Information Extraction and Text Summarization

Randil Pushpananda

University of Colombo School of Computing

No 35, Reid Avenue, Colombo 07

rpn@ucsc.cmb.ac.lk

Natural Language Processing

Introduction

- Both IE and Text Summarization are important natural language processing (NLP) tasks that involve processing and extracting meaningful content from text documents.

Information extraction

- is the process of automatically extracting structured information from unstructured text data.

- It extracts the specified type of ***entity, relationship, time, event, and other factual information*** from natural language text and forms a structured data output.

Information Extraction

- focuses on extracting structured information from unstructured text data.
- Aims to identify specific pieces of information within the text and convert them into a structured format, such as a database or a knowledge graph.

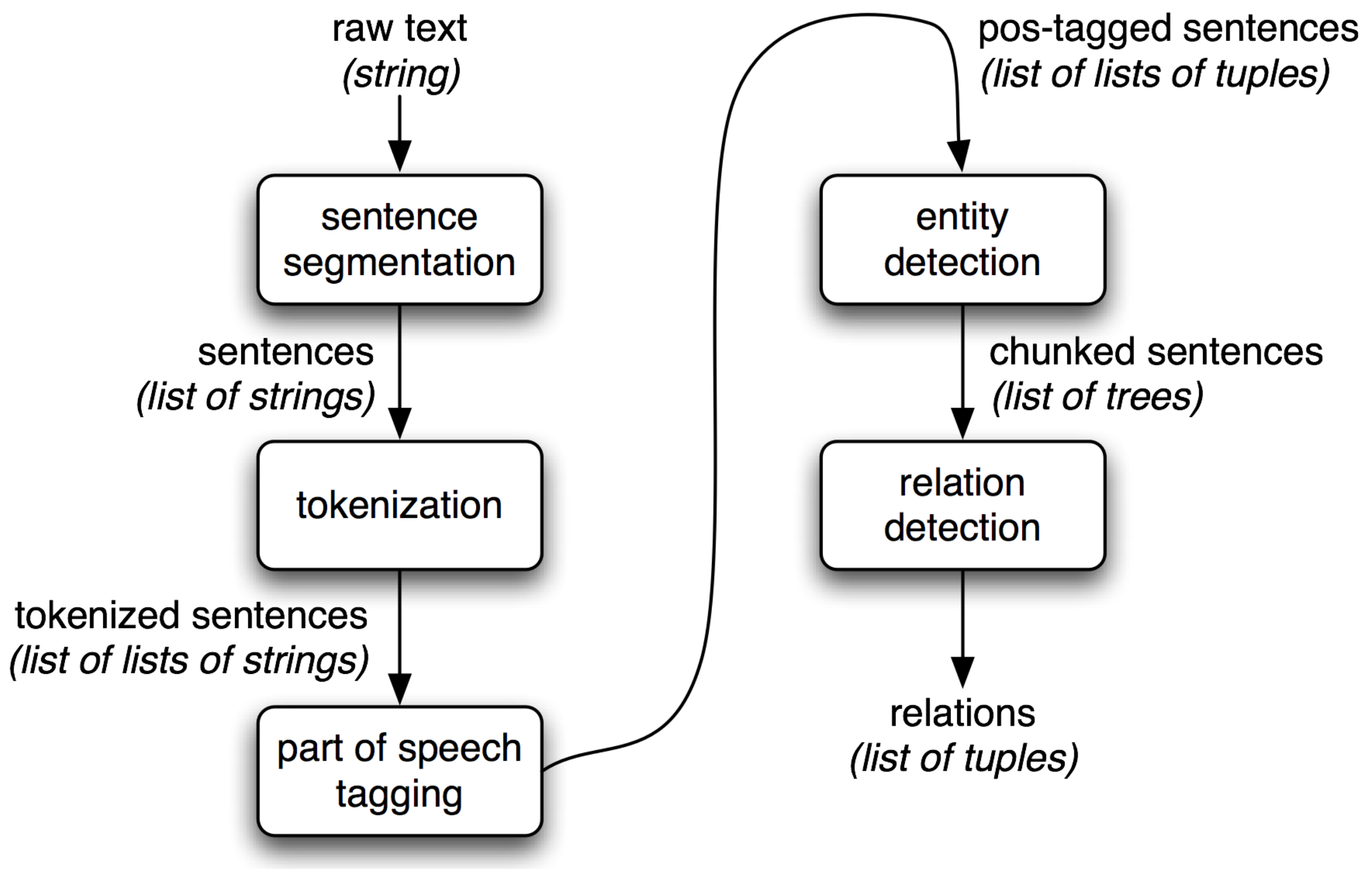
Information Retrieval

- The process of finding relevant documents or pieces of information from a large collection of documents in response to a user's query.
- About searching for and retrieving documents that match a user's information needs.

Information Extraction

- Usually from unstructured or semi-structured data
- Examples for unstructured datasets:
  - News Stories
  - Scientific Papers
  - Resumes
- Try to extract Entities, Relationships and finally try to build a Knowledge Base.
  - Who did, What, When, Where and Why

Pipeline Architecture for an IE System



raw text
(string)

sentence segmentation

sentences
(list of strings)

tokenization

tokenized sentences
(list of lists of strings)

part of speech tagging

pos-tagged sentences
(list of lists of tuples)

entity detection

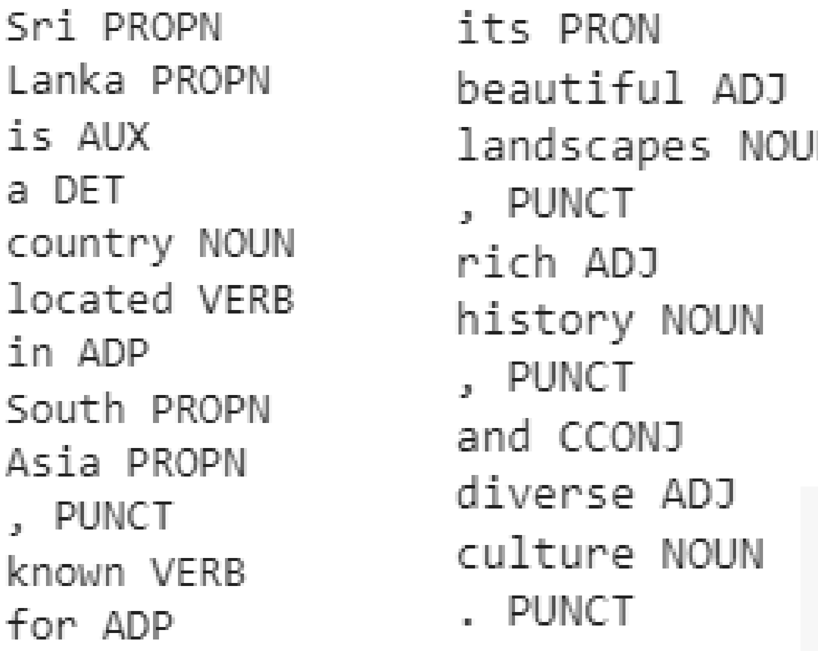
chunked sentences
(list of trees)

relation detection

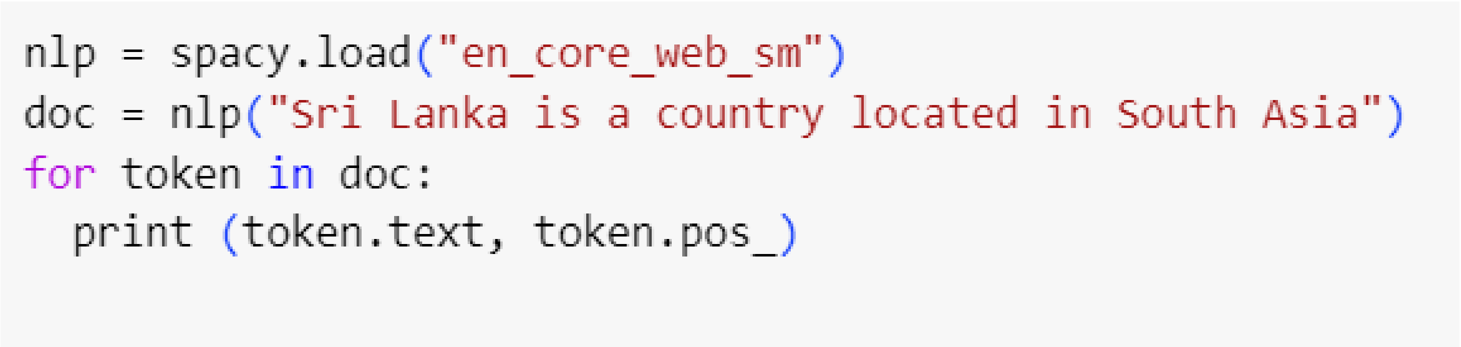
relations
(list of tuples)

Tokenization and POS Tagging

- Sri Lanka is a country located in South Asia, known for its beautiful landscapes, rich history, and diverse culture.



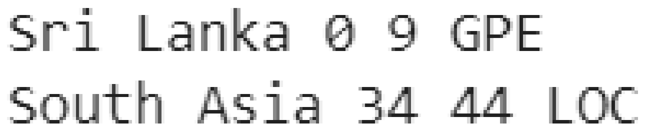
```
Sri PROPN
Lanka PROPN
is AUX
a DET
country NOUN
located VERB
in ADP
South PROPN
Asia PROPN
, PUNCT
known VERB
for ADP
its PRON
beautiful ADJ
landscapes NOUN
, PUNCT
rich ADJ
history NOUN
, PUNCT
and CCONJ
diverse ADJ
culture NOUN
. PUNCT
```



```
nlp = spacy.load("en_core_web_sm")
doc = nlp("Sri Lanka is a country located in South Asia")
for token in doc:
  print (token.text, token.pos_)
```

Entity Detection (NER)

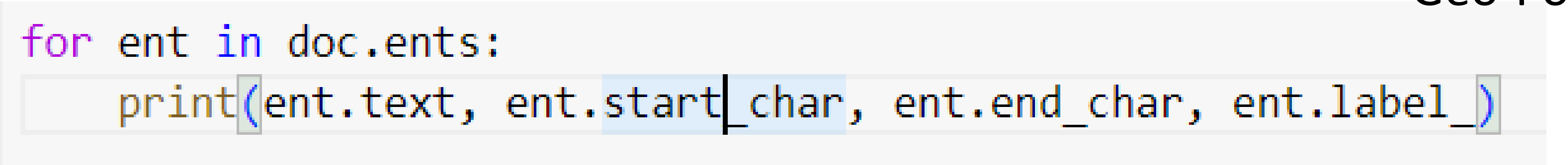
- Sri Lanka is a country located in South Asia, known for its beautiful landscapes, rich history, and diverse culture.



```
Sri Lanka 0 9 GPE
South Asia 34 44 LOC
```

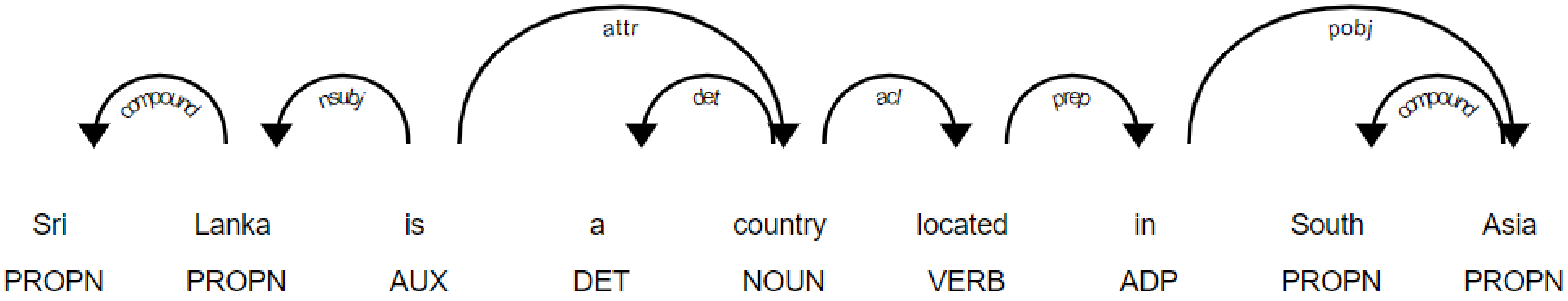
NE Types

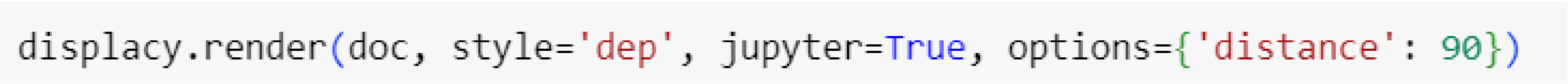
- People
- Locations
- Organizations
    - Teams
    - Newspapers
    - Company
- Geo-Political Entities



```
for ent in doc.ents:
    print(ent.text, ent.start_char, ent.end_char, ent.label_)
```

Relation Extraction (Subject and Object)

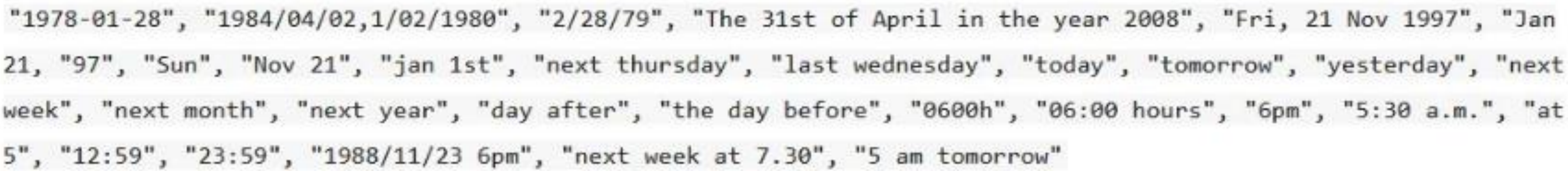




```
displacy.render(doc, style='dep', jupyter=True, options={'distance': 90})
```

Times and Events

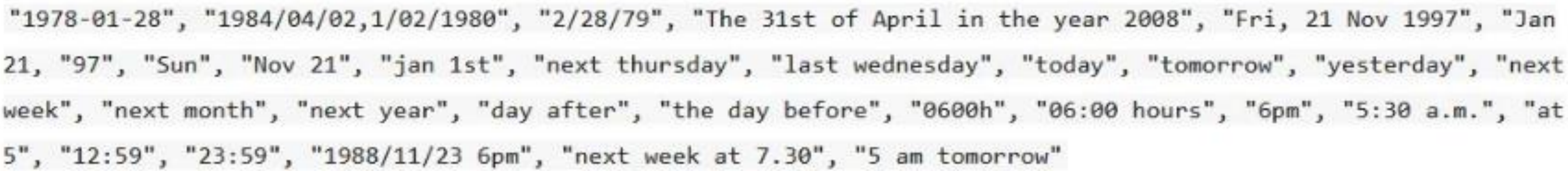
- Times
  - Absolute Expressions
  - Relative Expressions



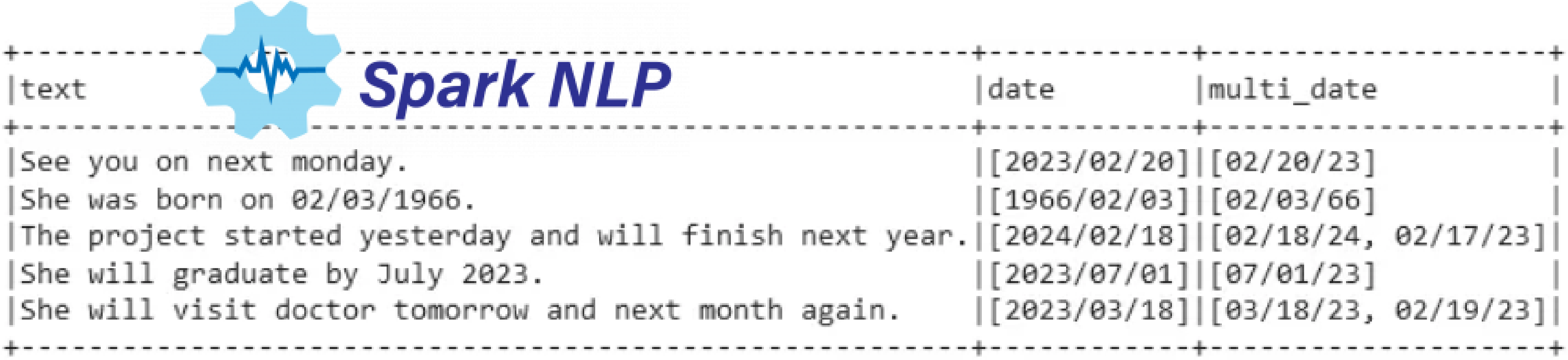
```
"1978-01-28", "1984/04/02,1/02/1980", "2/28/79", "The 31st of April in the year 2008", "Fri, 21 Nov 1997", "Jan
21, "97", "Sun", "Nov 21", "jan 1st", "next thursday", "last wednesday", "today", "tomorrow", "yesterday", "next
week", "next month", "next year", "day after", "the day before", "0600h", "06:00 hours", "6pm", "5:30 a.m.", "at
5", "12:59", "23:59", "1988/11/23 6pm", "next week at 7.30", "5 am tomorrow"
```

Times and Events

- Times
  - Absolute Expressions
  - Relative Expressions



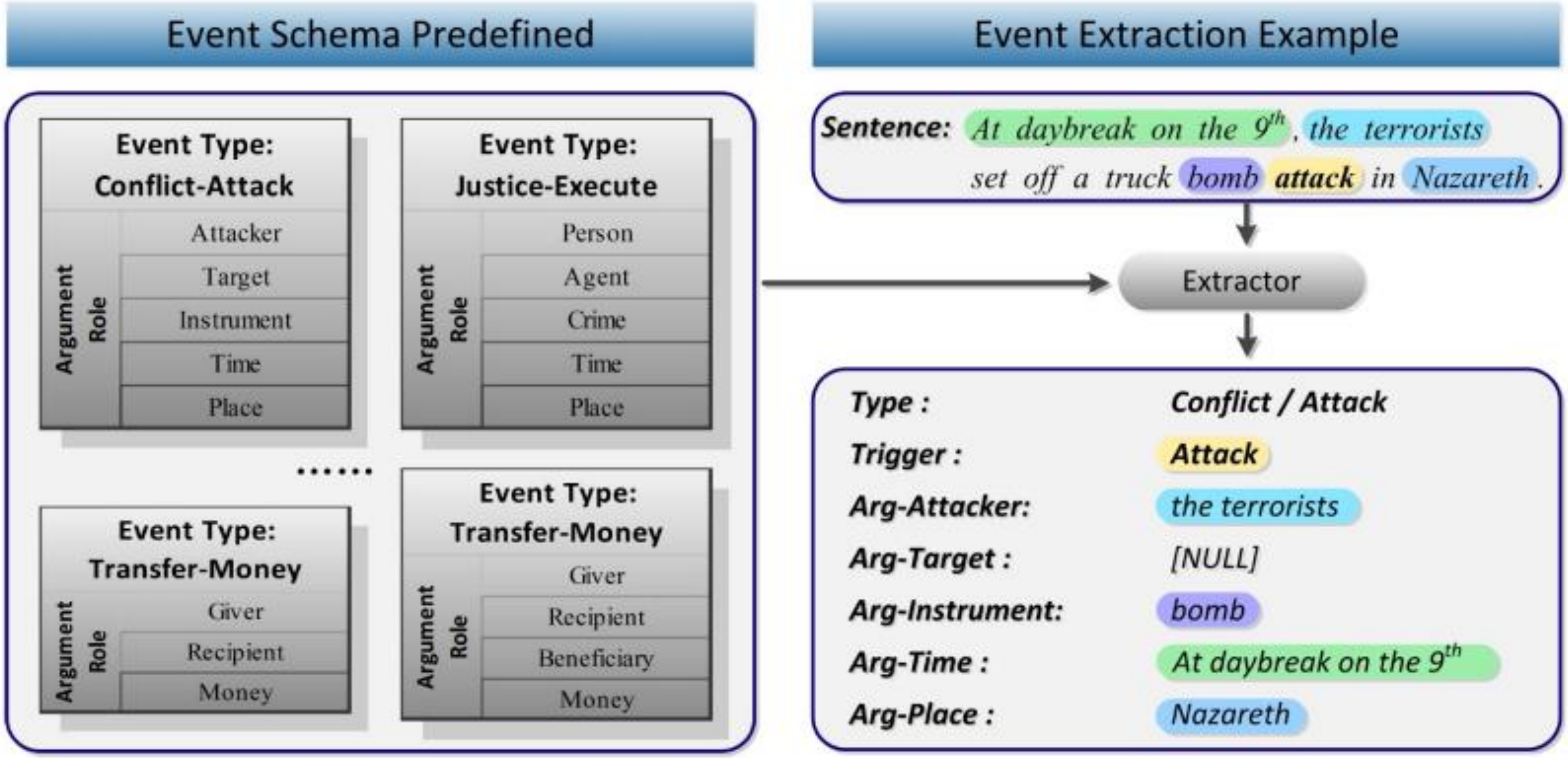
```
"1978-01-28", "1984/04/02,1/02/1980", "2/28/79", "The 31st of April in the year 2008", "Fri, 21 Nov 1997", "Jan
21, "97", "Sun", "Nov 21", "jan 1st", "next thursday", "last wednesday", "today", "tomorrow", "yesterday", "next
week", "next month", "next year", "day after", "the day before", "0600h", "06:00 hours", "6pm", "5:30 a.m.", "at
5", "12:59", "23:59", "1988/11/23 6pm", "next week at 7.30", "5 am tomorrow"
```



Spark NLP

```
+--------------------------------------------------------+------------+--------------------+
|text                                                    |date        |multi_date          |
+--------------------------------------------------------+------------+--------------------+
|See you on next monday.                                 |[2023/02/20]|[02/20/23]          |
|She was born on 02/03/1966.                             |[1966/02/03]|[02/03/66]          |
|The project started yesterday and will finish next year.|[2024/02/18]|[02/18/24, 02/17/23]|
|She will graduate by July 2023.                         |[2023/07/01]|[07/01/23]          |
|She will visit doctor tomorrow and next month again.    |[2023/03/18]|[03/18/23, 02/19/23]|
+--------------------------------------------------------+------------+--------------------+
```

Events



Event Schema Predefined

Event Type: Conflict-Attack

Argument Role: Attacker, Target, Instrument, Time, Place

Event Type: Justice-Execute

Argument Role: Person, Agent, Crime, Time, Place

......

Event Type: Transfer-Money

Argument Role: Giver, Recipient, Money

Event Type: Transfer-Money

Argument Role: Giver, Recipient, Beneficiary, Money

Event Extraction Example

Sentence: *At daybreak on the 9th, the terrorists set off a truck bomb attack in Nazareth.*

Extractor

Type :	*Conflict / Attack*
Trigger :	*Attack*
Arg-Attacker:	*the terrorists*
Arg-Target :	*[NULL]*
Arg-Instrument:	*bomb*
Arg-Time :	*At daybreak on the 9th*
Arg-Place :	*Nazareth*

An example of closed-domain event extraction, taken from Figure 1 by Xiang and Wang (2019)[1]

Sequence Labeling

- Many NLP problems can be considered as sequence labeling problems
  - POS – Part of Speech Tagging
  - NER – Named Entity Recognition
  - SRL – Semantic Role Labeling

- INPUT:
  - Sequence $W_1W_2W_3$
- OUTPUT:
  - Labeled Words

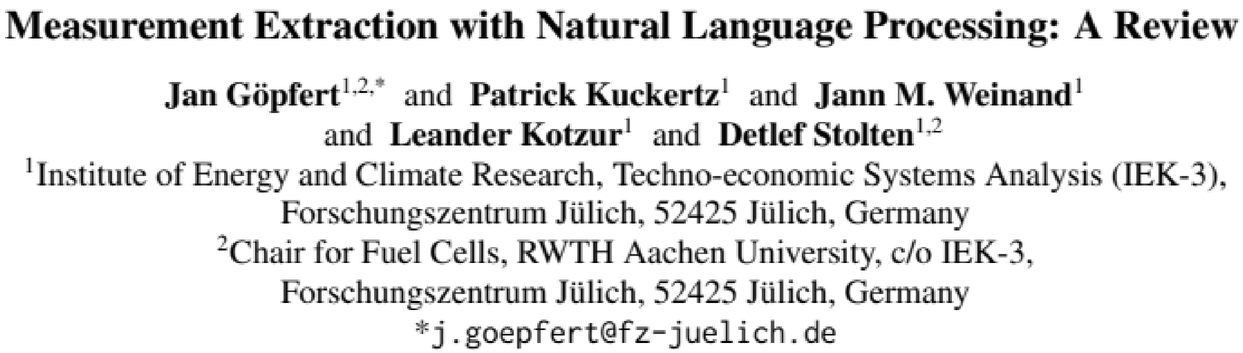
Sequence Labeling

- Many NLP problems can be considered as sequence labeling problems
  - POS – Part of Speech Tagging
  - NER – Named Entity Recognition
  - SRL – Semantic Role Labeling

- INPUT:
  - Sequence $W_1W_2W_3$
- OUTPUT:
  - Labeled Words

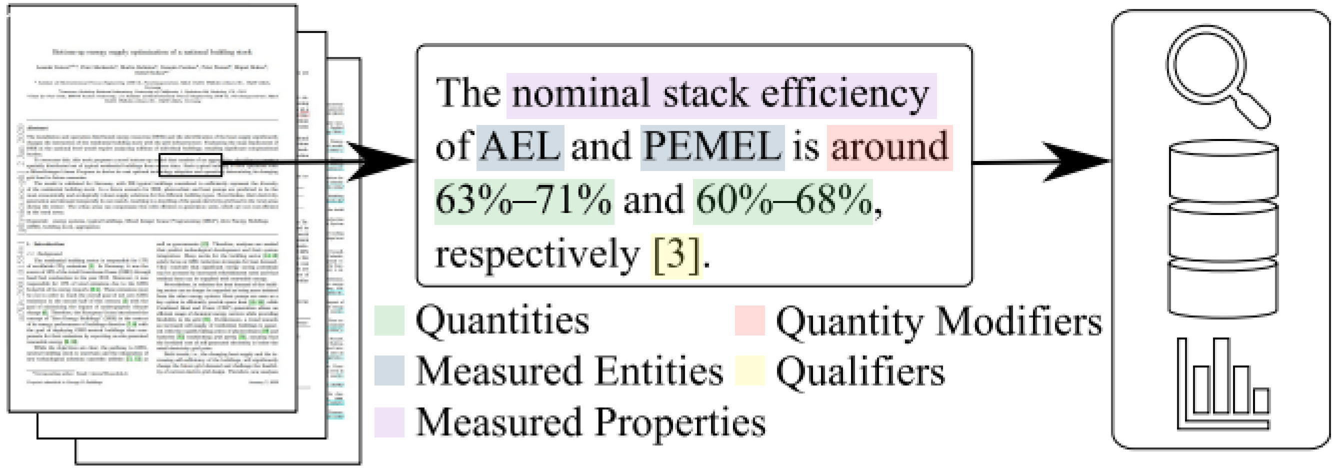
- Classification Methods:
  - Can use categories of previous tokens as features in classifying the next one

Extracting Quantitative data [New Trend]



Measurement Extraction with Natural Language Processing: A Review

Jan Göpfert[1,2,*] and **Patrick Kuckertz**[1] and **Jann M. Weinand**[1]
and **Leander Kotzur**[1] and **Detlef Stolten**[1,2]
[1]Institute of Energy and Climate Research, Techno-economic Systems Analysis (IEK-3),
Forschungszentrum Jülich, 52425 Jülich, Germany
[2]Chair for Fuel Cells, RWTH Aachen University, c/o IEK-3,
Forschungszentrum Jülich, 52425 Jülich, Germany
*j.goepfert@fz-juelich.de



The nominal stack efficiency of AEL and PEMEL is around 63%–71% and 60%–68%, respectively [3].

Quantity Extraction

- is the task of identifying quantities.
- A quantity (e.g., ‘1 kg’) is composed of a numeric value and, if applicable, a unit.

✓Numeric numbers (e.g., ‘27’)

✓Alphabetic number (e.g., ‘twenty-seven’)

✓Combinations (e.g., ‘2 million’)

✓Imprecise quantities (e.g., ‘a couple’)

✓Constants (e.g., ‘room temperature’ or ‘speed of light’)

Measurement Extraction

- adds to the identification of quantities by extracting their related measured properties and measured entities
- A measured property might be given implicitly.

✓Nominal Stack Efficiency

Quantity Extraction - Research Paper

Text summarization

- is the process of shortening lengthy textual content with the aim of producing a concise and coherent summary that highlights the document's key points.
- Why Text Summarization?
  - Pain point: information overload!
  - Key question: how to focus on the important (and avoid the ‘noise’)?
  - Crucial fact: Time available constant!
  - Strong focus on summarization: From executive summaries and abstracts to elevator pitches

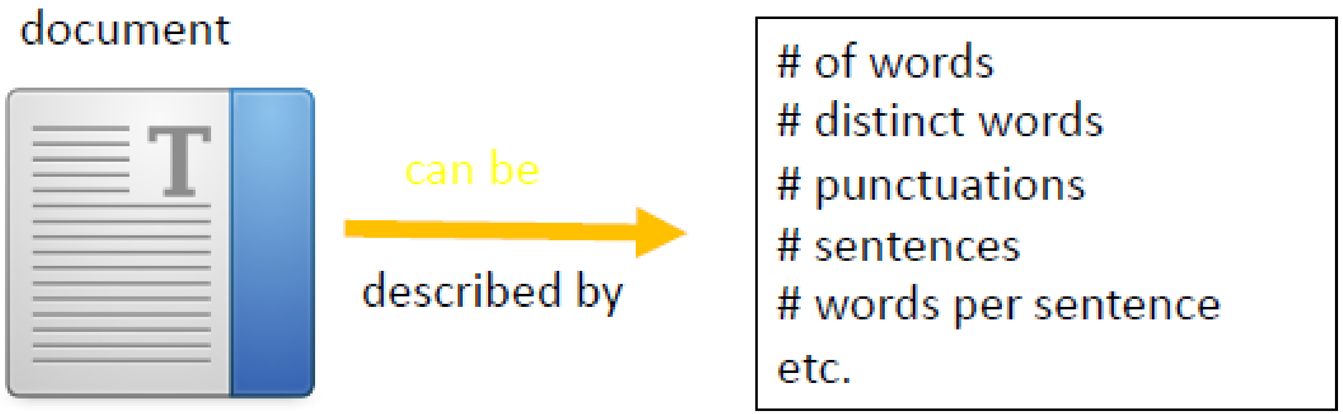
Exploring Text

How can we explore text?

Mean?

Standard Deviation?

No, but...



document

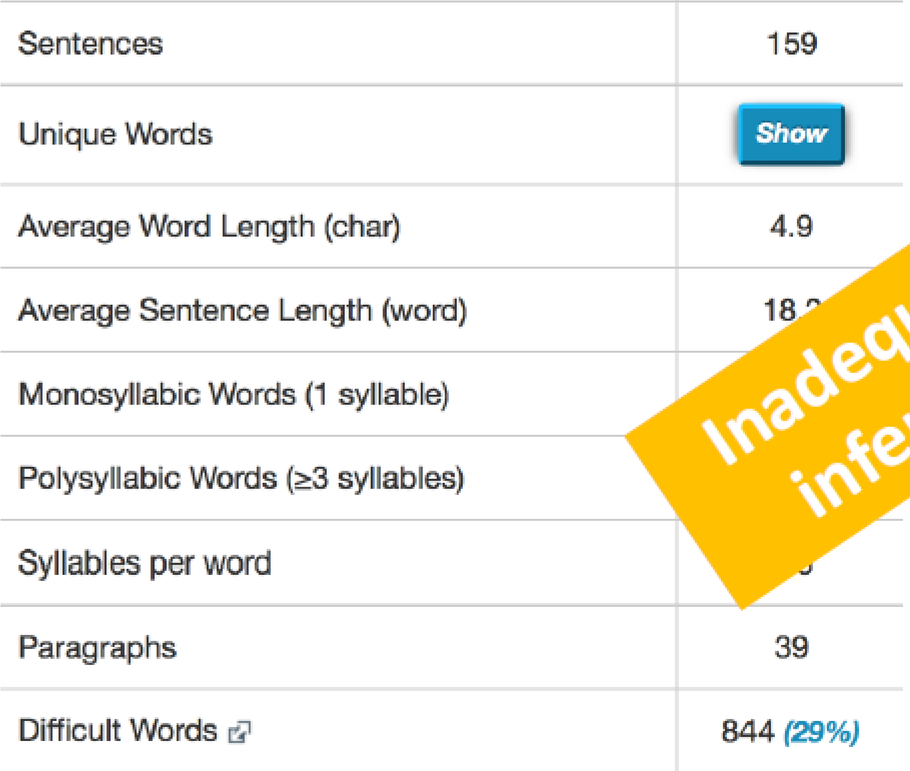
can be

described by

# of words
distinct words
punctuations
sentences
words per sentence
etc.

Exploring Text

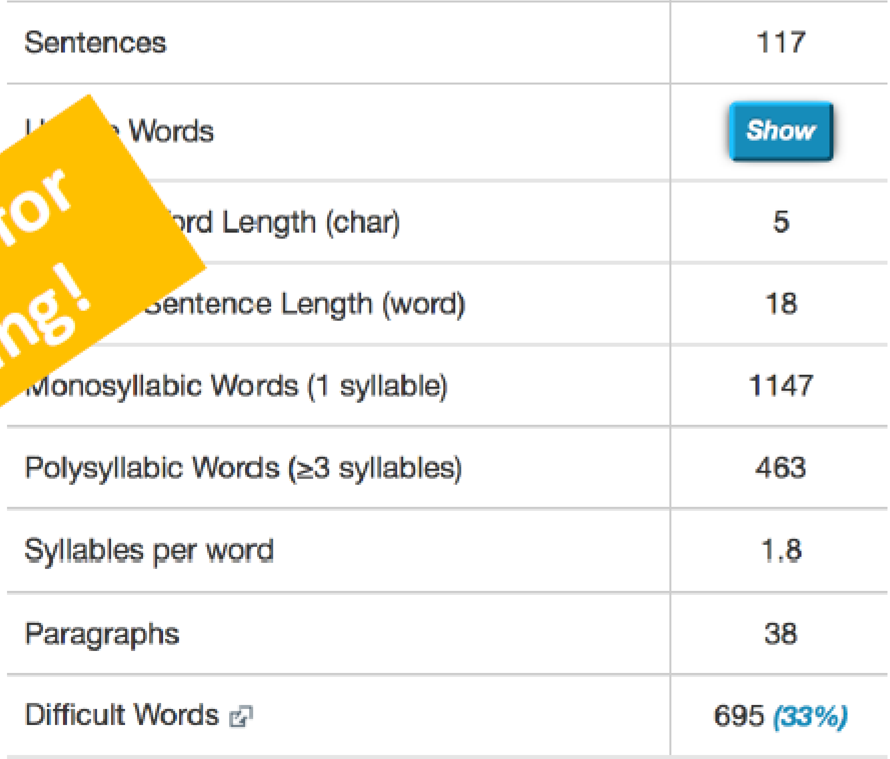
Example - description



Sentences	159
Unique Words	Show
Average Word Length (char)	4.9
Average Sentence Length (word)	18.[illegible]
Monosyllabic Words (1 syllable)	[illegible]
Polysyllabic Words (≥3 syllables)	[illegible]
Syllables per word	[illegible]
Paragraphs	39
Difficult Words	844 *(29%)*

MS 100 day speech
(2900 words)

Inadequate for inferencing!

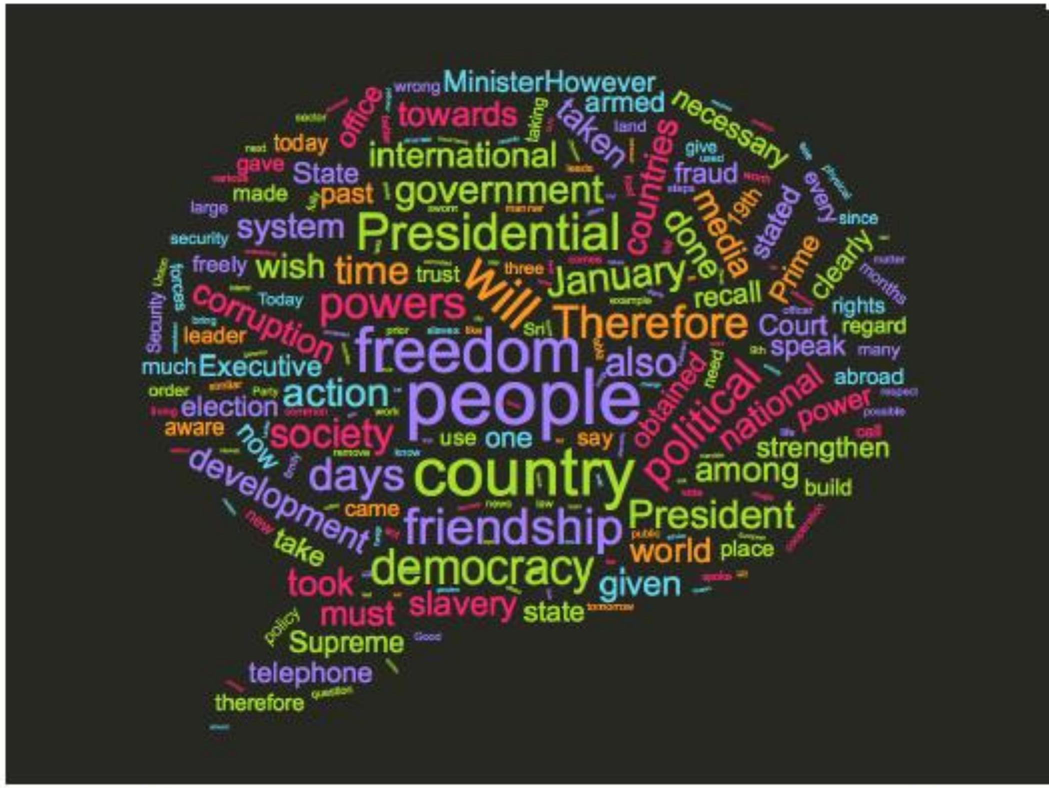


Sentences	117
U[illegible] Words	Show
[illegible]ord Length (char)	5
[illegible] Sentence Length (word)	18
Monosyllabic Words (1 syllable)	1147
Polysyllabic Words (≥3 syllables)	463
Syllables per word	1.8
Paragraphs	38
Difficult Words	695 *(33%)*

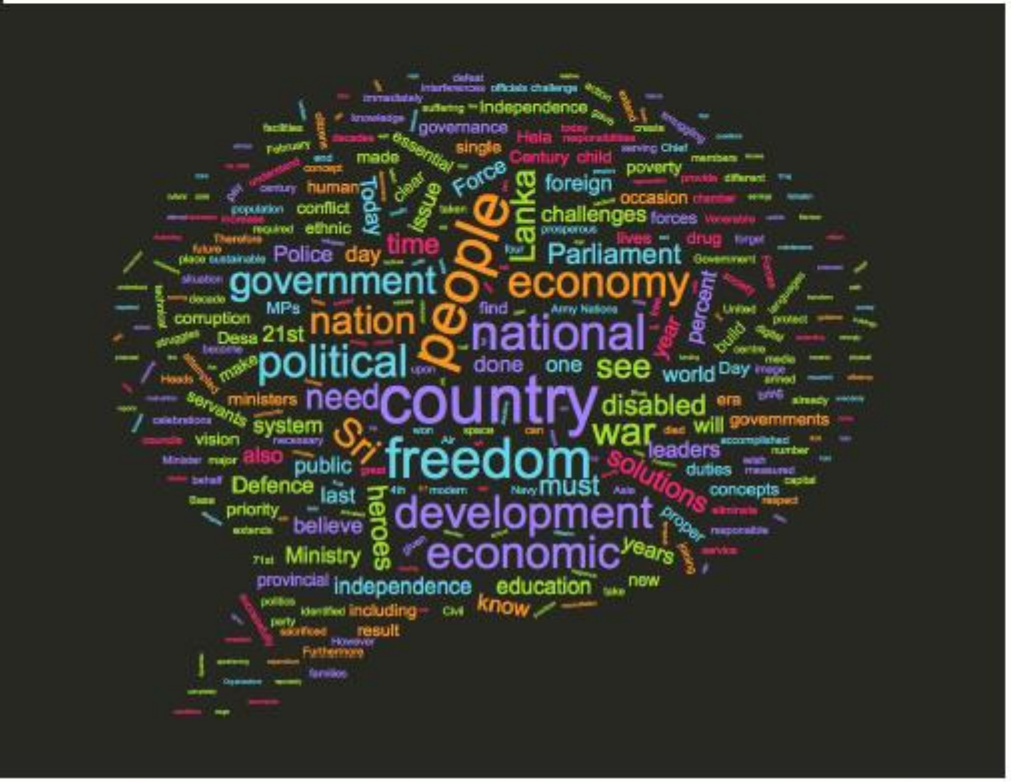
MS 71 independence day speech
(2100 words)

Exploring Text

Example - visualization



MS 100 day speech
(2900 words)



MS 71 independence day speech
(2100 words)

Automatic Text Summarization (ATS)

- Huge amount of news articles, Scientific papers, legal documents, etc.
- Manual text summarization consumes a lot of time, effort, cost, and even becomes impractical with the gigantic amount of textual content.
- Researchers have been trying to improve ATS techniques since the 1950s

Approaches

- Extractive
  - The extractive approach selects the most important sentences in the input document(s) then concatenates them to form the summary.
- Abstractive
  - The abstractive approach represents the input document(s) in an intermediate representation then generates the summary with sentences that are different than the original sentences.
- Hybrid
  - The hybrid approach combines both the extractive and abstractive approaches.

Main Objective of ATS

- To generate a concise summary of the key concepts from the input document while minimizing redundancy.
- The produced summary should be shorter in length than the input text and include the most important information in the input text.
- ATS systems can be classified as single-document or multi-document summarization systems.
- The former produces the summary from a single document while the latter generates the summary from a cluster of documents.

Architecture of an ATS system

Automatic text summarization: A comprehensive survey By Wafaa S. El-Kassas et.al.(2020)

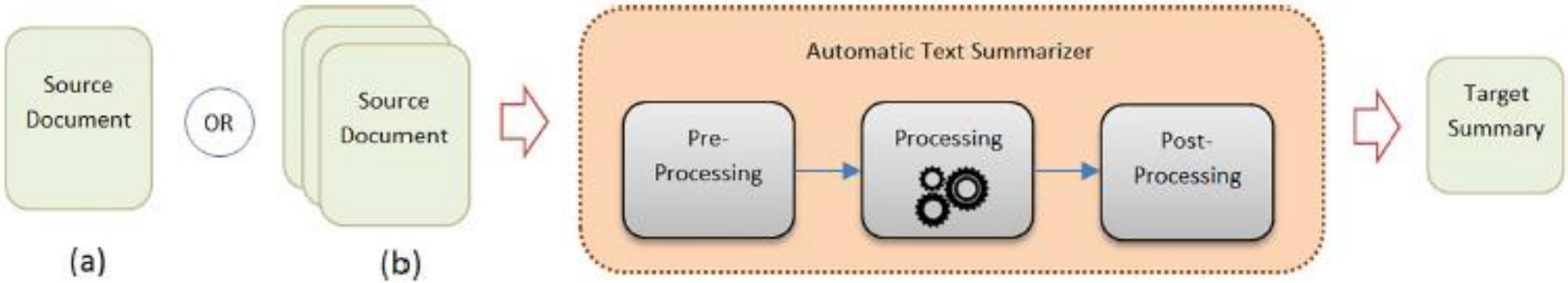
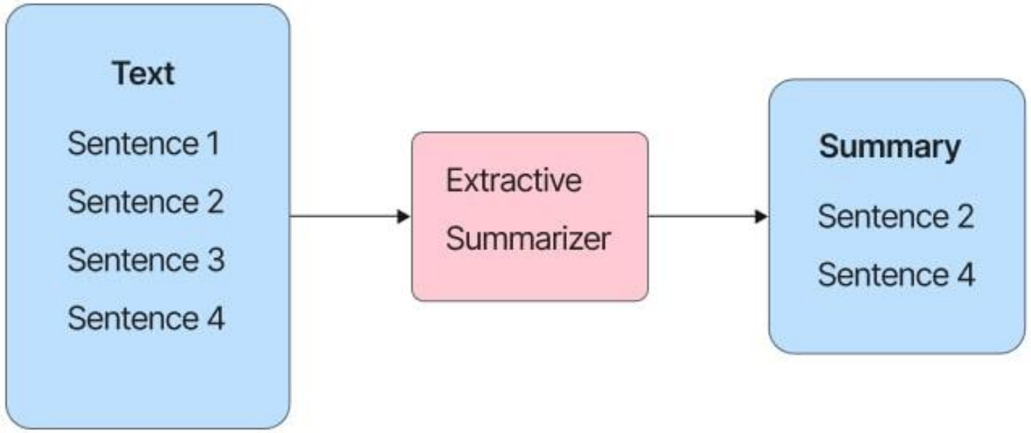


Fig. 1. (a) Single-document or (b) Multi-document, automatic text summarizer.

- Pre-Processing
- Processing
- Post-Processing

Extraction Based Summarization

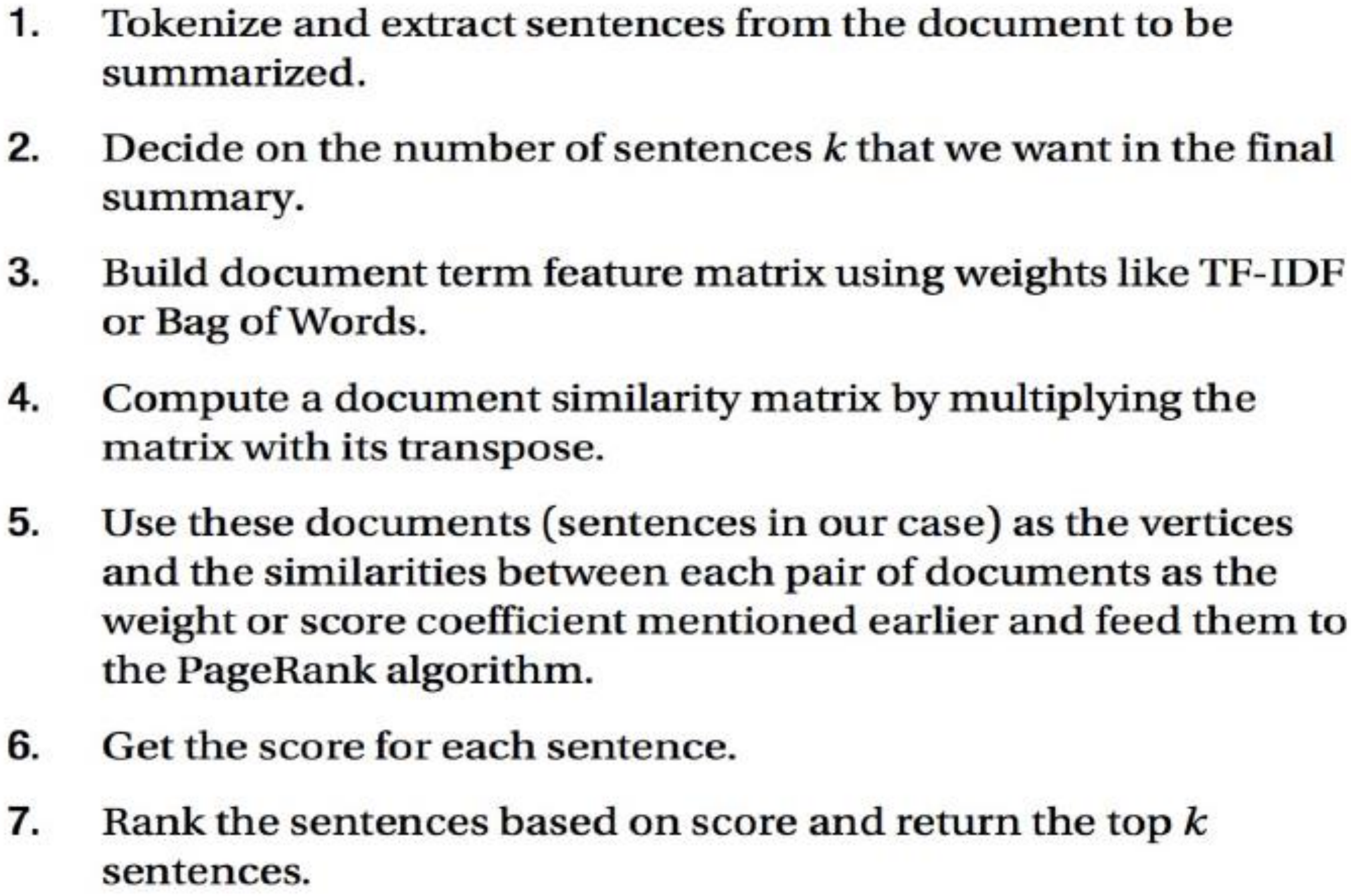
- Two Main Techniques
  - Text Rank Algorithm
    - unsupervised graph-based ranking algorithm
  - Latent Semantic Analysis (LSA)
    - Language Independent Semantic-Based Approach



TextRank algorithm

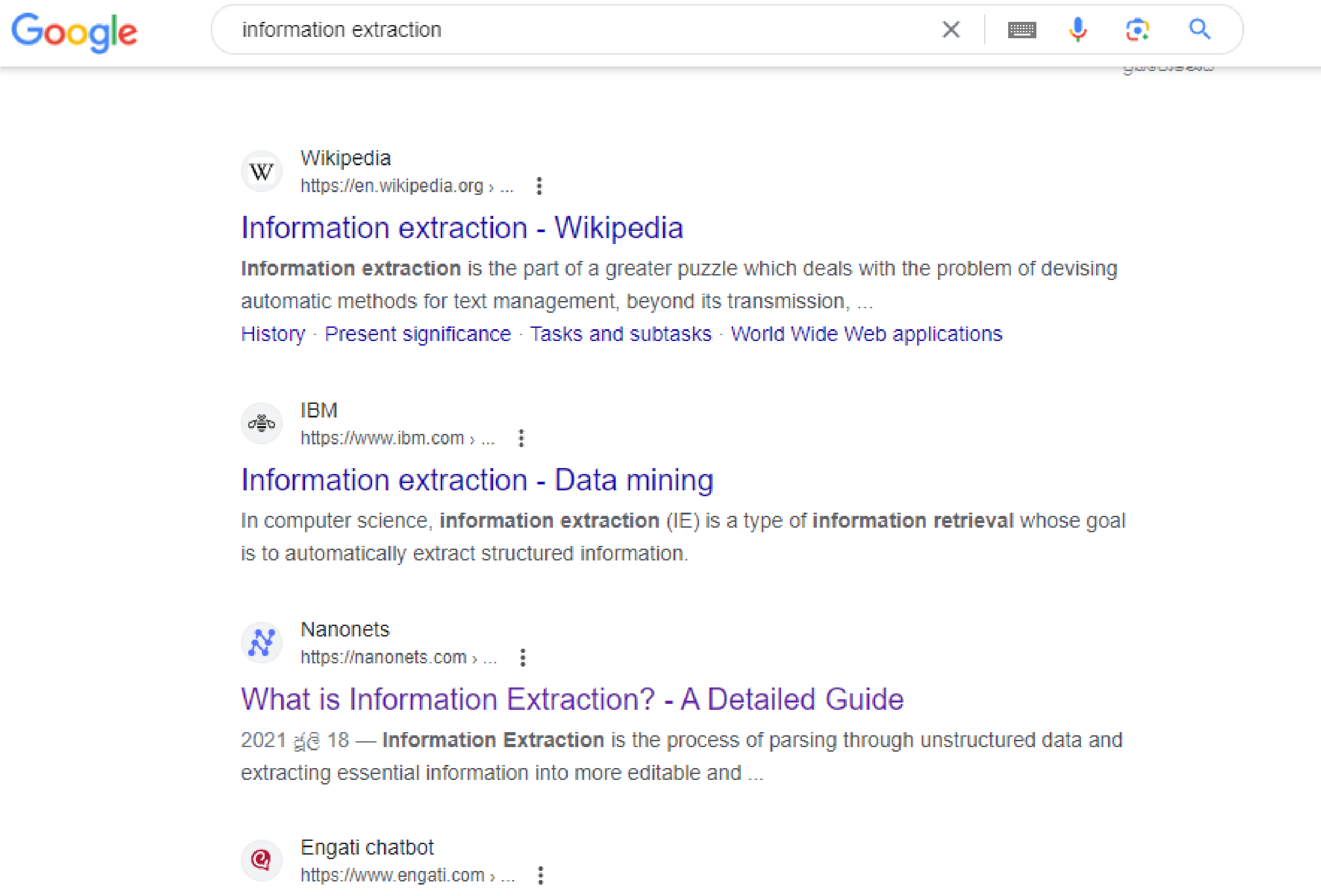
- Uses the popular PageRank algorithm
- PageRank is a graph-based algorithm to score web pages based on their importance
  - Importance is measured by how many pages link to a given page
  - As well as the importance of such pages
  - Hence it is an intrinsically iterative process
- In TextRank, sentences are the nodes/vertices and similarities between them are the links/edges
  - We then apply the PageRank algorithm to compute weights

TextRank algorithm



1. Tokenize and extract sentences from the document to be summarized.
2. Decide on the number of sentences *k* that we want in the final summary.
3. Build document term feature matrix using weights like TF-IDF or Bag of Words.
4. Compute a document similarity matrix by multiplying the matrix with its transpose.
5. Use these documents (sentences in our case) as the vertices and the similarities between each pair of documents as the weight or score coefficient mentioned earlier and feed them to the PageRank algorithm.
6. Get the score for each sentence.
7. Rank the sentences based on score and return the top *k* sentences.

Page Rank Algorithm



information extraction

Wikipedia
https://en.wikipedia.org › ...

Information extraction - Wikipedia

Information extraction is the part of a greater puzzle which deals with the problem of devising automatic methods for text management, beyond its transmission, ...

History · Present significance · Tasks and subtasks · World Wide Web applications

IBM
https://www.ibm.com › ...

Information extraction - Data mining

In computer science, **information extraction** (IE) is a type of **information retrieval** whose goal is to automatically extract structured information.

Nanonets
https://nanonets.com › ...

What is Information Extraction? - A Detailed Guide

2021 ජූලි 18 — **Information Extraction** is the process of parsing through unstructured data and extracting essential information into more editable and ...

Engati chatbot
https://www.engati.com › ...

- PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results.
- It is named after both the term "web page" and co-founder Larry Page.
- PageRank is a way of measuring the importance of website pages.

Page Rank Algorithm

$$PR(p_i;t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j;t)}{L(p_j)}$$

PR(p_i, t) → Page rank at t^th iterations for i^th webpage.

d → Damping Factor (Way to do teleportation)

L→ length of outgoing links

N→ length of webPages

- The probability, at any step, that the person will continue following links is a damping factor d.
- The probability that they instead jump to any random page is 1 - d

Latent Semantic Analysis/Indexing

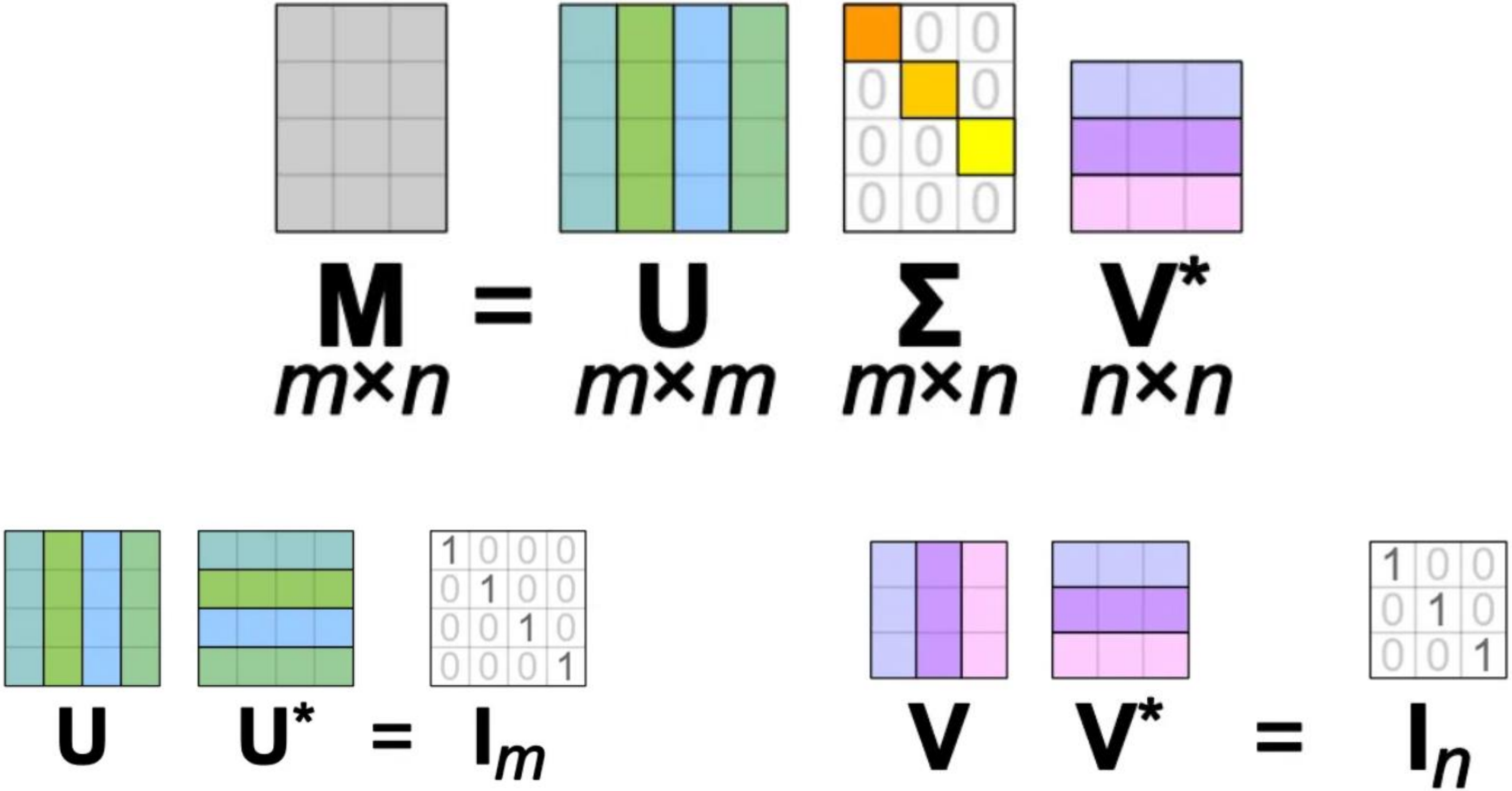
- Latent sematic indexing is used to analyze the relationship between a set of documents and terms contained in the document.
- Singular Value Decomposition (SVD); mathematical concept is used to compute set of matrices which give the similarity between the documents.

Singular Value Decomposition (SVD)

- Singular Value Decomposition is one of dimensionality reduction techniques.
- SVD is factorization of matrix into 3 matrices.

$$A = U\Sigma V^T$$

U is referred to a left singular vector,
∑is a singular values or eigen values,
V is a right singular vector.

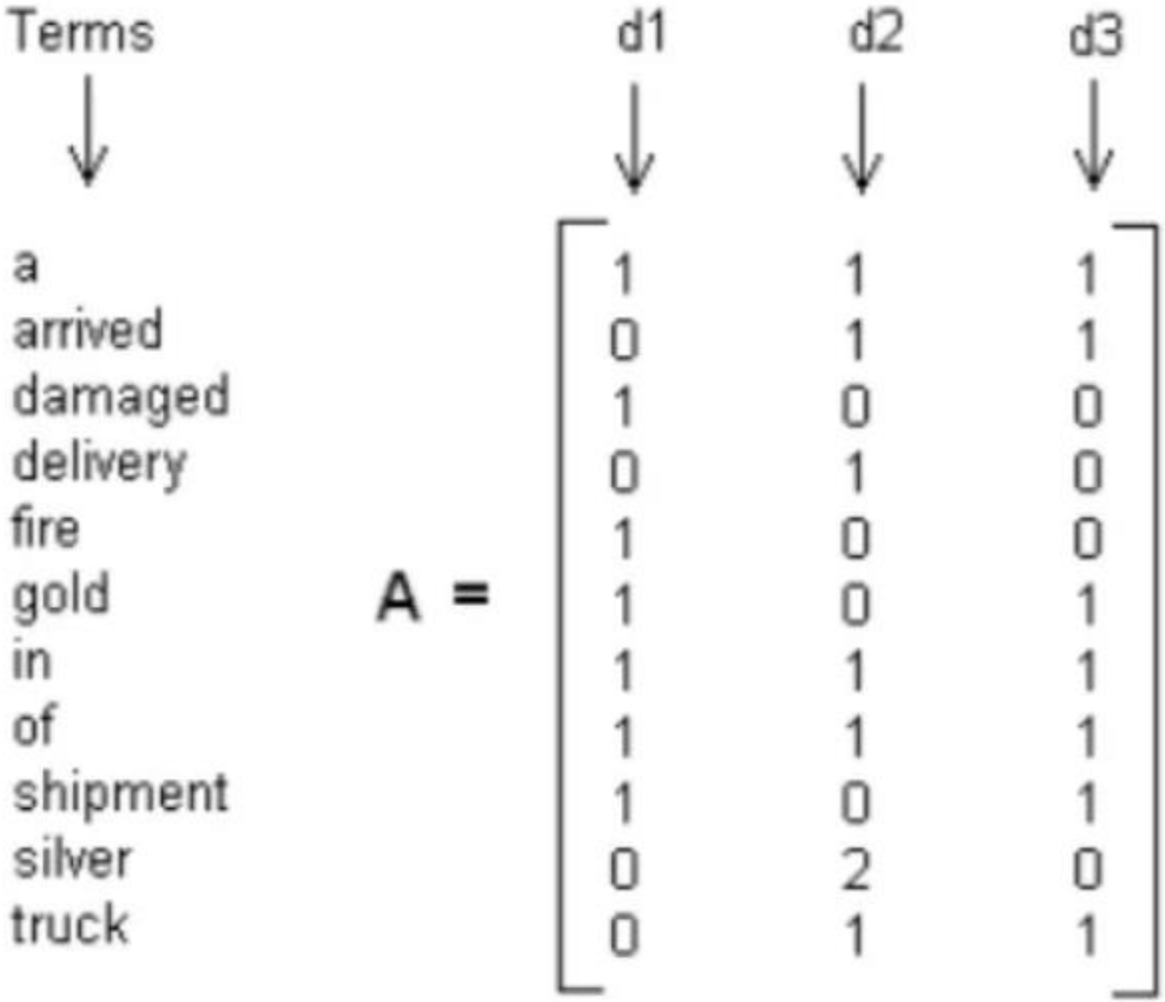


Calculation

d1: Shipment of gold damaged in a fire.

d2: Delivery of silver arrived in a silver truck.

d3: Shipment of gold arrived in a truck.



$$
A = \begin{array}{l|ccc}
\text{Terms} & d1 & d2 & d3 \\
\hline
\text{a} & 1 & 1 & 1 \\
\text{arrived} & 0 & 1 & 1 \\
\text{damaged} & 1 & 0 & 0 \\
\text{delivery} & 0 & 1 & 0 \\
\text{fire} & 1 & 0 & 0 \\
\text{gold} & 1 & 0 & 1 \\
\text{in} & 1 & 1 & 1 \\
\text{of} & 1 & 1 & 1 \\
\text{shipment} & 1 & 0 & 1 \\
\text{silver} & 0 & 2 & 0 \\
\text{truck} & 0 & 1 & 1
\end{array}
$$

$$A = USV^T$$

$$\mathbf{U} = \begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1576 & -0.3046 & -0.2006 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & 0.4078 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix} \qquad \mathbf{V}^{\mathbf{T}} = \begin{bmatrix} -0.4945 & -0.6458 & -0.5817 \\ 0.6492 & -0.7194 & 0.2469 \\ -0.5780 & -0.2556 & 0.7750 \end{bmatrix}$$

3.2 Vector-Space Models

The LSA method makes use of a term-sentence matrix. Different vector space models can be used to represent sentences in a term space. Some of these are briefly dicussed below:

Binary features This is the basic bag-of-words model where, each term in the document is a dimension in the space. A sentence has a feature value of 1 for a word if it contains a word and 0 otherwise.

Term Frequency Sometimes it is desirable to capture the number of occurences of a word in a particlar sentence. This can be used as a feature value in the term-sentence matrix.

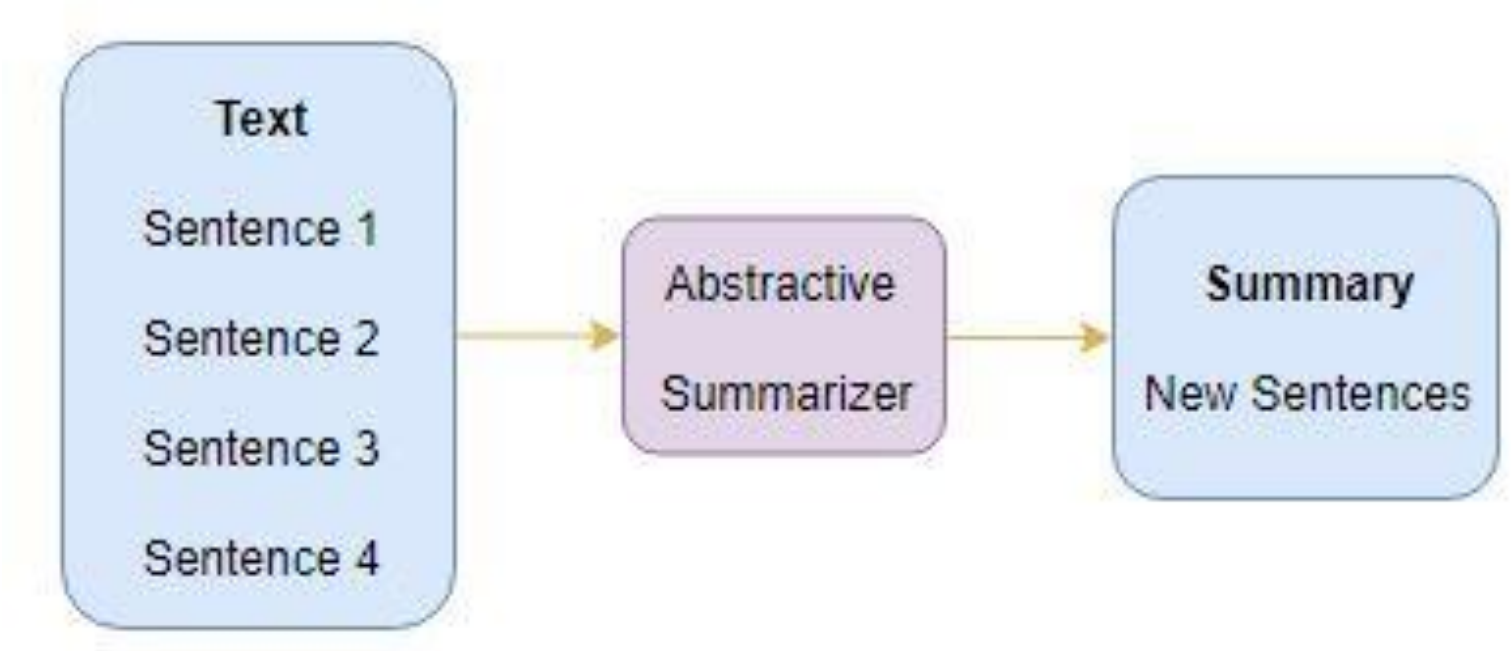
TF.IDF The previous models assume that all words are of equal importance when characterizing sentences. However, a word that occurs in several sentences may not be particularly useful in distinguishing one sentence from the rest. To capture this, a variant of TF.IDF, commonly used in Information Retrieval, can be used. In place of IDF, we use the inverse sentence frequency, calculated as follows:

$$IDF(w) = \log_2 \frac{n_w}{N} \tag{3}$$

where, n_w is the number of sentences in which w occurs and N is the total number of sentences in the document.

Abstractive Summarization

- This technique involves the generation of entirely new phrases that capture the meaning of the input sentence.
- Sequence 2 Sequence (encoder-decoder) model.



Applications of Sequence Models (RNNs)

Speech recognition		→	"The quick brown fox jumped over the lazy dog."
Music generation	Ø	→	
Sentiment classification	"There is nothing to like in this movie."	→	★☆☆☆☆
DNA sequence analysis	AGCCCCTGTGAGGAACTAG	→	AGCCCCTGTGAGGAACTAG
Machine translation	Voulez-vous chanter avec moi?	→	Do you want to sing with me?

Image by Andrew Ng on Coursera