Smart AI News Reader

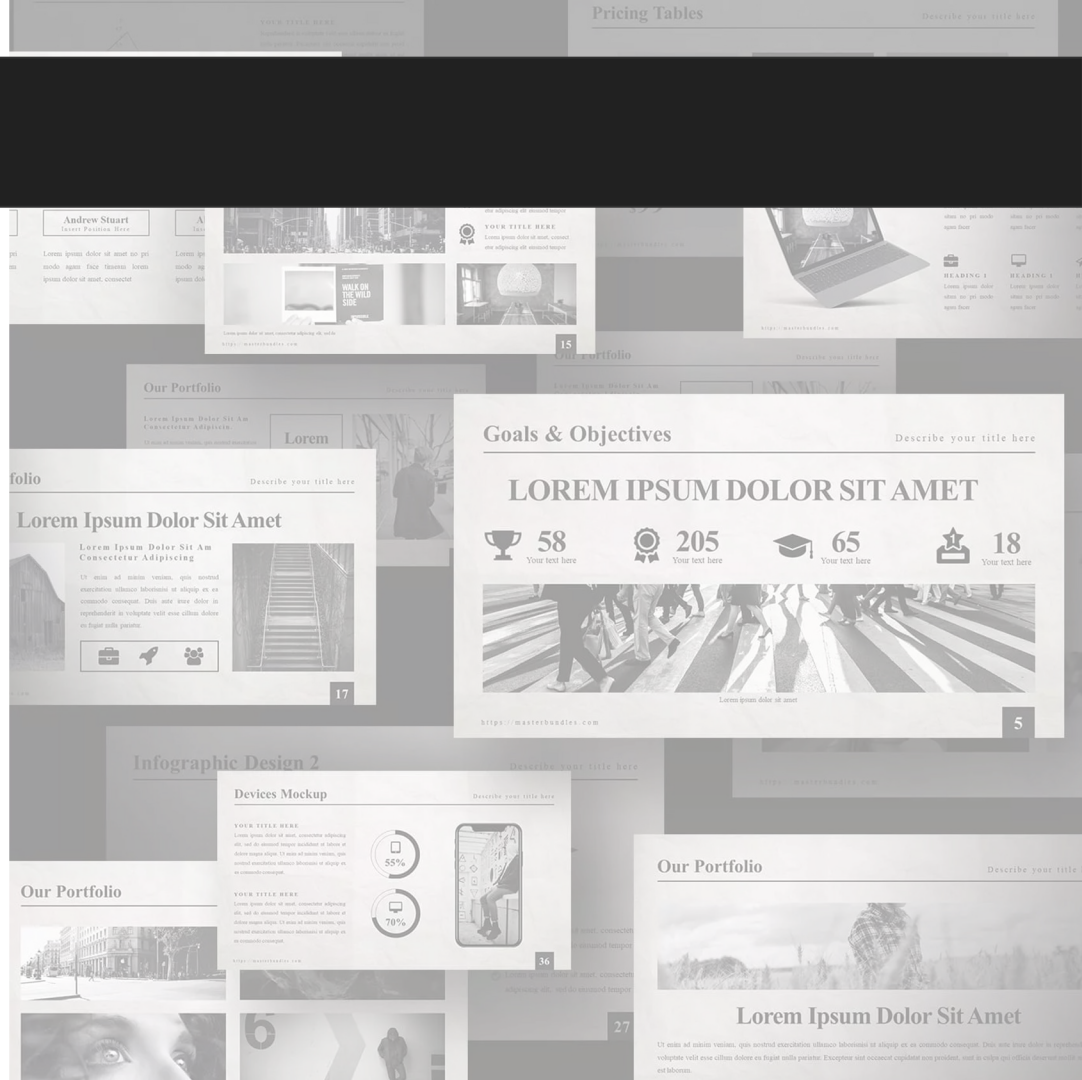
2023 Fall NLP Final Project

December 11 2023

Team 5. Arjun Bingly, HaeLee Kim, Nayaeun Kwon

Project Overview

1. Introduction
2. Methodology
    - News Fetch
    - Summarization
    - Zeroshot Classification
    - Keyword Extraction
    - Question & Answering
1. Demo (App)
2. Possible Improvement
3. Conclusion



Introduction

1. **Background**

- Unprecedented proliferation of digital content
- Overwhelming volume challenges traditional news consumption methods

1. **Objectives**

- Introduce the Smart News Reader application
- Emphasize the importance of NLP in news exploration
- Streamline information consumption in the contemporary era
- Address challenges posed by the information deluge

Methodology 1:

News Fetch

- **Intro**
  - A comprehensive class for extracting and parsing news articles from given URLs
- **Initial Attempts**
  - Tried *Beautifulsoup* for web scraping and HTML parsing
  - Limitation: Each news website uses different structuring
- **Implementations**
  - Incorporates the Newspaper3k library for efficient download and parsing
  - Parses rich features like images and videos

Methodology 2:

Article Summarization

- **Goal**
  - Condenses extensive news articles into concise and informative summaries.
- **Initial Attempts**
  - Tried *BERT-Large-CNN*

    : pre-trained on the extensive CNN news dataset

    : fine-tuning not necessary
- **Limitations**
  - Maximum token size

    : observed that most news article are larger than the maximum allowed text input size

Methodology 2:

Article Summarization

- **Implementation of LangChain**

  : Split text into maxtokensize with small overlap

  : Chunks processed separately to fit model constraints

  : Prevented loss of critical information in large articles

- **Limitation of Controlling Over Summary Size**

  : Lack of direct control over max summary size

  : Recursive summarization considered but with drawbacks of longer run-time and exaggerated errors

Methodology 3:

Zeroshot Classification

- **Goal**
  - Predicts the relevance of input labels to a given input without the need for explicit training on labeled data

- **Implementations**
  - BART-Large model fine-tuned on MLNI (“facebook/bart-large-mnli”) for sequence classification

  - MNLI (Multi-Genre Natural Language Interference) : Dataset for evaluating NLP models' ability to comprehend sentence relationships across diverse genres and contexts

Methodology 4:

Keyword Extraction

- **Goal**
  - The most relevant n-grams from a document
- **Initial Attempts**
  - Statistical methods like TF-IDF and YAKE
  - Limitation: Does not take context into consideration
  - Transformer based key phrase generation
  - Limitation: Does not guarantee that the key-phrase exists in the main body
- **Implementations**
  - Key BERT
  - Uses BART Sub-Word Tokenization and Cosine Similarity

Methodology 5:

Conversational Q&A

- Intro
  - Enable users to ask questions regarding the article in a conversational manner.
- Implementation
  - **Bert-Large Fine-Tuned on SQUAD**
    - Limitations

      1. Truncation due to max token limits

      2. Non conversational model

Methodology 5:

Conversational Q&A

- **Retrieval Augmented Modeling**: Solves the truncation issue
  - How it works
    - LangChain used to split the news article into smaller chunks with overlap.
    - The embeddings are stored in a vector database (Chroma).
    - Retrieval of most similar chunks based on user question.
  - Consideration
    - Optimal embedding model crucial for performance.
    - 'jinaai/jina-embeddings-v2-base-en' performed well but couldn't be integrated due to HuggingFace and LangChain constraints.
    - Hence used sentence-transformers/all-MiniLM-L12-v2

Methodology 5:

Conversational Q&A

- **Retrieval Augmented Modeling**
  - Limitation
    - Lacked conversational capabilities
    - Answers typically limited to phrases/sentences extracted from the text

Methodology 5:

Conversational Q&A

- **Larger LLM Models**: Solves the conversationality
  - Local approach, avoiding API reliance
  - Obtained Llama-2 model by META upon request
  - Tries Llama-2 7B chat and 13B chat, faced size limitations
  - Quantization Attempts and Model Selection
    - Attempted quantization to 4-bit integer precision using llama.cpp, faced dependency issues
    - Found already quantized model on HuggingFace: TheBloke/Llama-2-13B-chat-GGUF

Methodology 5:

Conversational Q&A

- **Larger LLM Models**
  - Implementation Differences with Llama-2 13B model
    - LangChain and llama-cpp-python used for conversational retrieval
    - Utilized CuBLAS for GPU inference
    - Utilized LangChains Conversational retrieval chain
  - How it works
    - Prompt for condensincing follow up questions
    - Prompt
    - Implemented LangChain Memory Buffer for tracking previous chat inputs and outputs

Methodology FAILED:

Translation many-to-one

- **Goal**
  - Ensures the accessibility of news articles across diverse linguistic audiences
- **Initial Attempts**
  - Many-to-One translation module, equipped with the MBART model
    : proficient in translating news articles from various source languages to English
  - NewsTranslator class handles language codes for translation mechanism
- **Implementation Failed**
  - Implementation caused some dependency related issues

App for Demo

Possible Improvement

1. The whole app takes a lot of computational resource and takes a fairly long time to run, optimizations of both computational resource and time could be done
2. The app uses multiple models, this could be condensed to a single LLM
3. Every input refreshes the whole app, even with caching more improvements needs to be made.
4. Better UI
5. Implementation of the translation feature
6. App currently only supports one user.

Conclusion

1. **Transformative News Experience**:
   Smart News Reader redefines news consumption through user-centric design and advanced NLP features.

1. **Versatile NLP Toolbox**:
   Summarization, QA, translation, zero-shot, and keyword extraction provide users with powerful tools for news interaction.

1. **Adapting to Evolving Needs:**
   Anticipates and meets user demands in the digital age, addressing immediate challenges and staying ahead.

1. **Innovative Benchmark:**
   Continuous integration of cutting-edge NLP techniques positions Smart News Reader as a pioneer in efficient and engaging news consumption.

Thank you