Capstone Project

Presentation on

Automated Q&A System

Presented by : 1. Priyabrata Mohanty
2. Rahul Sharma

Pipeline

- Problem statement
- Objective
- Data summary
- Data cleaning
- Machine learning model building
- Model deployment
- Conclusion
- Appendix
    - Data Sources
    - Data Assumptions

Why Automated Q&A System?

- Customer handling.
- FAQ and chatbot have become very vital for a growing business and the process that can be automatically done. That should be done automatically and use cases like these is the reason we need automated querying systems.

Problem statement

To build a Deep Learning model for the Automated Question & Answering system, that would the question from the user as an input & display the recommended answer as the output.

Data summary

- We used wide spectrum of data to test our automated Q&A system.
- The data comprises, a book about machine learning
- Blogs which were scraped using python scripts.

Data cleaning & Processing

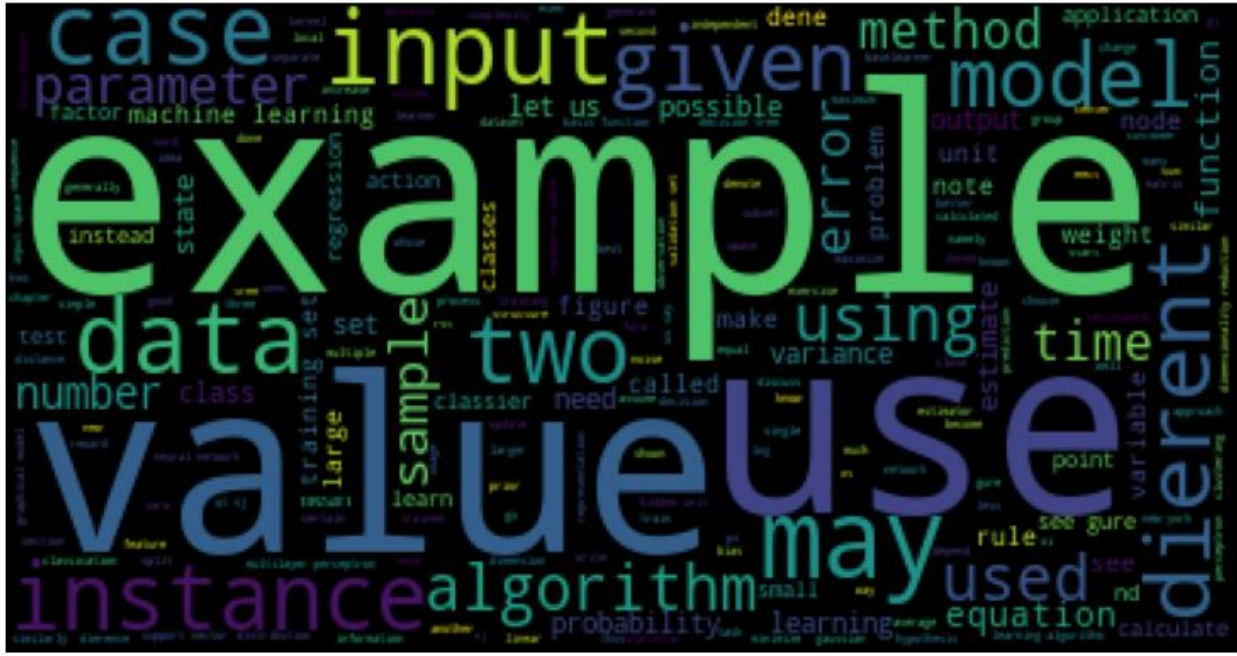
- **Merged** different articles together to form whole dataset.
- Removal of **stopwords**, **punctuations** & **unwanted** characters (e.g. “\n”, “\xa0”) from data.
- Converted the texts into **lower case** & **tokenization**.
- Removal of unwanted empty spaces in the sentences & added “.” at the end of every sentence .


Exploratory Data analysis

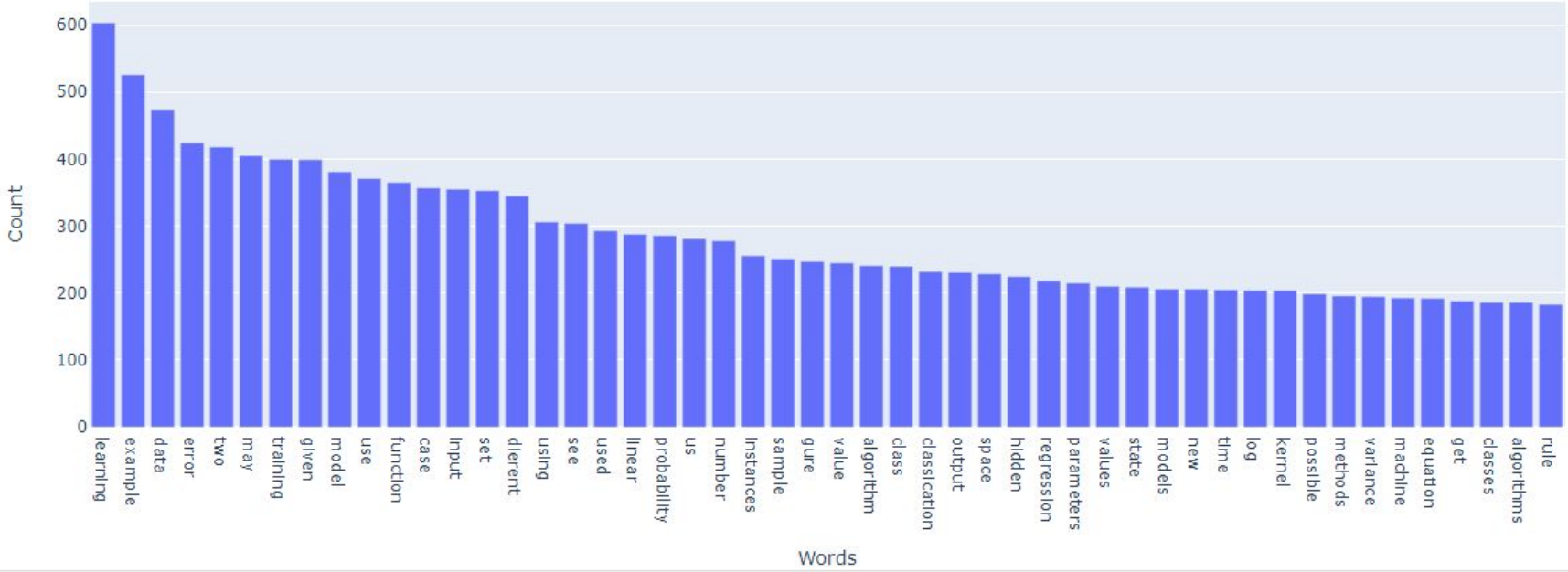
Important Words In The Article

According to this word cloud:

- **Example, Data, Value, Input, Algorithm, Model, Instance, Case, Equation** are the most important words in this article.



Word VS Frequency Graph



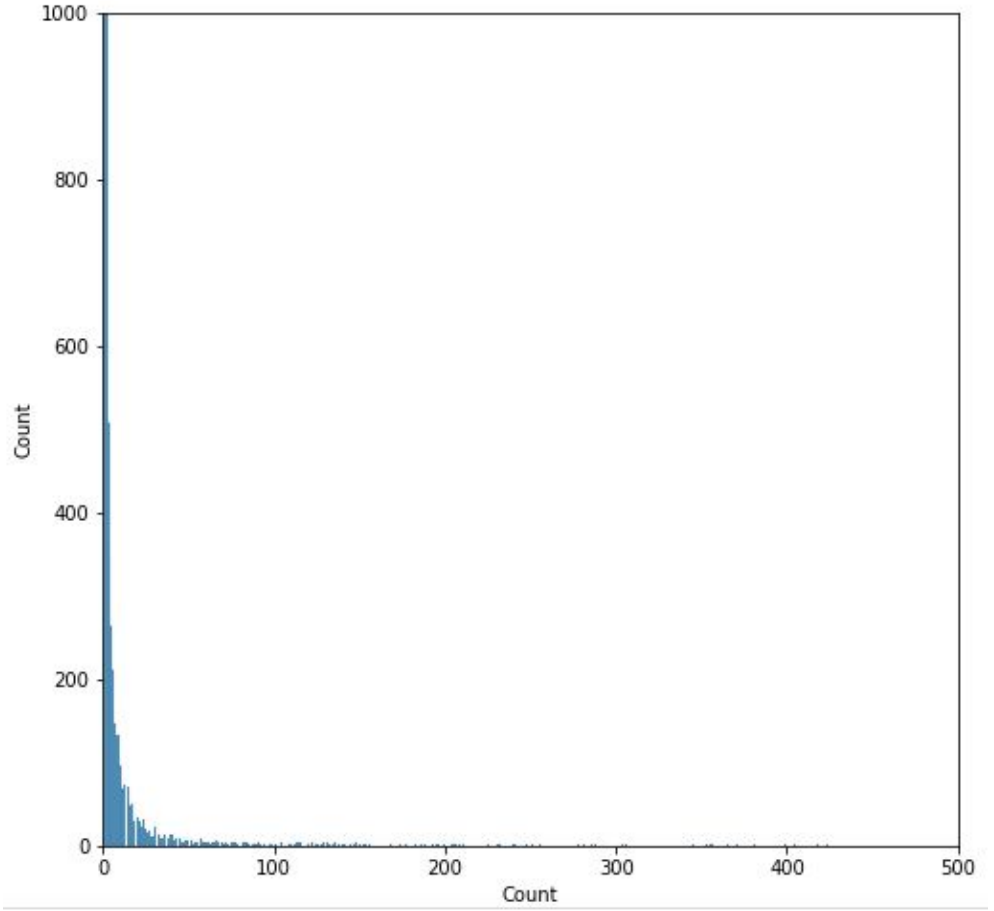
According to above graphs:

- **Learning** is the most frequent word with frequency of **603**
- Other most frequent words are **Example, Data**, **Error**, **Training**, **Model**.

Word Counts of articles Distribution

According to above graph:

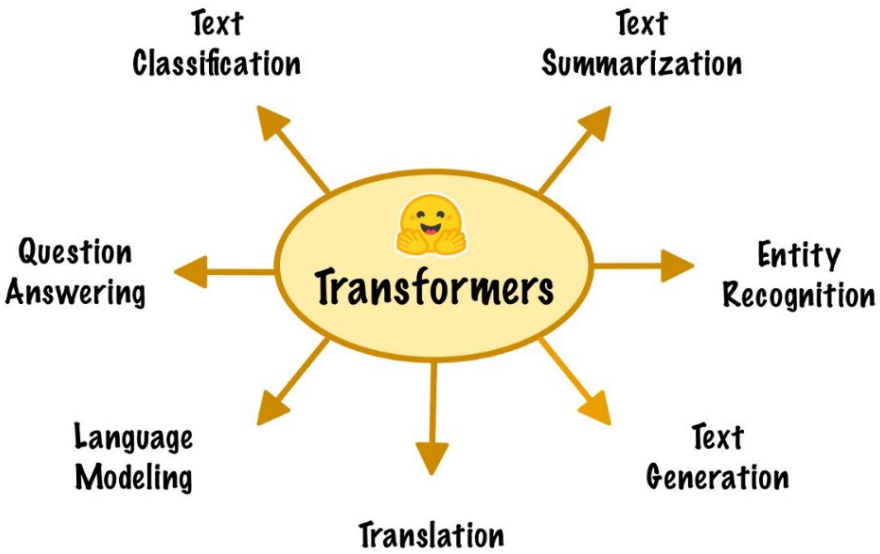
Most of the articles word counts is in the range of 0-100.



Deep Learning Model Building

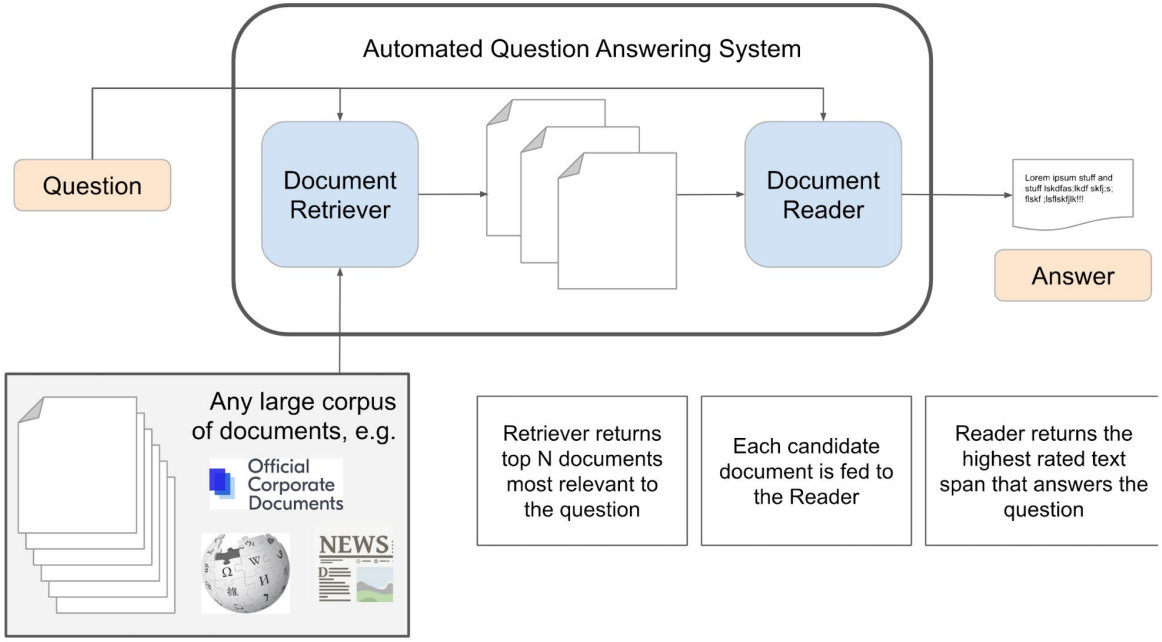
HuggingFace Transformers

- The **HuggingFace Transformer** library is one the most popular libraries used for Natural Language Processing (NLP).
- It not only can do the Question answering but also the Text summarization, Entity recognition, Text generation, Translation and many more.



Extractive Question Answering

- It is the task of extracting an answer from a text given a question. An example of a question answering dataset is the **SQuAD** dataset, which is entirely based on that task. If you would like to fine-tune a model on a SQuAD task.



Pipeline For Transformer Model

Following is a general pipeline for any transformer model:

Tokenizer definition →Tokenization of Documents →Model Definition →Model Training →Inference

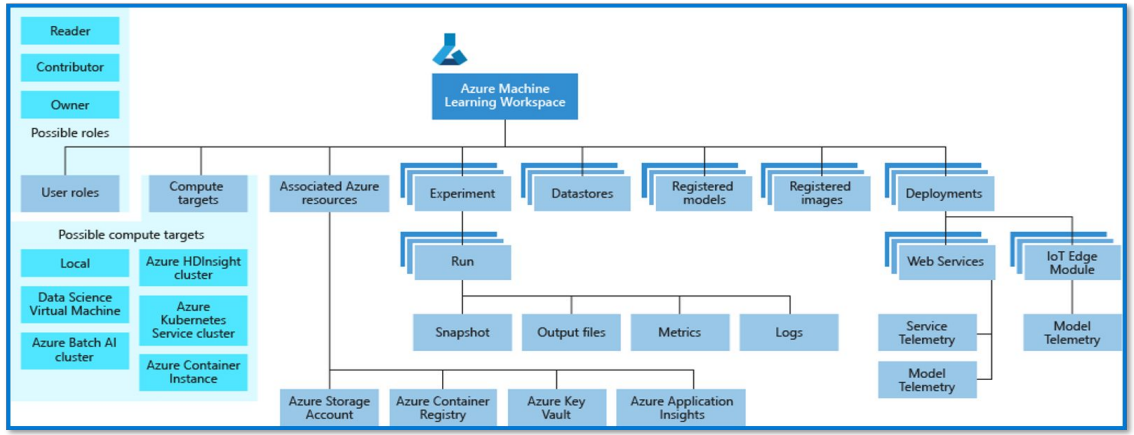
- It leverages techniques of deep learning and uses a concept called **masked learning**.
- This model can look at long-range dependencies in sentences and helps in understanding context much deeper. Depending on the context in the sentence, it creates different vectors for the same words, that are semantically different.

Deployment Strategies

- Docker - For starter we tried to deploy directly through the Azure. Which went in vein but after that we built a docker image to see if it works and it worked.
- Heroku - As the transformers needed one of pytorch or tensorflow in the system. These libraries were heavy on size and complexities. After trails we got to know that Heroku only let's deploy upto 500 mbs
- Azure - Azure deployment went smoothly after the docker image got built successfully.

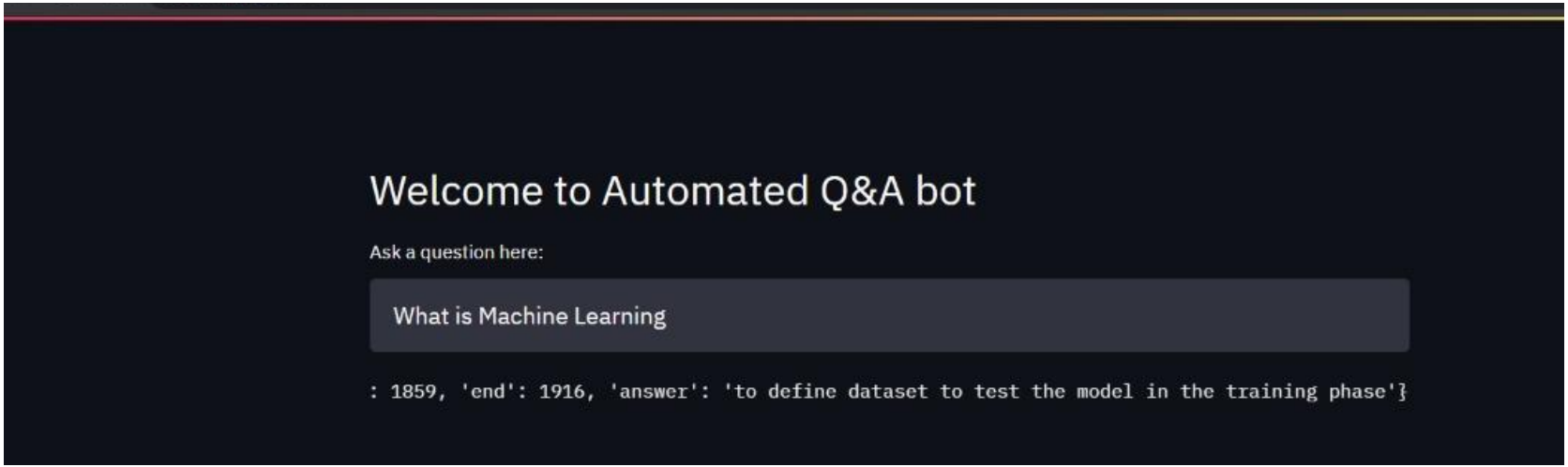
Model Deployment

- After building the model, deployed it in **microsoft Azure** so that I could be accessible to everyone.
- Also we could monitor its performance and the logs to resolve any issue in future if caused.



```
from azureml.core import Workspace
ws = Workspace.create(name='Demo',
                      subscription_id='12345678-1234-1234-a0e3-b1a1a3b06324',
                      resource_group='Contoso',
                      location='eastus2')
```

Automated Q&A Bot Testing



- A simple question was asked to the Automated Q&A bot based on Machine Learning after building & training the Deep Learning model.
- It displayed the relevant answer with the starting & ending positions of the sentence in our document.

Conclusion

- We need to keep our dataset **updated** always so that whenever someone asks any questions related to new topics the Q&A bot would be able to provide any relevant answers.
- The **stopwords** must not be removed from the dataset otherwise it would provide the incomplete answers to the users.
- If the document size or area of topics asked by the users is increased then we might need huge amount of data to provide any relevant answers which might be more computationally expensive.

Appendix - Data sources

- Here is a snapshot of data dictionary:

The data comprises, a book about machine learning.

Blogs which were scraped using python scripts.

Appendix - Data Methodology

We conducted a thorough analysis, Deep Learning model building for the Automated Q&A system. The process includes:

- **Data cleaning** – Removal of unwanted characters, spaces from data.
- **EDA** - Understanding data using different visualization methods.
- **Model Building** – Built & tested the deep learning models to do the Automated Q&A bot.
- **Model Deployment** – Deployed the model in Microsoft Azure.

Thank You !