# **Text Summarization**

# **NLP**

# **ELC002**

Text Summarization and Question Answering System for COVID-19 Data

ELC ACTIVITY – 4th Semester (2022)

* Bisman Kaur(102017051)

1. Developed a language-independent extractive text summarization system that generates a summary for [given document](https://drive.google.com/drive/folders/1u6EgCVC7ZOJjxkIcT-UZGz_dreCiDpqW?usp=sharing).

Library used: Spacy,NLP

Developed an extractive text summarization using various features and reduced the length of text to a short and crisp summary using NLP in python on unlabelled data. By calculating word frequencies, sentence tokens and heapq.

2. Developed a question answering system that can answer When/Where/Who type questions from a given set of documents.

Library used: nltk.data,Wikipedia,wordnet

Text summarization in NLP is the process of summarizing the information in large texts for quicker consumption.

As information is available in abundance for every topic on the internet, condensing the important information in the form of a summary would benefit a number of users. An automatic text summarization system generates a summary, i.e., short length text that includes all the important information of the document.

Formally, Summarization can be defined as a task of producing a concise and fluent summary while preserving key information and overall meaning.

We have developed the system using Python’s library nltk (The Natural Language ToolKit).

There are two NLTK libraries that are necessary:

* from nltk.corpus import stopwords
* from nltk.tokenize import word\_tokenize, sent\_tokenize

Sentence Features:

1. Cue - Phrases:

* Sentences starting with ‘in conclusion’, ’example’, ’therefore’, ‘our investigation’, ‘the article describes’, ‘the best’, ‘the most important’, ‘in particular’, ‘according to’, ‘significantly’ are of great importance in a text document.
* The sentence consisting of cue phrase or words are assigned high score.
* The score can be calculated by the formula:

Cuescore=number of cue words in sentence/maximum cue score

1. Numerical data

* Sentence containing numerical data like event dates, transaction having numerical data, year, age, etc. are important.
* These sentences must be included in the summary.

1. Sentence Length

* Sentences which are too long or too short (as compared to a threshold) are not valuable for the summary.
* So, some penalty is added to these sentences.

1. Sentence Position

* Sentences closer either from the start or end of a document are considered as important and strong candidate for the summary.

**PART-II**

Develop a question answering system that can answer When/Where/Who type questions from a given set of documents.

I am trying to build intelligent systems that don’t just return documents related to the question, but extract relevant information within the documents and puts forth the detailed answer, like one that a human would have come up with.

This task comes under the fields of:

* [Information Retrieval](https://nlp.stanford.edu/IR-book/pdf/01bool.pdf)(IR)
* [Information Extraction](https://www.ontotext.com/knowledgehub/fundamentals/information-extraction/) (IE)
* Natural Language Processing (NLP).

The design of a question answering system has specific vital components. There are three distinct modules used in a question-answering system:

* **Query Processing Module:** Classifies questions according to the context. This module identifies the context and focus, classifies the type of question, and sets the answer type’s expectations.
* **Document Processing Module:** Information retrieval module that focuses on gathering relevant documents.
* **Answer Processing Module:** Once the relevant documents are retrieved, they need to be parsed through to obtain an accurate and appropriate answer.

**Evaluating the answers obtained**

We want the question answering system to produce relevant, correct, and complete answers to the point. Hence many evaluation metrics were developed to measure such ambiguous terminologies. A few of those metrics are F1-scores, precision, recall, etc.