

# *Text Summarization - LexRank*

Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 1

## *What is a summary?*

A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (*Hovy, 2008*)

# Text Summarization

## *What is a summary?*

A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (*Hovy, 2008*)

## *What is text summarization?*

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task. (*Mani and MayBury, 2001*)

# Text Summarization

## *What is a summary?*

A summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). (Hovy, 2008)

## *What is text summarization?*

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user or task. (Mani and MayBury, 2001)

Humans have an incredible capacity to condense information down to the critical bit.

*“He said he is against it.”*

*Calvin Coolidge, on being asked what a clergyman preaching on sin said.*

## *Goal of a Text Summarization System*

To give an overview of the original document in a shorter period of time.

# Automatic Text Summarization

## Goal of a Text Summarization System

To give an overview of the original document in a shorter period of time.

## Summarization Applications

- *outlines or abstracts* of any document, news article etc.
- *summaries* of email threads
- *action items* from a meeting
- *simplifying* text by compressing sentences

# Application: Generating Snippets

## Robert O'Neill taking credit for killing Osama bin Laden sparks debate

Hindustan Times - 1 hour ago

Some special operations service members and veterans are unhappy that one of their own has taken credit publicly for killing Osama bin Laden.

## It's been special knock as wait has been long: Rayudu

Hindustan Times - 1 hour ago

An elated Ambati Rayudu said Friday that his maiden hundred in international cricket will certainly be a "special one" as it took a long time to come.

# Application: Generating Snippets

## Robert O'Neill taking credit for killing Osama bin Laden sparks debate

Hindustan Times - 1 hour ago

Some special operations service members and veterans are unhappy that one of their own has taken credit publicly for killing Osama bin Laden.

## It's been special knock as wait has been long: Rayudu

Hindustan Times - 1 hour ago

An elated Ambati Rayudu said Friday that his maiden hundred in international cricket will certainly be a "special one" as it took a long time to come.

what is the relation between pressure and velocity

Web

Images

Videos

News

More ▾

Search tools

About 1,10,00,000 results (0.49 seconds)

### fluid dynamics - Relation between pressure, velocity and ...

[physics.stackexchange.com/.../relation-between-pressure-velocity-and-ar...](http://physics.stackexchange.com/.../relation-between-pressure-velocity-and-ar...) ▾

In a nozzle, the exit **velocity** increases as per continuity equation as given by Bernoulli equation (incompressible fluid). **Pressure** is inversely proportional to ...

### Chapter 9: Fluid Dynamics

[francesa.phy.cmich.edu/people/andy/physics110/book/.../Chapter9.htm](http://francesa.phy.cmich.edu/people/andy/physics110/book/.../Chapter9.htm) ▾

From practical experience we know that the velocity of fluid through the small ... we found a qualitative **relationship between pressure and velocity** in a fluid flow.

### Bernoulli's Equation

[https://www.princeton.edu/~asmits/Bicycle\\_web/Bernoulli.html](https://www.princeton.edu/~asmits/Bicycle_web/Bernoulli.html) ▾

... can give great insight into the balance **between pressure, velocity** and elevation. ...  
When streamlines are parallel the **pressure** is constant across them, except ...

### Pressure Vs velocity | Student Doctor Network

[forums.studentdoctor.net](http://forums.studentdoctor.net) › ... › MCAT Study Question Q&A ▾

Jul 21, 2009 - 8 posts - 3 authors

Velocity increases with a decrease in pressure. Velocity... ... If you want to think of the **relationship between pressure and velocity**, you can use ...



## Genres of Summary

- Extract vs. Abstract  
*...lists fragments of text vs. re-phrases content coherently.*
- Single document vs. Multi-document  
*...based on one text vs. fuses together many texts.*
- Generic vs. Query-focused  
*...provides author's view vs. reflects user's interest.*

## Genres of Summary

- **Extract** vs. Abstract  
...*lists fragments of text vs. re-phrases content coherently.*
- **Single document** vs. Multi-document  
...*based on one text vs. fuses together many texts.*
- **Generic** vs. Query-focused  
...*provides author's view vs. reflects user's interest.*

Query-focused summarization can be thought of as a complex question answering system

# *Summarization: Main stages*

## *Content Selection*

Choose sentences to extract from the document

# *Summarization: Main stages*

## *Content Selection*

Choose sentences to extract from the document

## *Information Ordering*

Choose an order to place them in the summary

# *Summarization: Main stages*

## *Content Selection*

Choose sentences to extract from the document

## *Information Ordering*

Choose an order to place them in the summary

## *Sentence realization*

Simplify the sentences

# *Summarization: Main stages*

## *Content Selection*

Choose sentences to extract from the document

## *Information Ordering*

Choose an order to place them in the summary

## *Sentence realization*

Simplify the sentences

## *Removing Redundancy*

Increase diversification by removing redundant sentences

# Summarization: Main stages

## *Content Selection*

Choose sentences to extract from the document

## *Information Ordering*

Choose an order to place them in the summary

## *Sentence realization*

Simplify the sentences

## *Removing Redundancy*

Increase diversification by removing redundant sentences

The most basic algorithm only does the first stage, *content selection*.

# *Unsupervised content selection; Luhn (1958)*

## *Intuition*

Choose sentences that have salient or informative words



# Unsupervised content selection; Luhn (1958)

## Intuition

Choose sentences that have salient or informative words

## Two approaches to define salient words

- *tf-idf*: weigh each word  $w_i$  in document  $j$  by tf-idf

$$\text{weight}(w_i) = \text{tf}_{ij} \times \text{idf}_i$$

- *Topic signatures*: choose a smaller set of salient words, specific to that domain

$$\text{weight}(w_i) = 1 \text{ if } w_i \text{ is a specific term (use mutual information)}$$

# Unsupervised content selection; Luhn (1958)

## Intuition

Choose sentences that have salient or informative words

## Two approaches to define salient words

- *tf-idf*: weigh each word  $w_i$  in document  $j$  by *tf-idf*

$$\text{weight}(w_i) = \text{tf}_{ij} \times \text{idf}_i$$

- *Topic signatures*: choose a smaller set of salient words, specific to that domain

$$\text{weight}(w_i) = 1 \text{ if } w_i \text{ is a specific term (use mutual information)}$$

## Weighing a sentence

$$\text{weight}(s) = \frac{1}{|S|} \sum_{w \in S} \text{weight}(w)$$

# LexRank: A Graph-based approach

## Text Document

Computation is a process following a well defined model ...  
A computation can be seen as a purely physical phenomena ...  
...

processing

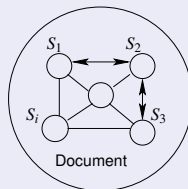
$S_1 \rightarrow \{(computation, 0.1), (process, 0.15), \dots\}$   
 $S_2 \rightarrow \{(computation, 0.1), (seen, 0.05), \dots\}$   
 $S_3 \rightarrow \dots$

Machine-readable format

## Document Representation

### Underlying Hypothesis

Sentences that convey the theme of the document are more similar to each other

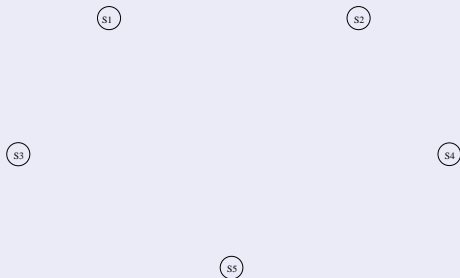


## Finding the most salient sentences

# Sentence Centrality Measure

## *Finding the most salient sentences*

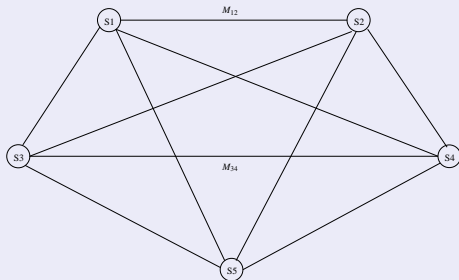
A document graph is constructed with sentences as the vertices



# Sentence Centrality Measure

## Finding the most salient sentences

A sentence similarity function is used to calculate the edge weights.

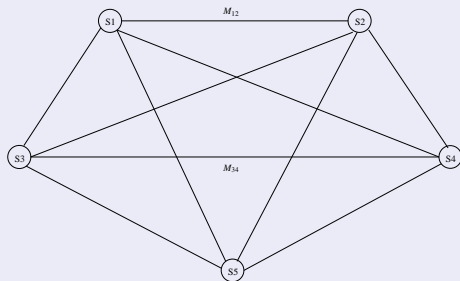


$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

# Sentence Centrality Measure

## Finding the most salient sentences

PageRank based algorithm is used to compute the sentence centrality vector  $I$ .



$$\tilde{M} = \begin{bmatrix} 0.0 & 0.5 & 0.0 & 0.4 & 0.1 \\ 0.5 & 0.0 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 & 0.0 \\ 0.4 & 0.0 & 0.4 & 0.0 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

$$I_j = \mu \cdot \sum_{\forall k \neq j} I_k \cdot \tilde{M}_{k,j} + \frac{1-\mu}{|S|}$$

$$I = \begin{bmatrix} 0.22 & 0.18 & 0.2 & 0.3 & 0.1 \end{bmatrix}$$

# Removing Redundant Sentences

## Maximal Marginal Relevance

- An iterative method for content selection from a selected list of important sentences
- Iteratively choose the best sentence to insert in the summary that is minimally redundant with the summary so far ( $Sum$ )

$$Inf(s)_{MMR} = \max_{s \in D} (Inf(s) - \lambda \cdot sim(s, Sum))$$

where  $Inf(s)$  denotes the informativeness score of a sentence