



The City College
of New York

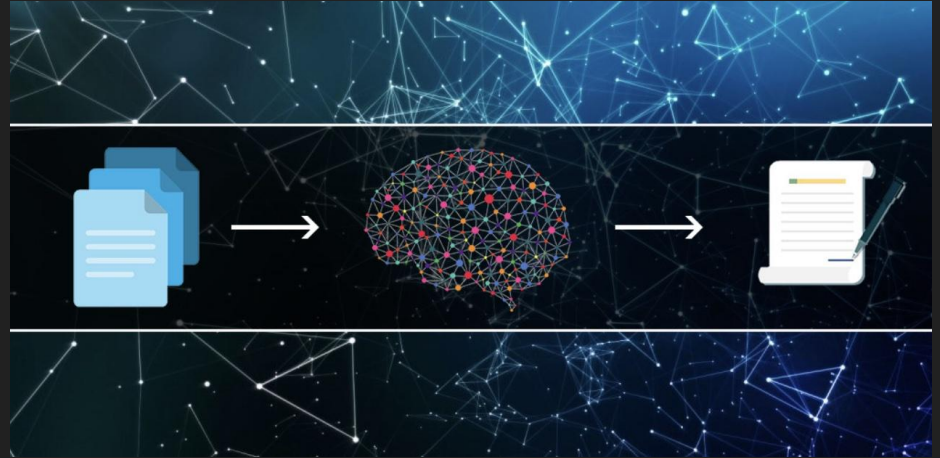


NetDesign: Text Summarizer and RAG Playground

Senior Design Project
The City College of New York
By: Usman Abbas, Evan Perez, Jair Ruiz

What is Text Summarization

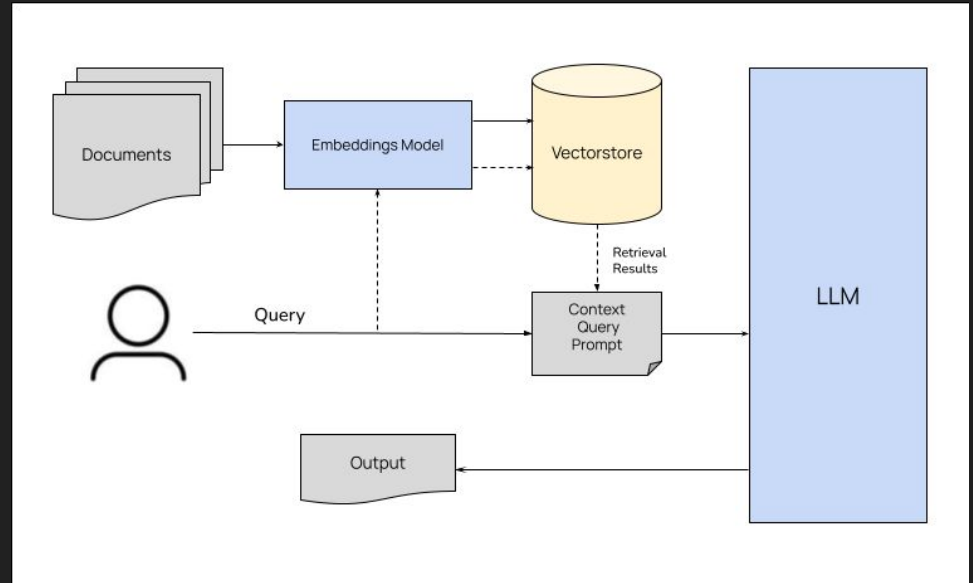
- Take large pieces of text and be able to provide high level details from the context



Source:
<https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/>

What is RAG

- Optimization of LLM response
- Uses external sources
- Tons of use-cases

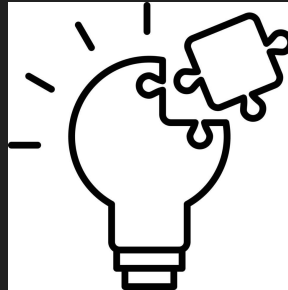


Problem

- Many companies need text summarization capabilities
- Not a simply way to compare text summary techniques such as RAG
- Manual summarizations are very time consuming

Solution

- Build a text summarizer using sequence to sequence model
- Compare model results with various RAG pipelines
- Create one resource for text summarization tools



Objectives

- Train Text Summarization model with Seq2Seq
- Test model on dataset and evaluate performance
- Compare results against LLM based RAG pipelines

Data

- Multiple Datasets
- Preprocessing required
- Each entry containing “Source Document” and “Ground Truth Summary”

Source Document

(@entity0) wanted : film director , must be eager to shoot footage of golden lassos and invisible jets . <eos> @entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie (the hollywood reporter first broke the story) . <eos> @entity5 was announced as director of the movie in november . <eos> @entity0 obtained a statement from @entity13 that says , " given creative differences , @entity13 and @entity5 have decided not to move forward with plans to develop and direct ' @entity9 ' together . <eos> " (@entity0 and @entity13 are both owned by @entity16 . <eos>) the movie , starring @entity18 in the title role of the @entity21 princess , is still set for release on june 00 , 0000 . <eos> it 's the first theatrical movie centering around the most popular female superhero . <eos> @entity18 will appear beforehand in " @entity25 v. @entity26 : @entity27 , " due out march 00 , 0000 . <eos> in the meantime , @entity13 will need to find someone new for the director 's chair . <eos>

Ground truth Summary

@entity5 is no longer set to direct the first " @entity9 " theatrical movie <eos> @entity5 left the project over " creative differences " <eos> movie is currently set for 0000

Source: <https://paperswithcode.com/dataset/cnn-daily-mail-1>

Methodology

- Preprocessing
 - Remove duplicate/null values
 - Remove punctuation
 - Split data for training/validation/testing
- Training
 - Training Seq2Seq Model on context and summary dataset
- Evaluate Results
 - Graph validation loss
 - Observe summary results



Challenges

- **Data Preprocessing**
 - Defining thresholds
 - Complexity: Managing diverse data formats and quality.
 - Scalability: Scaling preprocessing steps for large datasets.
- **Developing text evaluation scores**
 - Setting consistent evaluation metrics.
- **Organizing different RAG frameworks**
 - Developing higher order functions to take different RAG components
 - Compatibility: Ensuring different frameworks work together.
 - Functionality: Developing versatile functions for varied components.



What we know / Need to Learn

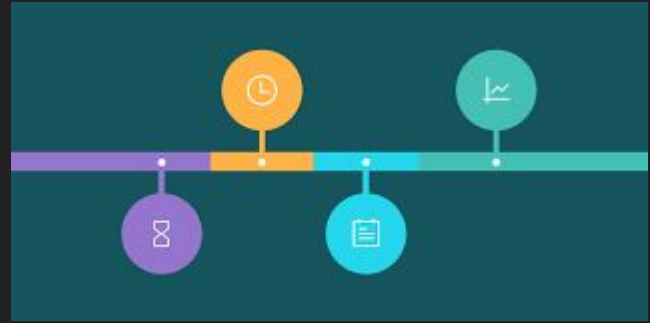
- What we know
 - RAG pipelines
 - Model training and evaluation
 - Natural Language Processing
- What we need to Learn
 - Seq2Seq modeling
 - LSTM
 - RAG playground application workflow
 - Data preprocessing

Simulation Environment

- Google colab
 - Simple to use
 - Access to cloud GPU
 - Large storage
 - Adaptability and organization for prototyping
- Migrate to Streamlit for interface
 - Fast Iterations: Quick refreshes for prototype updates.
 - Session State: Manages user settings across interactions.
 - Community Plugins: Extensive tools and add-ons.
 - Simple Deployment: Easy setup for app launching.
 - Data Visualization: Effective display tools for model insights.



Timeline



- Data preprocessing
- Define model architecture and overall system

- Application development
- RAG playground

