**Report IBM Project**

**On**

**Text summarization & Visualization**

*Submitted by,*

Zeus Saikia

Abhinab Saikia

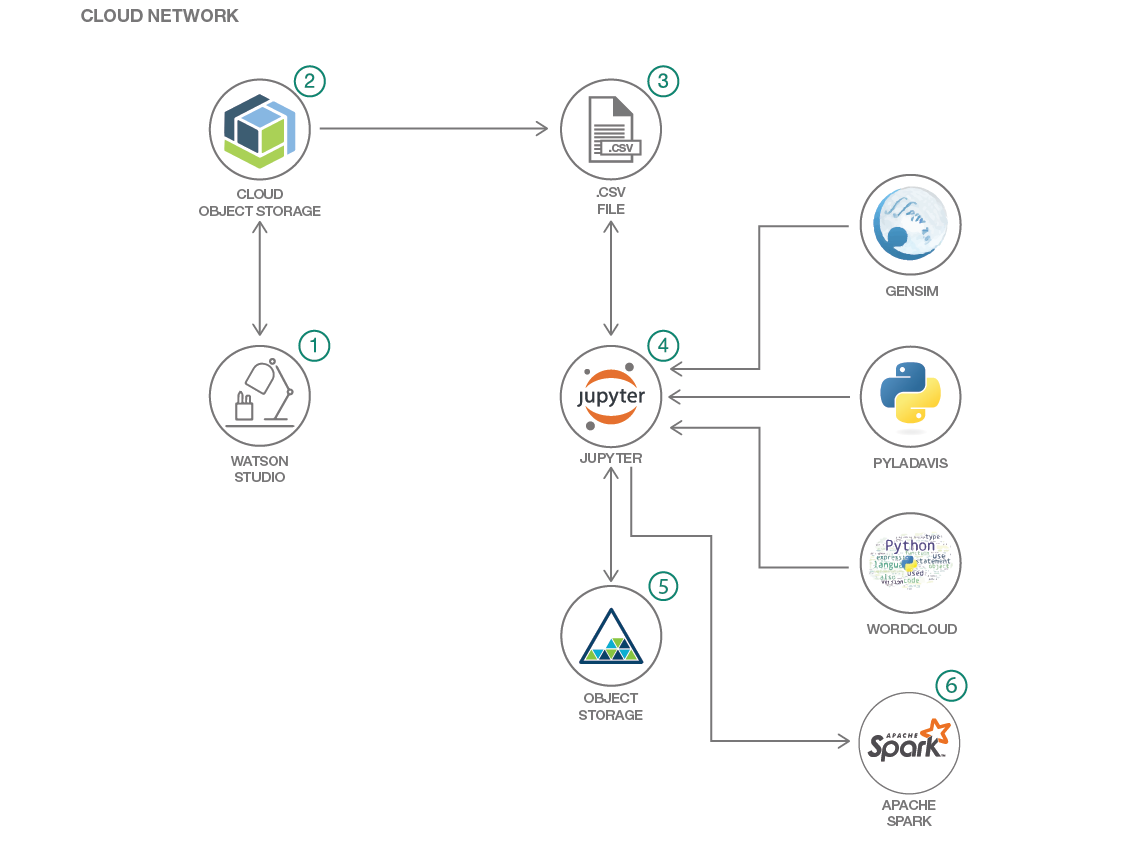
Bhargab Kataki

Indrajeet Kaur Sehra

Asna Brahma

**UNDERSTANDING THE PROJECT:**

Text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document. We will demonstrate a methodology to summarize & visualize text using Watson Studio. There are two methods to summarize the text, extractive & abstractive summarization. We will focus on extractive summarization which involves the selection of phrases and sentences from the source document to make up the new summary. Techniques involve ranking the relevance of phrases in order to choose only those most relevant to the meaning of the source. Some of the advantages of text summarization are below. We will also demonstrate different methods to visualize the data which can aid in providing quick peek of the data. Visualizations can stand in many relations to texts – and, as research into learning with pictures has shown, they can become particularly valuable when they transform the contents of the text.

Summaries reduce reading time. When researching documents, summaries make the selection process easier.Text summarization improves the effectiveness of indexing.Text summarization algorithms are less biased than human summarizers. Personalized summaries are useful in question-answering systems as they provide personalized information.Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of texts they are able to process.

**APPLICATION DEVELOPMENT ARCHITECHTURE:**

1. IBM Watson Studio: Watson Studio democratizes machine learning and deep learning to accelerate infusion of AI in your business to drive innovation. Watson Studio provides a suite of tools and a collaborative environment for data scientists, developers and domain experts. It

analyzes data using RStudio, Jupyter, and Python in a configured, collaborative environment that includes IBM value-adds, such as managed Spark.

With the help of IBM Watson, we can do the following :

* Allowing us to access pre-trained models from the Watson Services, such as Watson Visual Recognition.
* Enabling us to bring in non-structured data.
* Further automating and providing insight into model management.
* Continuing to provide us with a choice of best-in-breed data science/ML tools
* Strengthening our drag-and-drop interface to build analytics models using SPSS Modeler.
* Enabling us to visualize the insights with dynamic dashboards.

1. IBM Cloud Object Storage: Cloud object storage is a format for storing unstructured data in the cloud. Object storage is considered a good fit for the cloud because it is elastic, flexible and it can more easily scale into multiple petabytes to support unlimited data growth.

The architecture stores and manages data as objects compared to block storage, which handles data as blocks, and logical volumes and file storage which store data in hierarchical files. An IBM Cloud service that provides an unstructured cloud data store to build and deliver cost effective apps and services with high reliability and fast speed to market. This code pattern uses Cloud Object Storage. The object storage software design includes a globally unique identifier for each object along with rich, customizable metadata. The metadata is separated to enable other capabilities such as application- and user-specific data for indexing, interfaces that can be directly programmed by the application, a global namespace and more flexible data management policies.

1. Jupyter Notebooks: An open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line interface in a shell.

1. Python: Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.

It is used for web development (server-side), software development, mathematics, system scripting. Python can be used on a server to create web applications. It can be used alongside software to create workflows.

It can connect to database systems. It can also read and modify files.

1. Data Science: Data science is the process of using algorithms, methods, and systems to extract knowledge and insights from structured and unstructured data. It can be used to make predictions and decisions using analytics and machine learning.
2. Analytics: It uncover insights with data collection, organization, and analysis. Analytics delivers the value of data for the enterprise.
3. Text Ranking: Gensim is a Python library for topic modelling, document indexing and similarity retrieval with large corpora. Target audience is the natural language processing (NLP) and information retrieval (IR) community. The gensim implementation is based on the popular TextRank algorithm.
4. WordCloud: It is used for identifying and visualizing the key words in the document. The wordcloud library is MIT licenced, but contains DroidSansMono.ttf, a true type font by Google, that is apache licensed. The font is by no means integral, and any other font can be used by setting the font\_path variable when creating a WordCloud object.
5. pyLDAvis: It is a Python library for interactive topic model visualization. pyLDAvis is designed to help users interpret the topics in a topic model that has been fit to a corpus of text data. The package extracts information from a fitted LDA topic model to inform an interactive web-based visualization.

Scalability Scope of Project:

We can proceed with this project in this following way:

Image collection summarization is another application example of automatic summarization. It consists in selecting a representative set of images from a larger set of images. Video summarization is a related domain, where the system automatically creates a trailer of a long video. This also has applications in consumer or personal videos, where one might want to skip the boring or repetitive actions. Similarly, in surveillance videos, one would want to extract important and suspicious activity, while ignoring all the boring and redundant frames captured.

At a very high level, summarization algorithms try to find subsets of objects (like set of sentences, or a set of images), which cover information of the entire set. These algorithms model notions like diversity, coverage, information and representativeness of the summary.