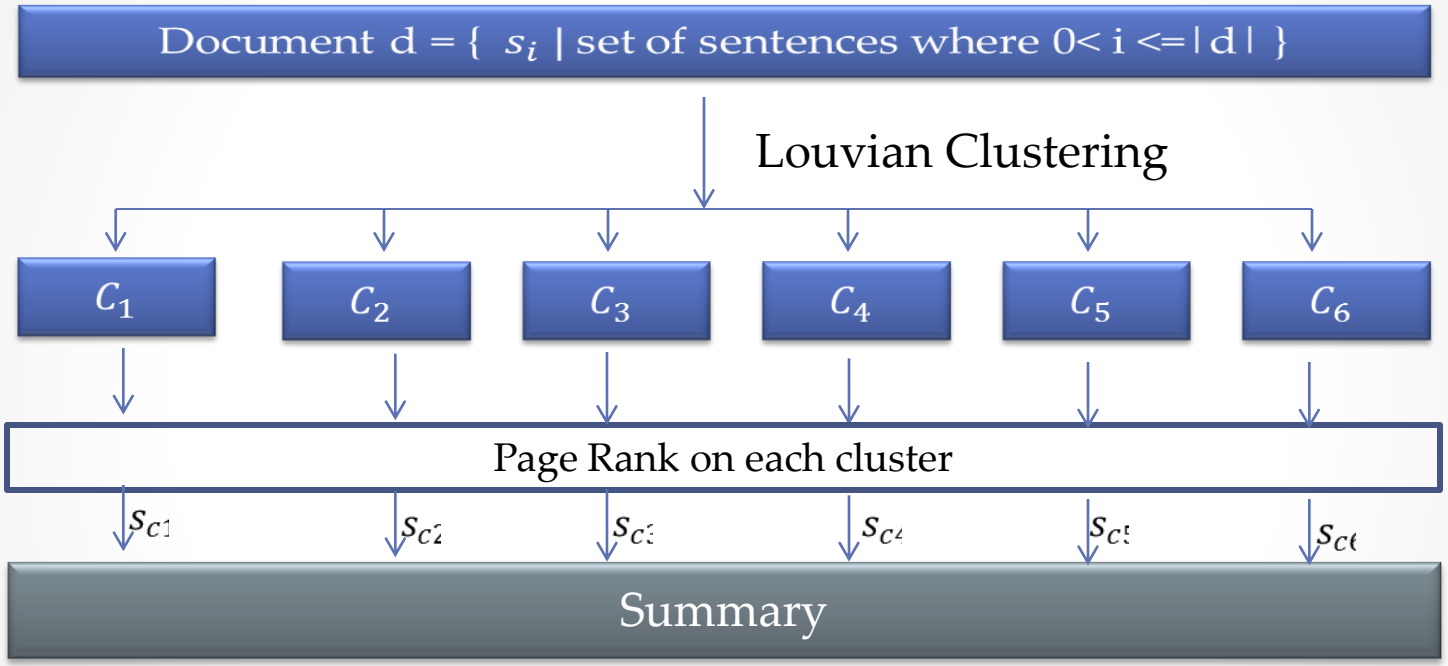
Methods Implementations

- KL Divergence
- Probsum Method
- **Clustering with Page Rank.**
- **Merging of summaries**
- LDA
- Doc2Vec with Louvain

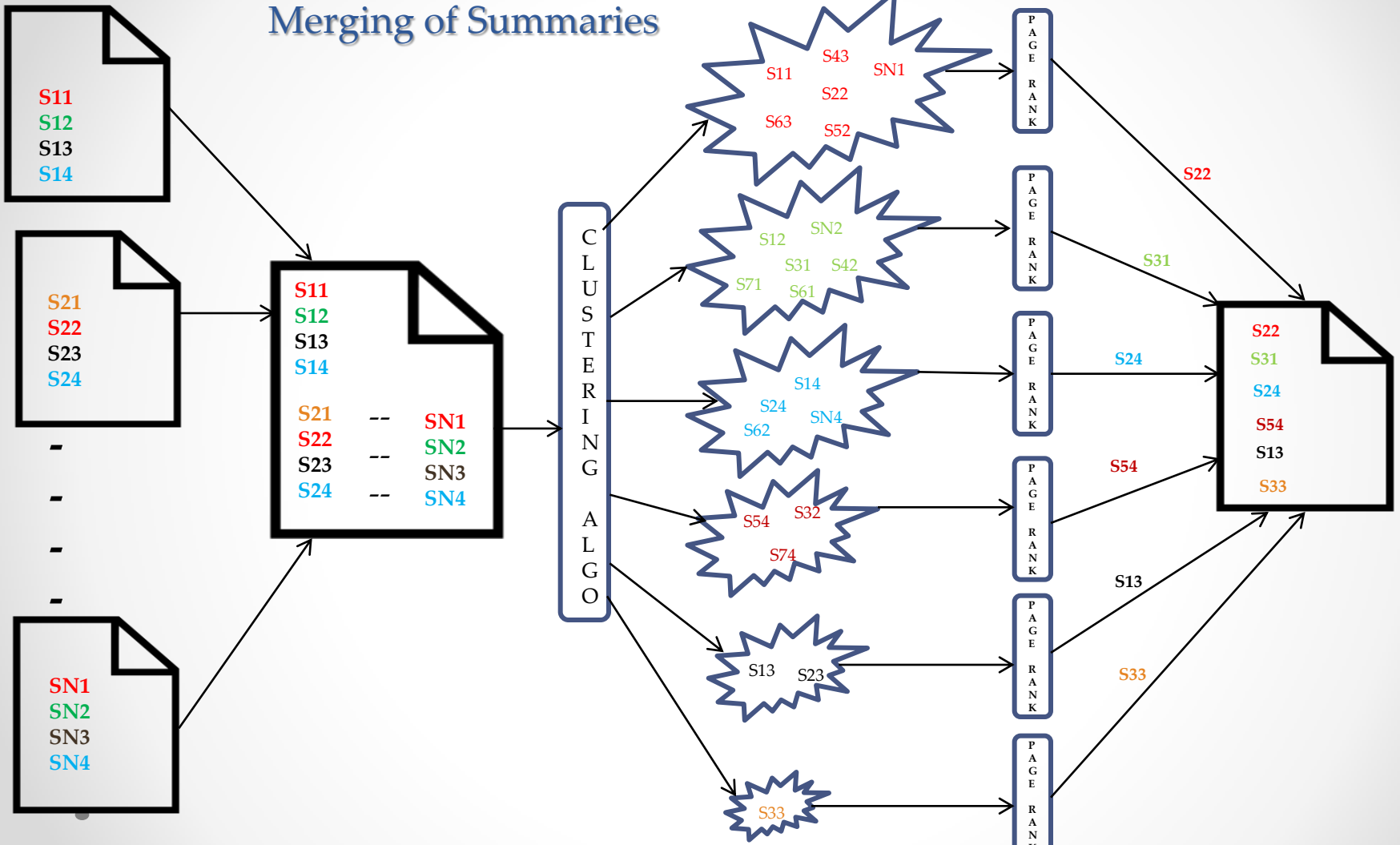
C-Lex Rank Algorithm(Hierarchical Clustering)



C-Lex Rank Algorithm(Louvian Clustering)



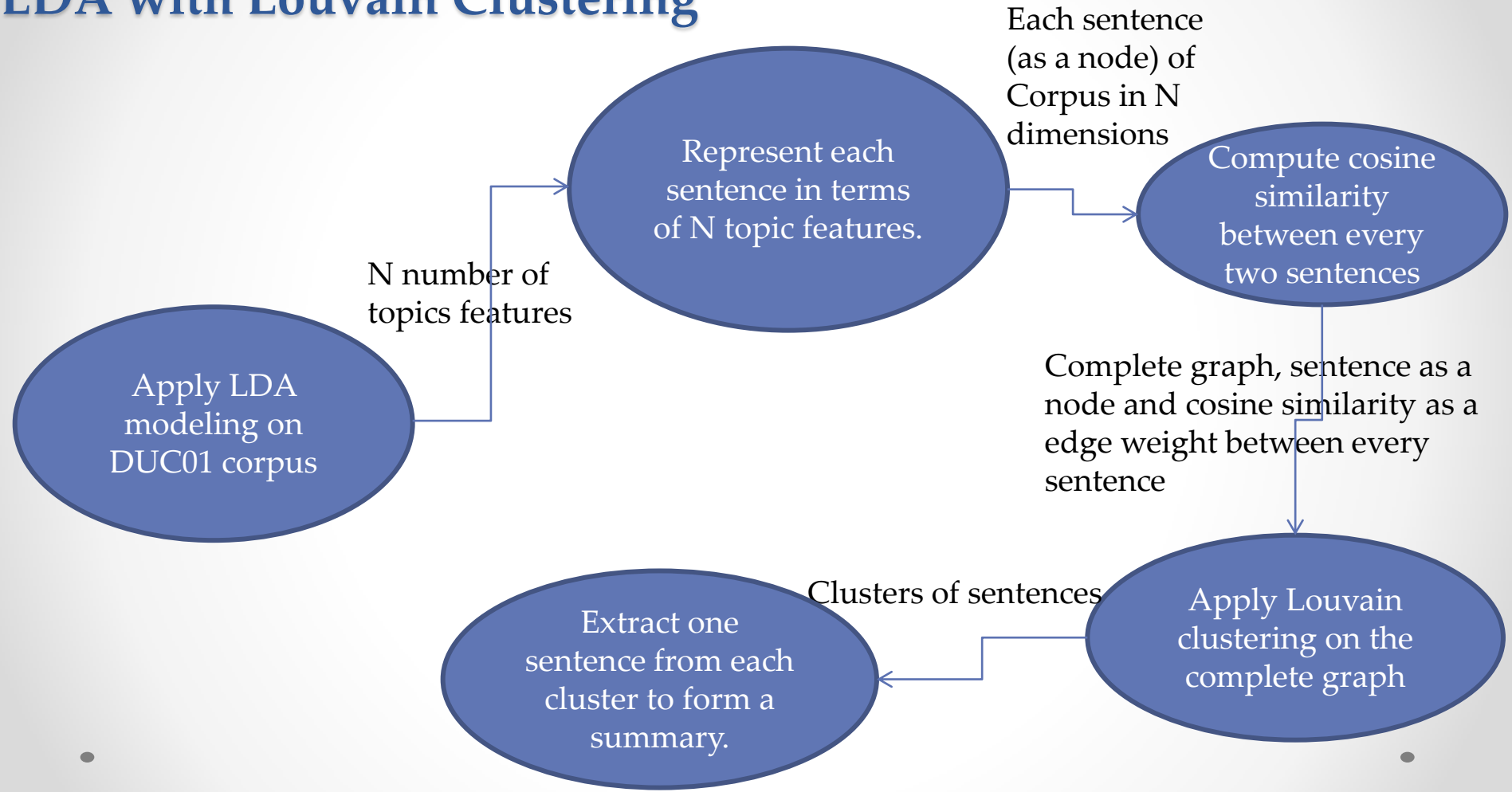
Merging of Summaries



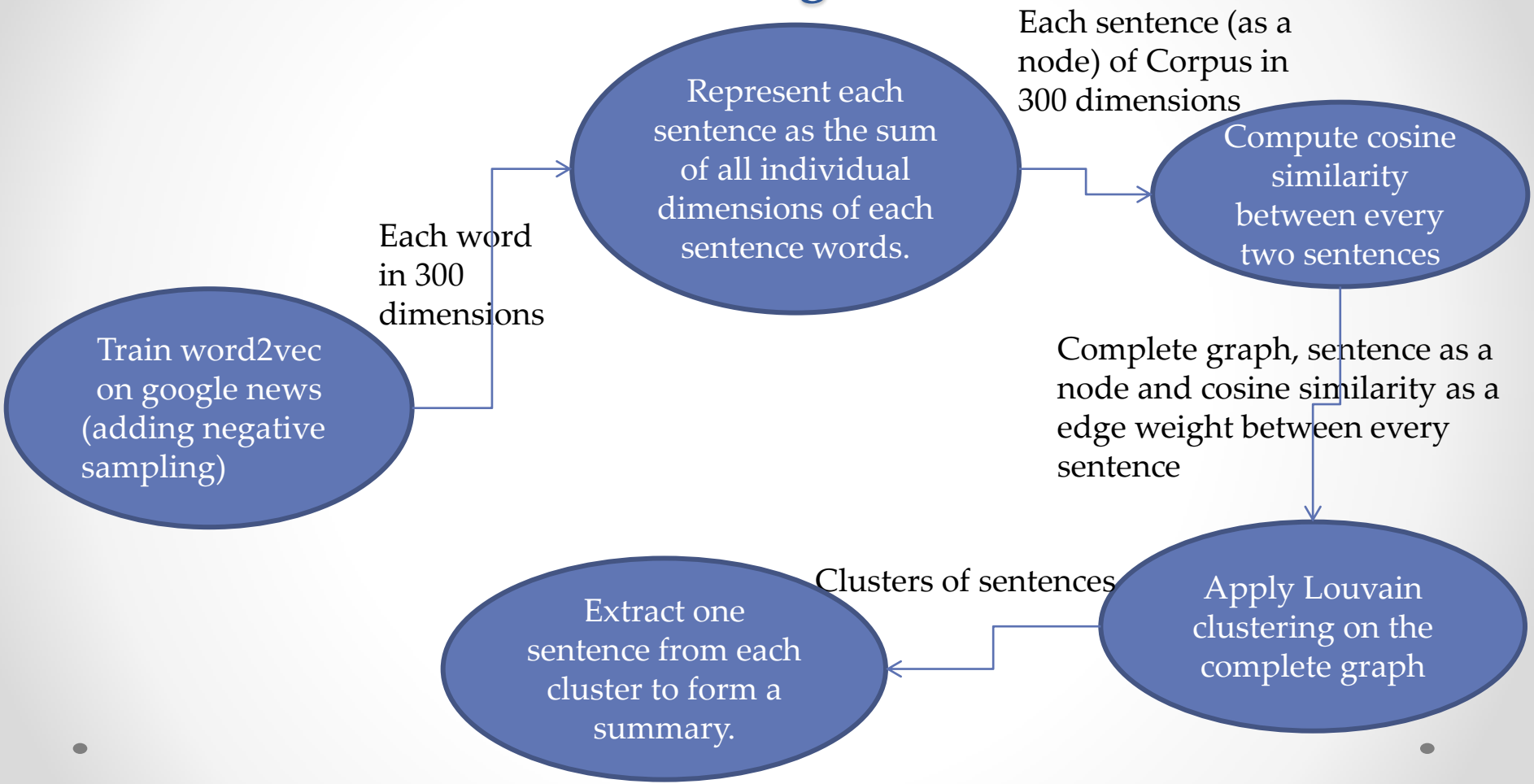
Observed Results:

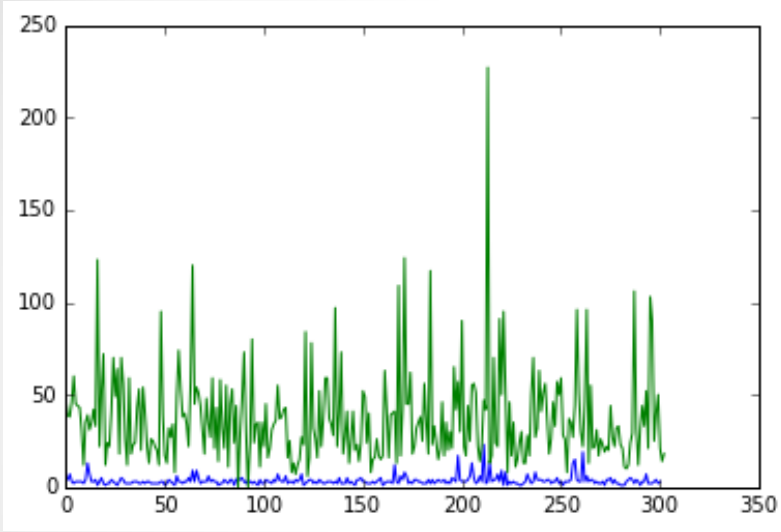
Method	Variations	Summary length	Recall	Precision	F1-Score
C-Lex Rank	6 Clusters	6 Sentences	0.557679	0.294905	0.376476
	4 Clusters	4 Sentences	0.495156	0.301213	0.365756
	6 Clusters	100	0.449767	0.366811	0.398644
	Normalized distance		0.42875	0.381393	0.398752
Merge	Without NNP(6 Clusters)	100	0.42082	0.411558	0.41271
	withNNP(6 Clusters)	100	0.420004	0.411214	0.412166

LDA with Louvain Clustering

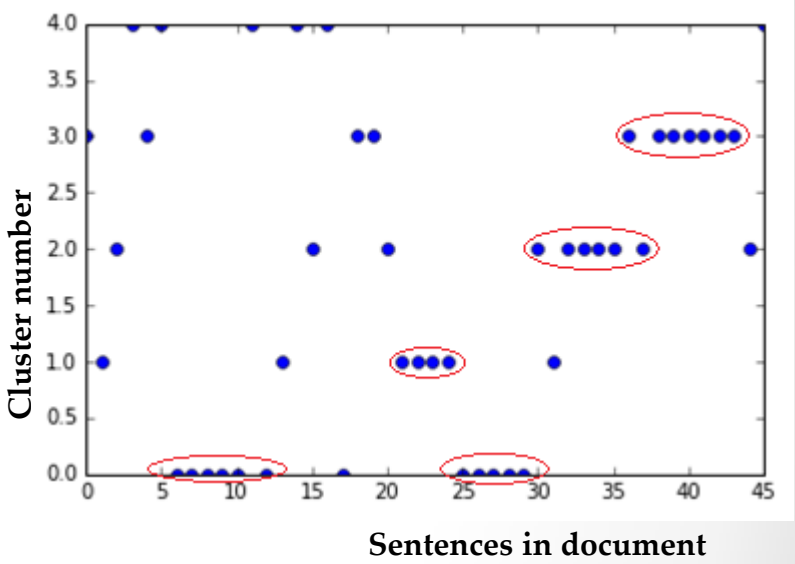


Word2Vec with Louvain Clustering





For each document, total number of sentences in document(green) and in summary(blue).

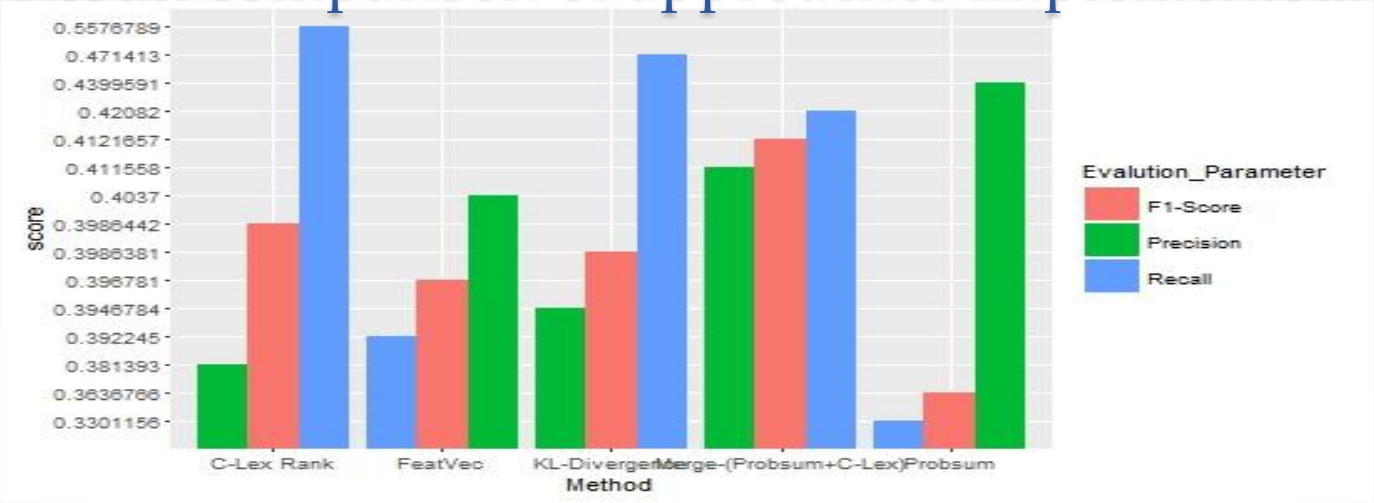


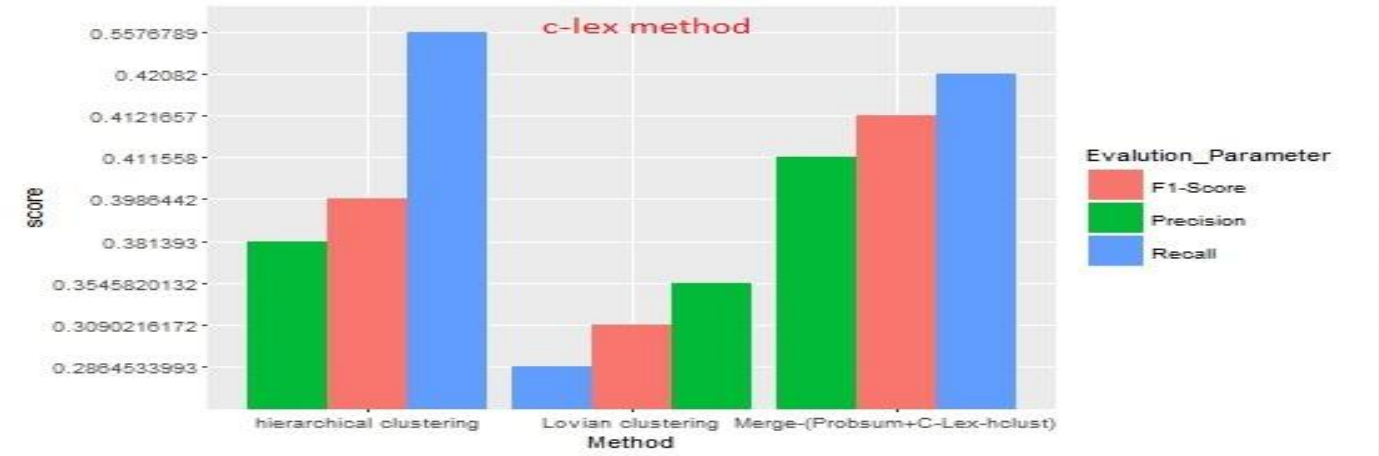
For a particular document, topical similarity by Louvain clustering.

Observed Results:

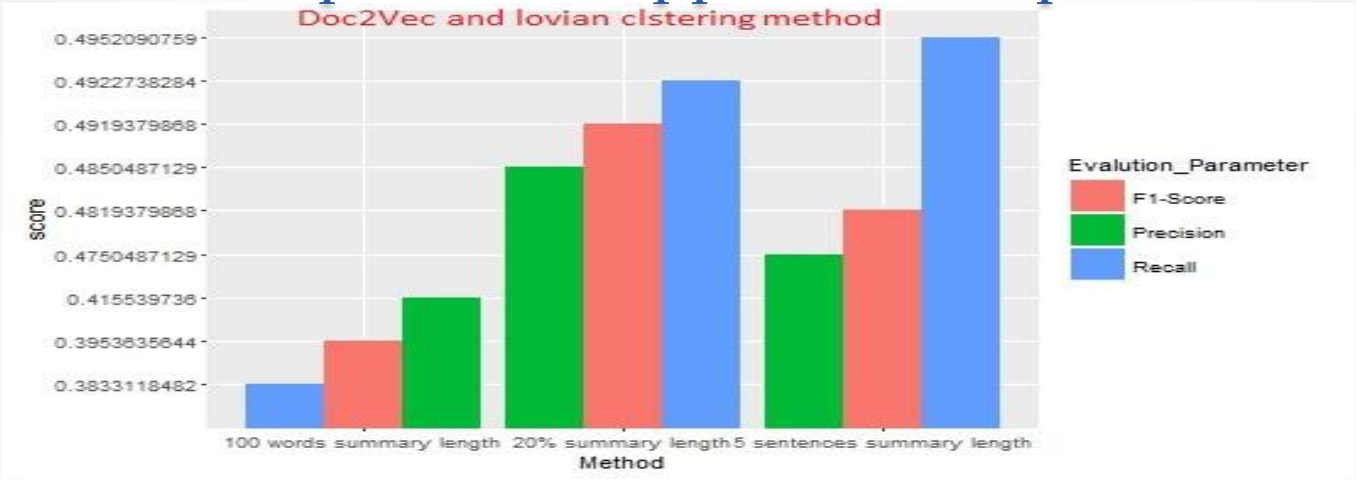
			Recall	Precision	F1 Score
Doc2Vec and lovian clustering		20%	0.4922738284	0.4850487129	0.4919379868
		5 scentences	0.4952090759	0.4750487129	0.4819379868
		100 Words	0.3833118482	0.415539736	0.3953635644
			Recall	Precision	F1 Score
Clex with Lovian		100	0.2864533993	0.3545820132	0.3090216172
			Recall	Precision	F1 Score
LDA		10 Topics - 100 words summary	0.3854543794	0.326488795	0.353529721
		20 Topics - 100 words summary	0.318519604	0.4817679208	0.4089458086
		30 Topics - 100 words summary	0.2503753135	0.4077232013	0.2924837294

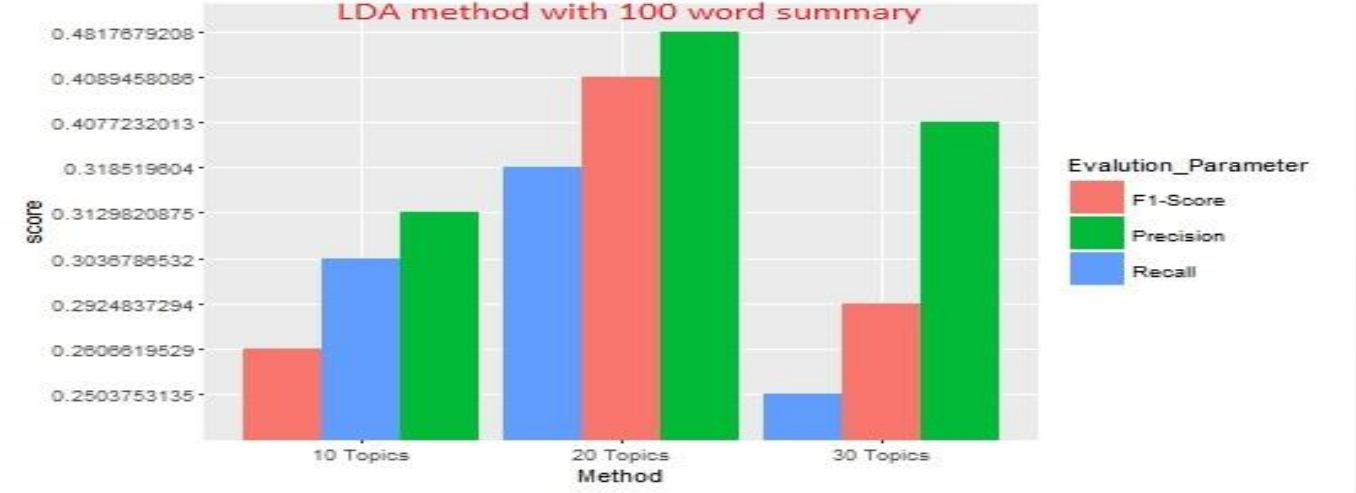
Result comparison of approaches implemented:



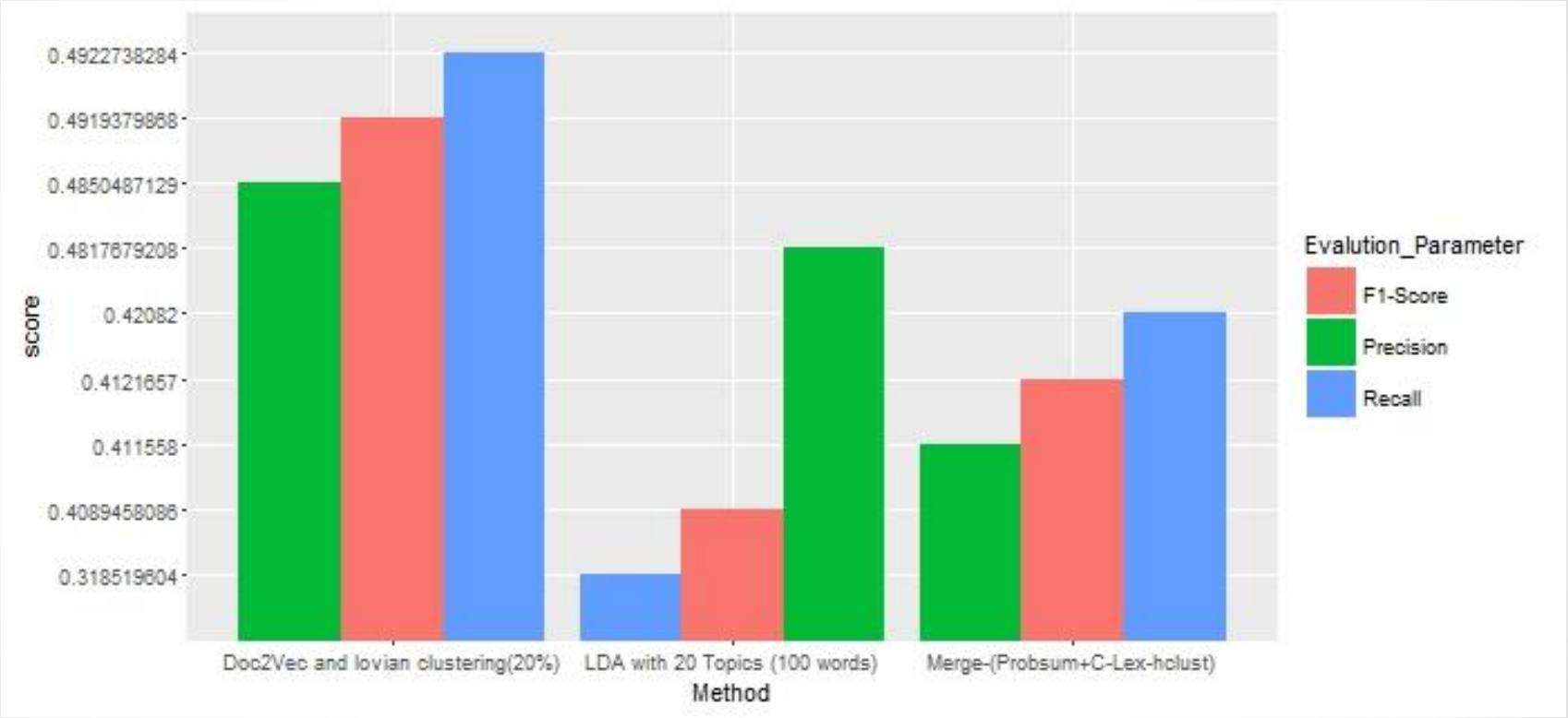


Result comparison of approaches implemented:





Result comparison of approaches implemented:



References

- **Aria Haghighi, Lucy Vanderwende**

Exploring Content Models for Multi-Document Summarization

Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pages 362–370,

http://www.aclweb.org/anthology/N09-1041

- **Vahed Qazvinian, Dragomir R. Radev, Saif M. Mohammad**

Generating Extractive Summaries of Scientific Paradigms

http://www-personal.umich.edu/~vahed/papers/iopener.pdf

- **Ani Nenkova, Lucy Vanderwende, Kathleen McKeown**

A Compositional Context Sensitive Multidocument Summarizer

http://www.cis.upenn.edu/~nenkova/papers/fp285-nenkova.pdf

Thank You