CS410 Project Report

Campuswire Search Module based on Question and answering
Team SM : Suejung Kang, Minsoo Kim

INTRODUCTION

Our project focused on improving the campuswire search module based on question and answering. Campuswire is where students spend most of their time during the semester. We ask questions about course materials, project topics, and other logistics.

Campuswire search module is mostly based on matching words, it finds materials that contain some of the words from the query. However, we search for answers using words different from the other students have asked, and want to ask detailed questions. This word matching-based search can not answer such questions in a sentence that contains similar but different words from posts.

We used sentence embedding to represent queries and posts. And used similarity scoring to find the similar questions posted by students. So this new module finds a similar question to what we ask and these posts will show the answer for it.

PROJECT SCOPE

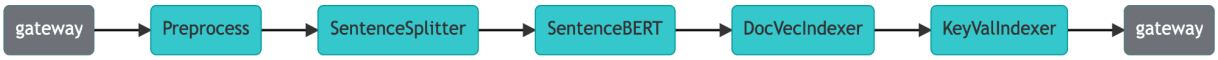
Our project scope was implementing a new search module based on question answering and making a restful API that our colleagues can try to use. This project report will focus on the implementation of the search module.

What we made

- Crawl data from campuswire 2021 fall CS410 feed.
- Implement search module based on sentence embedding
- Deploy restful API and web demo.

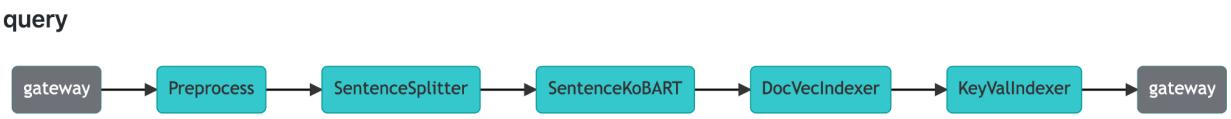
IMPLEMENTATION

Indexing



- 0. Saving Crawling data from campus wire(approximately 2000 Q-A set)

- 1. Preprocess text data
- 2. Sentence split in order to get embedding vectors from questions which has several sentences
- 3. load pre-trained sentence bert model(all-mpnet-base-v2) from hugging-face and put onto forward function and getting embedding vector from pooling layer
- 4. Indexing the crawled data by sentence chunking with segmenter(!?.) and aggregate to get 768 dimension embedding vector and keep important columns for fast searching



- 5. Define distance with cosine-similarity between two sentences
- 6. When query comes, preprocessing and getting embedding by sentence splitt again as done in indexing part

USAGE OF SEARCH MODULE

DEMO

- Demo : https://uiuc-cs410-demo.ngrok.io/

INSTALL FROM GITHUB

- Github : https://github.com/MinsooKim15/CS410_QuestionAnswering

Setup steps

1. Clone repository
2. Install requirements
    a. Python (at least 3.8)
    b. pip3 install -r requirements.txt
3. Index, this could take up to 30 minutes.
    a. python app.py -t index
4. Start Search Module
    a. python app.py -t query
5. Type in your question

CONTRIBUTIONS

1. Suejung Kang : Implement search module
2. Minsoo Kim : Crawled data from campuswire.

REFERENCES

- **JINA github: https://github.com/jina-ai/jina**
- **Question Answering example: https://github.com/yuanbit/jina-financial-qa-search-template**
- **https://towardsdatascience.com/how-to-build-a-production-ready-financial-question-answering-system-with-jina-and-bert-48335103043f**
- **FinBERT : https://arxiv.org/pdf/1908.10063.pdf**
- **https://huggingface.co/sentence-transformers/all-mpnet-base-v2**
- **https://arxiv.org/pdf/1908.10084.pdf**