CSCE 5290: Natural Language Processing

Project Proposal

Title: ***AnsicleXsummerizeR*** **(AXR)**

GitHub Link: *https://github.com/BlueTiddern/Article_summarizer_NLP_IR*

Group 4:

Pavan Yarlagadda	: 11658983
Giridhara Srikar Chittem	: 11703343
Uday Kiran Chimpiri	: 11727109
Manushree Buyya	: 11660966

Motivation:

We humans are social beings, as the world continues to run, a lot of changes occur. One of the best ways to keep up to date with current affairs is to read through the news articles. Here is the main catch, English is a complex language, and the meaning can vary with just a character or a small sequence of characters before/after the words, also when we need specific information and being able to view related text or paragraphs in the articles through a question-answering system can make our lives easy while saving a lot of time.

Now the reader has the relevant answer to their question, will this satisfy the users' need? Yes, but the answers retrieved in this way *may* lose their context. Providing a complete summary for each of the query key terms organized by topic considering the related terms is a great way to provide linked information.

The main motivation of this project is to target this exact problem statement by building a pipeline that effectively processes the question that will target the related sections as the output for the *QuestionAnswering* system (model 1). These lists of targeted and related articles will be the input topic modelling (model 2) and organization of articles by topic, this will be the input for contextual summarizer (model 3) module to generate a concise summary for each of the query keyword topics upholding the contextual information.

By combining the capabilities of *QuestionAnswering* and *ContextualSummerization* we can produce not only the valuable answers but also a quick summary of the relevant paragraphs. I believe this is very important when it comes to inferring information from articles where not only the 'point' but also understanding the information around it is important.

Significance:

News articles and related forms of information multiply exponentially each day, this rise in volume and complexity will pose a roadblock for the readers, specifically - journalist, researchers, students and analytic teams. Traditional search engines and retrieval algorithms provide us with general results that are flat or isolated.

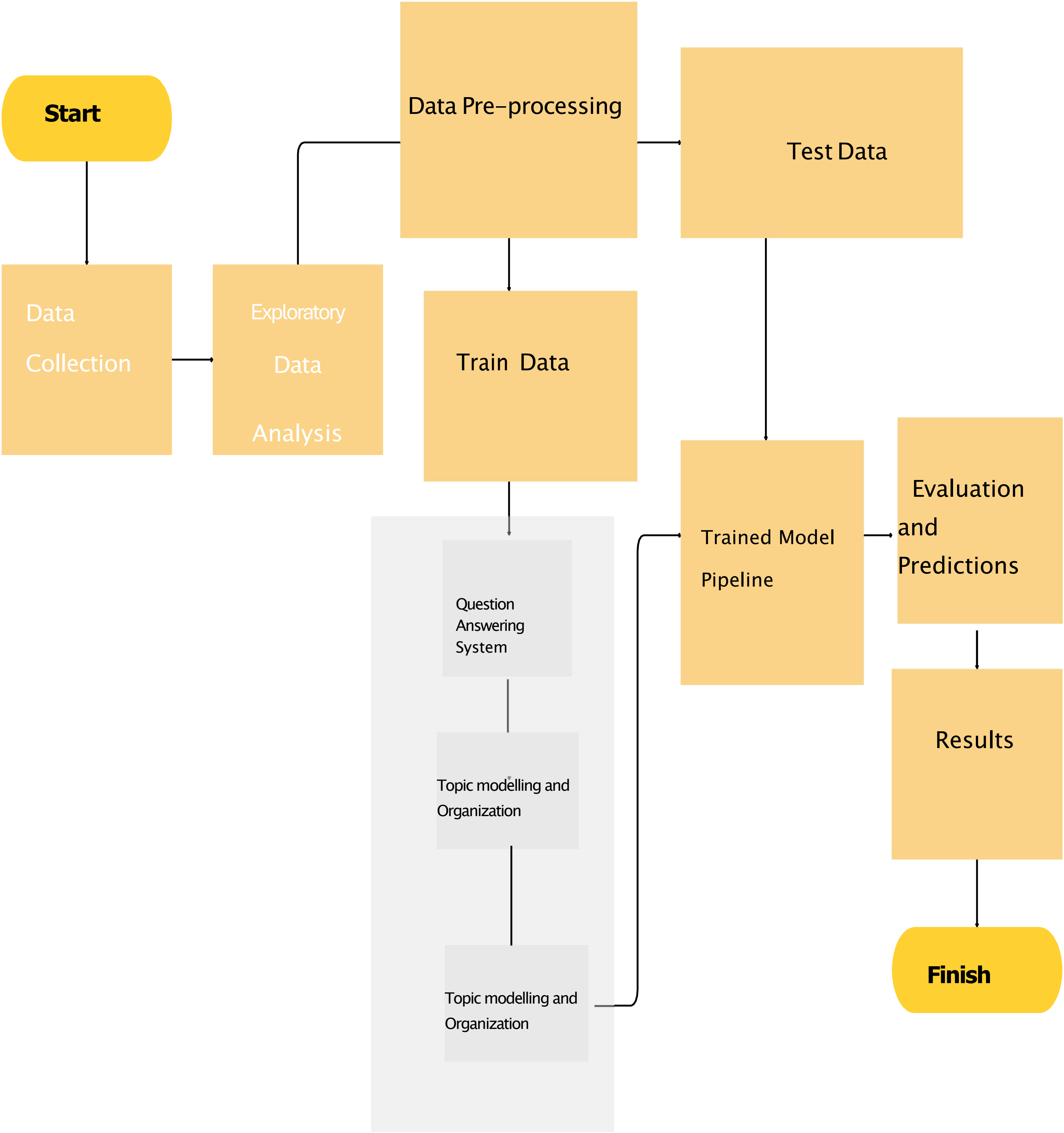
As we see more developments in NLP and information retrieval, we can meet the vast information demands by leveraging these technologies offered to produce more fine-grained answers along with their contextual understanding. This can be a crucial development in areas like Journalism, research/education as they can grab both the specific data with broader related details, this makes sure we get accurate understanding of the information while saving time.

Objectives:

The main objective is to build a pipeline that does question-answering and contextual summarization sequentially. It is aimed at articles or news readers who want to quickly retrieve information precisely along with contextual summary for a particular piece of information. Below are Objectives for the project

Data Collection	Retrieve the data for model training and testing.
Exploratory Data Analysis	To understand the data distribution, to look for possible challenges like noise and imbalances. Graphs this analysis will be included.
Data preprocessing	Handling special characters, tokenization, handling padding and truncating. Making Data ready.
Test – Train split	Splitting the data into training and test sets for model training and evaluation.
Question Answering System	Train a model effectively to retrieve a list of articles which are related to the key terms in the Query. Each key term has a set of relevant retrieved documents. Evaluations on results
Topic Organization	The retrieved relevant documents will be organized by a topic. Each article will be segmented into a topic modeled, based on word frequencies and distributions
Contextual Summarization System	These sets of retrieved relevant documents sets by topic will be passed to summarizer to generate a topic vice summary.
Validation and Display	The generated summaries evaluated and are displayed topic vice giving an additional context to the asked question.

Project Flow Diagram:


Start
Data Collection
Exploratory Data Analysis
Data Pre-processing
Test Data
Train Data
Question Answering System
Topic modelling and Organization
Topic modelling and Organization
Trained Model Pipeline
Evaluation and Predictions
Results
Finish

Features:

- **Data set**:

  - ➔ Data set considerations: Initially we will be basing the proposal on Google Deep mind dataset. CNN and Daily Mail news pieces are available with a compiled question set, stories set with complete articles and raw HTML data.
  - ➔ Data set source: DMQA (nyu.edu)
  - ➔ Data set Meta data: CNN – 90K stories (151 MB) and corresponding 380K questions (207 MB). Daily Mail - 197k stories (358 MB) and 879k Questions (505 MB). Above mentioned are Compressed sizes.
  - ➔ Data is in English; the question file type has @entity encodings.
  - ➔ During the EDA process the distribution of the data set will be viewed, a few supporting graphs will be plotted. The data set will be checked for potential challenges
  - ➔ Data processing, tokenization and data organization; will be followed to get the data set ready for retrieval and model training.

- **Model Components**:

  - ➔ Question-Answering System (document Retrieval model), here a QA model -1 will be implemented, to retrieve the articles (doc ID's) and make a list of articles for topic vice modeling. Possible models for Model 1, Evaluations of the results included.
  - ➔ Topic Organizer (topic modeling model), model 2, here we will be modeling the topics for all the relevant retrieved documents and place each retrieved document under a topic modelled. Possible models – LDA, LSA, BERT topic.
  - ➔ Contextual Summarizer (summarizer model), the organized topic vice segregation of the article will be used here to generate summary that is related to each topic, Model 3 (!TBD) BART/RoBERTa/ small BERT. Evaluations of the results included.
  - ➔ A Final Validation/predictions and display component to display the final summaries for each topic.

- **Uniqueness:** Our projects unique selling point lies in how the models are integrated to bring out the summaries for each key term used in a question. The output is not just an answer but a list of summaries for each key term giving an overall context for the question asked.

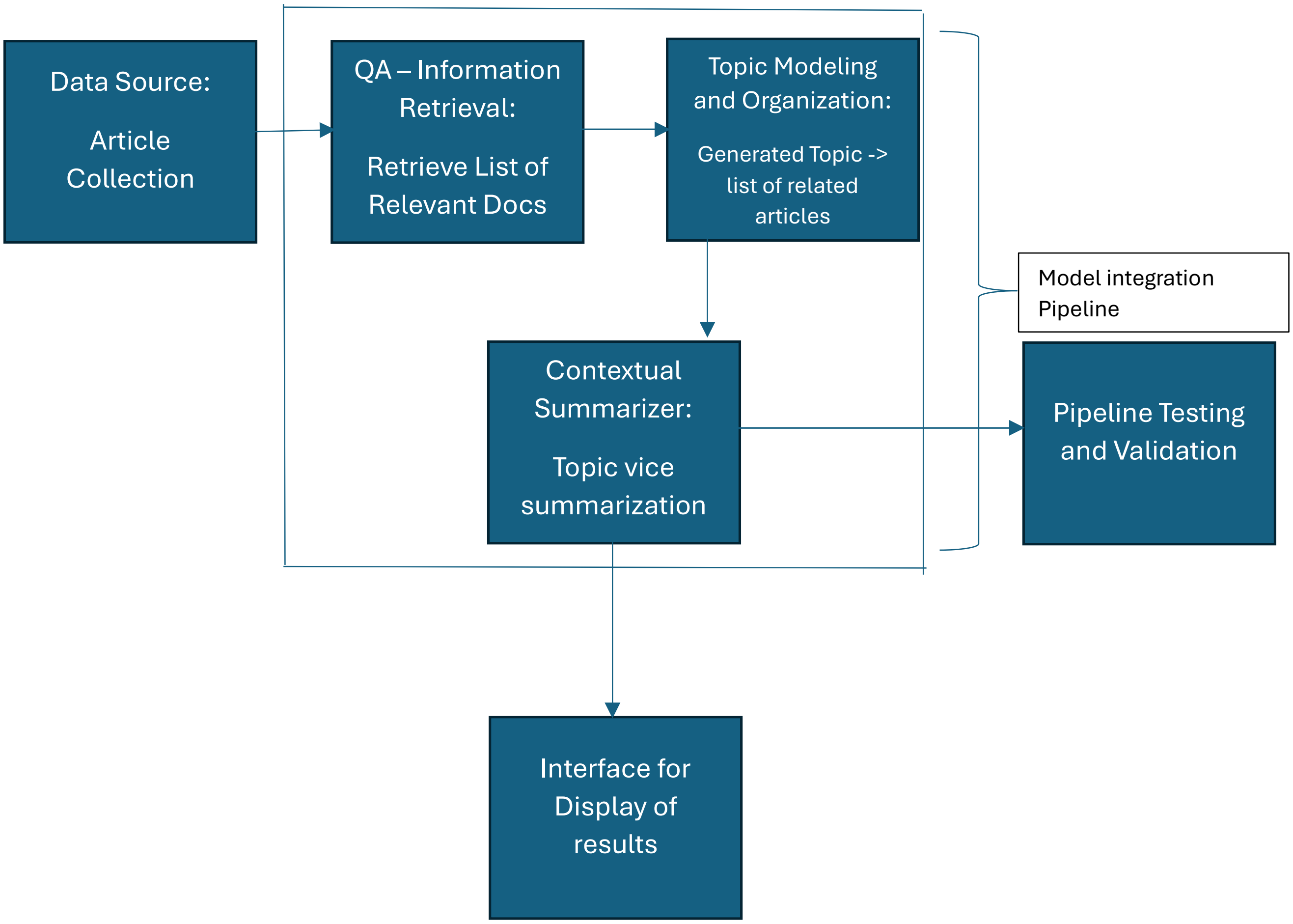
- **Evaluation Metrics:**

  ➔ The Question–Answering system (document Retrieval) will be checked for how relevant the retrieved article list is, possible metrics – Accuracy, F1, Recall, Precision.

  ➔ A combination of Coherence Score and human interpretation will be used to evaluate the topic organization.

  ➔ The contextual summarizer will be evaluated on how well the generated summary, coherence and relevancy is. Metrics such as ROUGE/ BLEU/BERT scores will be considered.

Core Component Pipeline:



Data Source:
Article Collection
QA – Information Retrieval:
Retrieve List of Relevant Docs
Topic Modeling and Organization:
Generated Topic -> list of related articles
Contextual Summarizer:
Topic vice summarization
Model integration Pipeline
Pipeline Testing and Validation
Interface for Display of results

Deliverable	**Description**
Data library	Collect data from sources and place them in appropriate directory structure.
EDA phase	Distribution of data and potential inconsistency analysis.
Data Preprocessing Module	Clean, format, tokenize and make data set for testing and training.
QA System IR Retrieval system	Handling the data retrieval and generating a list of articles related key terms in the query. Evaluation of the model is also handled.
Topic modeling and organization of articles	The list of articles is processed to model topics and organize them under topics
Contextual Summarizer system	These topic article lists are process to give the topic vice summary for our final output.
Final Validations and Display of results	A complete validation will be performed and the display for the final output will be designed

Tentative Milestones:

- ➔ Week 1 & week 2: Data collection and storage, EDA and Data preprocessing/splitting. Update documentation for project phase 1.

- ➔ Week 3 - 4: QA model construction and experimentation on which model to be implemented in document retrieval. Evaluation and validations to be done on IR retrieval.

- ➔ Week 4: Working on the topic modeling and organization for summary generation.

- ➔ Week 5 -6: Construction of the contextual summarizer. Model experimentation and suitability testing. Evaluation and validation to be done on summaries generated. Update documentation for project phase 2.

- ➔ Week 7: Test the model-pipeline complete validations, building the interface for results.

- ➔ Week 8 – End of project: Make adjustments or optimizations.