TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

HỌC THỐNG KÊ

ĐỒ ÁN CUỐI KÌ

Nhóm sinh viên thực hiện:

20120040 – Nguyễn Quang Gia Bảo

20120136 – Huỳnh Tuấn Nam

20120158 – Trần Hoàng Anh Phi

Giảng viên hướng dẫn: Ngô Minh Nhựt

Lê Long Quốc

Thành phố Hồ Chí Minh, tháng 6 năm 2023

THÔNG TIN NHÓM

MSSV	Họ và tên
20120040	Nguyễn Quang Gia Bảo
20120136	Huỳnh Tuấn Nam
20120158	Trần Hoàng Anh Phi

BẢNG PHÂN CÔNG CÔNG VIỆC VÀ ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH

MSSV	Công việc phân công	Mức độ hoàn thành
20120040	Tìm hiểu nội dung đồ án	60%
20120136	Hỗ trợ môi trường để train và thực hiện deploy lên môi trường web, viết báo cáo.	95%
20120158	Thực hiện quá trình tiền xử lý dữ liệu, train các mô hình và deploy lên môi trường web, viết báo cáo.	100%

Đánh giá toàn bộ Project

Đạt % về mặt output, xử lí dữ liệu và thực hiện các yêu cầu.

Mục lục

I/ Thông tin chung:

1. Dựa theo cuộc thi Zalo AI Challenge: Task E2E QUESTION ANSWERING

https://challenge.zalo.ai/portal/e2e-question-answering

Theo trang chủ của cuộc thi:

- Khi tìm kiếm thông tin trên Internet, chúng ta sẽ nhận được rất nhiều đáp án, tuy nhiên phần lớn trong số chúng không phải thứ ta cần. Do đó, 1 hệ thống End-to-end Question Answering sẽ làm hài lòng những người đang tìm câu trả lời cụ thể, trực tiếp cho câu hỏi.
- Trong cuộc thi này, người tham gia sẽ xây dựng một hệ thống có thể tìm câu trả lời cho mỗi câu hỏi trong Vietnamese Wiki Corpus. Câu trả lời có thể là một thực thể wiki (Wikipedia Page) hoặc một con số/ngày tháng.

INPUT & OUTPUT:

- Question: Một câu hỏi tiếng Việt.
- Answer: Một thực thể Wiki (Wikipedia Page), một con số hoặc một ngày tháng trong Wikipedia Corpus
  + Nếu câu trả lời là 1 Wikipedia entity thì nó phải bắt đầu bởi "wiki/" và title của page.
  + Nếu câu trả lời là ngày tháng thì câu trả lời phải theo format: “ngày … tháng … năm ...", “tháng … năm …", “năm ….”. Ví dụ: năm 1942, ngày 2 tháng 9 năm 1945.
  + Nếu câu trả lời là 1 con số, thì đó là 1 số nguyên.

Examples:

Question	Expected Answer
Tổng thống đầu tiên của Mỹ là ai ?	wiki/George_Washington
Ai là người thiết kế ra dinh độc lập ?	wiki/Ngô_Viết_Thụ
Nhà thơ Xuân Quỳnh sinh năm bao nhiêu ?	năm 1942
Đại Việt sử ký toàn thư của Ngô Sĩ Liên năm 1479 gồm bao nhiêu quyển	15
Nước nào vô địch World Cup 2026	null

Cách giải được tham khảo từ top 1 solution của cuộc thi: https://github.com/Telegram-Zalo/zac2022-e2e-qa

Ngôn ngữ lập trình: Python.

Môi trường hỗ trợ: Visual Studio Code, Jupyter-Notebook, Kaggle.

Loại file: .py, .ipynb, .csv, …

Các thư viện, framework hỗ trợ chính: tensorflow, transformer, sklearn, …

2. Dataset được sử dụng từ trang chủ của cuộc thi:

https://challenge.zalo.ai/portal/e2e-question-answering

Gồm có:

- Wikipedia Corpus: Kho ngữ liệu của Wikipedia Vietnam, gồm khoảng 1 triệu Wikipedia Page chứa title và toàn bộ nội dung của mỗi trang
- Training data: end2end dataset có mô tả như sau:

Training data description:

- id : id of the item
- question : a natural question
- text: the text, maybe the answer to the question, may be not
- is_long_answer: true or false indicates whether the text could be a long candidate answers to the question
- short_candidate: the span answer extracted from the text (in the case the text is an answer of the question)
- short_candidate_start: the start position of the short candidate in the text.
- answer: the expected answer for the question. Please see the detail of section input & output.

id	FULL_ANNOTATION	PARTIAL_ANNOTATION	FALSE_LONG_ANSWER
id	X	X	X
question	X	X	X
title	X	X	X
text	X	X	X
is_long_answer	X	X	X
short_candidate_start	X		
short_candidate	X		
answer	X		
	~4.800 items	~3.700 items	~12.300 items

Testing data:

Public Test: ~500 questions

Private Test: ~500 questions

- Format của test data:

```
{
    "data": [
        {
            "id": "test1",
            "question": "Ai là người thiết kế ra dinh độc lập ? "
        },
        {
            "id": "test2",
            "question": "Nước nào vô địch World Cup 2026 "
        },
        {
            "id": "test3",
            "question": "Đại Việt sử ký toàn thư của Ngô Sĩ Liên năm 1479 gồm ba
o nhiêu quyển "
        }
    ]
}
```

- Format của submission data:

```
{
    "data": [
        {
            "id": "test1",
            "question": "Ai là người thiết kế ra dinh độc lập ? ",
            "answer": "wiki/Ngô_Viết_Thụ"
        },
        {
            "id": "test2",
            "question": "Nước nào vô địch World Cup 2026 ",
            "answer": null
        },
        {
            "id": "test3",
            "question": "Đại Việt sử ký toàn thư của Ngô Sĩ Liên năm 1479 gồm ba
o nhiêu quyển ",
            "answer": "15"
        }
    ]
}
```

Link to Wikipedia dumps online : https://dumps.wikimedia.org/viwiki/20220620/viwiki-20220620-pages-articles.xml.bz2

Link to cleaned Wikipedia: https://dl-challenge.zalo.ai/e2e-question-answering/wikipedia_20220620_cleaned.zip

Link to training data & public test & sample submission file of public test: https://dl-challenge.zalo.ai/e2e-question-answering/e2eqa-train+public_test-v1.zip

Link to private test (TBA): https://dl-challenge.zalo.ai/e2e-question-answering/e2eqa-private_test_v3.zip

Zalo QA 2019: https://www.kaggle.com/datasets/duykhanh99/zalo2022-e2e-qa

II./ Solution:

Gồm có 5 bước chính:

1. Cắt data wikidump thành các sliding windows kích thước 256.
2. Tìm candidate contexts bằng thuật toán BM25.
3. Rank lại top 200 condidate contexts bằng model BERT sentence pair dựa trên kiến trúc Transformers: phân tích các cặp câu và xác định mối quan hệ giữa chúng.
4. Tìm candidate answers từ contexts, chọn kết quả cuối cùng bằng majority vote + community detection w/Louvain.
5. Tìm top 100 úng cử viên cho answer bằng BM25, rank lại bằng một model BERT sentence pair khác để tìm kết quả cuối cùng.

1. Quá trình thực hiện đồ án:

Preprocess (Thu thập và tiền xử lí):

Lần lượt chạy các notebook trong thư mục “/notebook”:

- **0.0-create-sliding-window.ipynb**: tạo các sliding window cho dữ liệu (file *wikipedia_20220620_cleaned_v2.csv*), kết hợp một số bước tiền xử lý: xoá appending title, tokenize, xoá dấu câu,..
- **0.1-find-dirty-data.ipynb:** Tìm các dữ liệu "dirty" bằng cách so sánh các dữ liệu trong file dữ liệu thô *wikipedia_20220620_cleaned_v2.csv **(1)*** và file dữ liệu đã clean *zac2022_train_merged_final.json **(2).*** Nếu title của file 2 không nằm trong file 1 (nghĩa là chỉ có 1 kết quả) thì title đó là dữ liệu sạch, ngược lại thì đó là

"dirty" (có nhiều kết quả cho 1 title). Kết hợp các bước tiền xử lý và chia **chunk_size** để tăng tốc độ.

- **0.2-create-stage1-ranking.ipynb:** label cho tập dữ liệu *zac2022_train_merged_final.json*, nhãn **1** nếu đó là FULL_ANNOTATION hoặc là long answer, ngược lại là nhãn 0. Chia tập Validation = 15% của các FULL_ANNOTATION.
- **0.3-create-stage2-ranking.ipynb:** Sử dụng Pyspark để xử lý phân tán cho tập dữ liệu lớn, nhằm tăng tốc độ tính toán (tiền xử lý dữ liệu). Sử dụng thuật toán BM25 để tìm các ứng cử viên cho mỗi question, kết hợp label cho dữ liệu, mỗi group chính là tập các ứng cử viên, chỉ có 1 giá trị chính xác trong mỗi group này.
- **0.4-find-redirects.ipynb:** parse file xml để tìm các redirect link và map với dữ liệu đã tiền xử lý.

Sau khi thực hiện chạy lần lượt các notebook trên, các files kết quả được tạo ra:

- wikipedia_20220620_cleaned_v2.csv
- entities.json
- train_stage1_ranking.csv
- train_stage2_ranking.csv
- zac2022_train_merged_final.json

Training models (Huấn luyện các model):

Chạy các notebook:

- **1.0-train-bm25_stage1.ipynb**: Tạo từ điển, sử dụng tf-idf kết hợp huấn luyện mô hình bm25 và lưu lại các file.
- **1.1-train-bm25_stage2.ipynb**: Tiền xử lý và train bm25 (title, text) trên tập dữ liệu thô.

Các thư mục outputs sẽ được tạo ra sau khi hoàn tất việc huấn luyện:

- bm25_stage1
- bm25_stage2

Model Ranking:

Chạy các notebook:

- **1.2-train-pairwise-stage1.ipynb**: Sử dụng mô hình pre-trained model https://huggingface.co/nguyenvulebinh/vi-mrc-base và thêm các layer Dropout, , Linear kết hợp với một số kĩ thuật load data để train trên tập dữ liệu

train_stage1_ranking.csv đã tiền xử lý ở bước trước đó. Sử dụng Kfold cross-validation để train mô hình (chia dữ liệu thành nhiều fold, train trên n-1 fold và kiểm tra trên fold còn lại)

- **1.3-train-pairwise-stage2.ipynb:** Tương tự với bước trên nhưng trên tập dữ liệu *train_stage2_ranking.csv* và sử dụng kĩ thuật GroupKFold trên tập dữ liệu có group (tránh lặp lại các groups). Ở giai đoạn này vì mô hình khá lớn, mặc dù đã sử dụng TPU của kaggle nhưng 1 session chỉ có 9 giờ, do đó chỉ thực hiện train trên 2 epochs.

Sau khi hoàn thành, các file outputs sẽ được tạo ra:

- pairwise_v2.bin
- pairwise_stage2_seed0.bin
- bm25_stage1
- bm25_stage2

Model QA:

- **1.4-robust-qa-model.ipynb**: sử dụng pre-trained model để train mô hình MRCQuestionAnswering kết hợp với bộ dữ liệu Zalo Wiki 2019.

Deploy mô hình lên ứng dụng web:

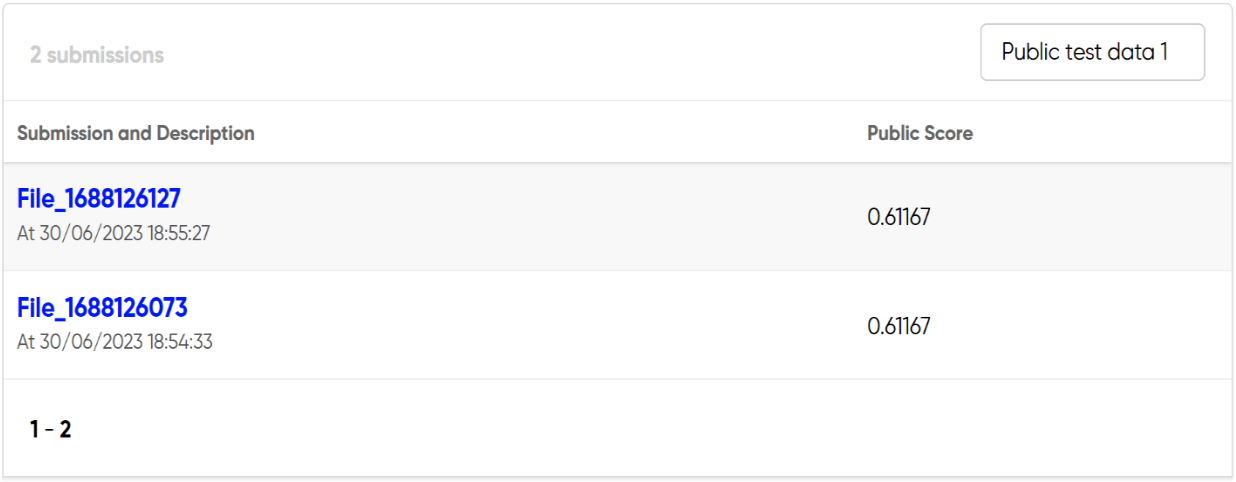
- Sử dụng **streamlit** để deploy ứng dụng lên môi trường web, cài đặt các hàm cần thiết để inference kết quả từ prompt của người dùng nhập vào

2. Kết quả thu thập được:

Kết quả thu thập thông qua việc trả lời các câu hỏi từ bộ public test được lưu trong file “submission/submisson.json”.

Kết quả thu thập được sau khi qua nhiều lần chạy thử. Môi trường thực hiện (deploy app hoặc thực hiện inference cần phải có nhân CUDA).

Kết quả submit public test trên trang chủ của cuộc thi



2 submissions

Public test data 1

Submission and Description	Public Score
File_1688126127 At 30/06/2023 18:55:27	0.61167
File_1688126073 At 30/06/2023 18:54:33	0.61167

1 - 2

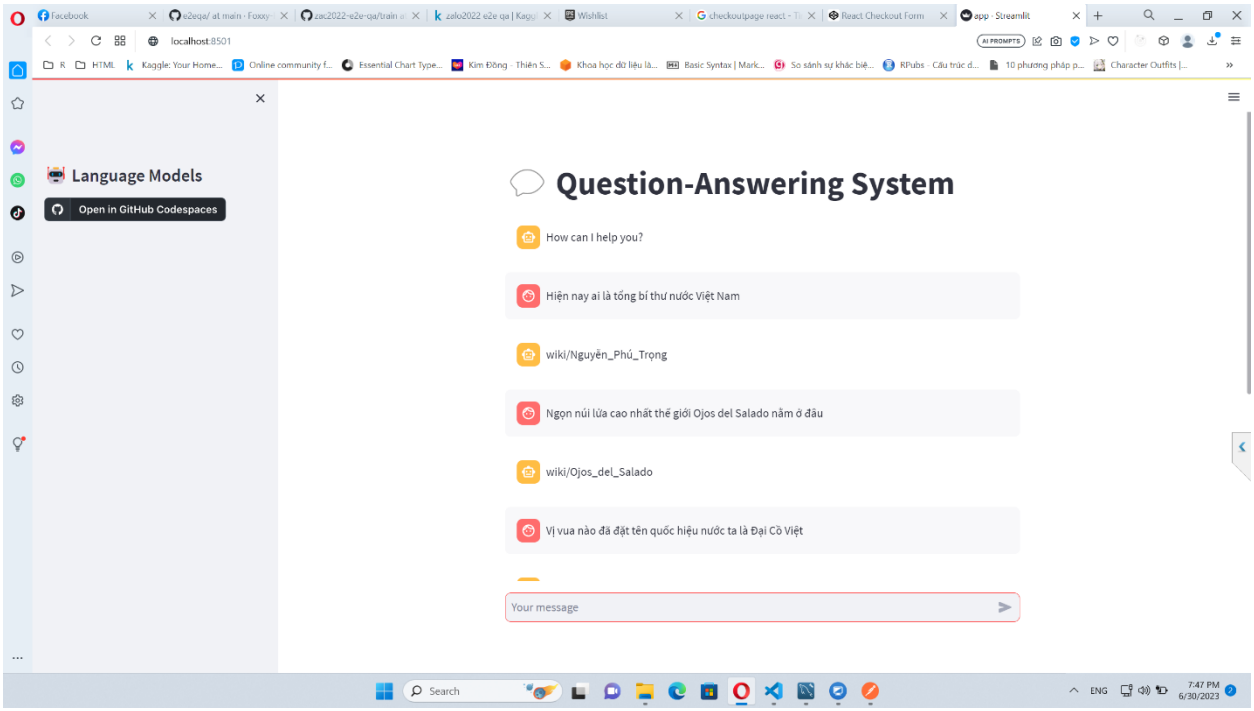
Kết quả được tính bằng Accuracy theo công thức:

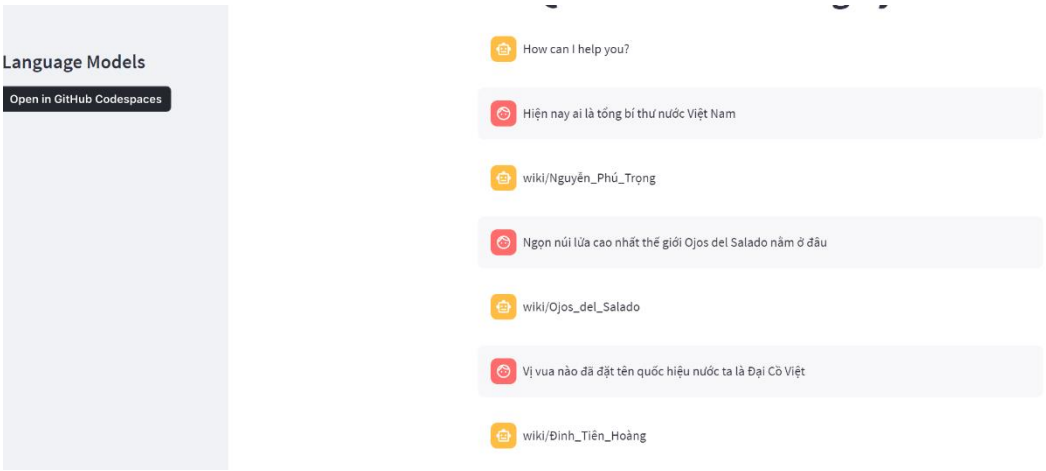
$$Accuracy = \frac{n_{correct}}{total\_item}$$

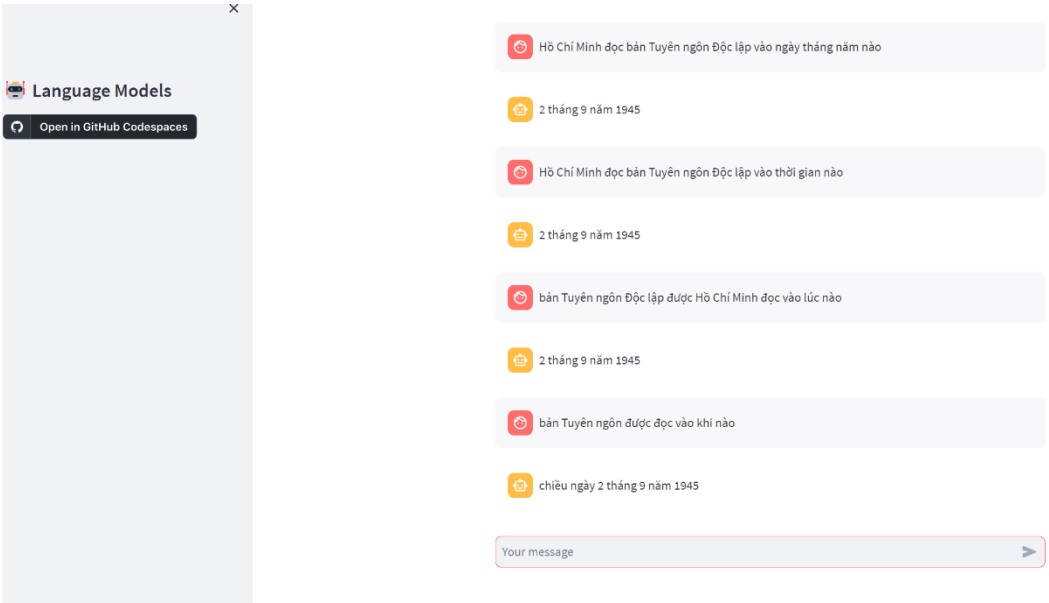
Trong đó:

- $n_{correct}$: Số mẫu đúng
- Total_item: Tổng số mẫu

Một số kết quả deploy model lên môi trường web:


Language Models
Open in GitHub Codespaces
Question-Answering System
How can I help you?
Hiện nay ai là tổng bí thư nước Việt Nam
wiki/Nguyễn_Phú_Trọng
Ngọn núi lửa cao nhất thế giới Ojos del Salado nằm ở đâu
wiki/Ojos_del_Salado
Vị vua nào đã đặt tên quốc hiệu nước ta là Đại Cồ Việt
Your message


Language Models
Open in GitHub Codespaces
How can I help you?
Hiện nay ai là tổng bí thư nước Việt Nam
wiki/Nguyễn_Phú_Trọng
Ngọn núi lửa cao nhất thế giới Ojos del Salado nằm ở đâu
wiki/Ojos_del_Salado
Vị vua nào đã đặt tên quốc hiệu nước ta là Đại Cồ Việt
wiki/Đinh_Tiên_Hoàng



III. Resources

Video demo:

- https://www.youtube.com/watch?v=dDLh55HYrfg&ab_channel=Miruku

Data + models:

- https://drive.google.com/drive/u/0/folders/1-ZX_puDxcRZqS41B5g6s0NHIyc1_JRaV
- Tải 2 folder data.zip + models.zip rồi giải nén vào thư mục e2eqa (hoặc tải folder.zip và chạy trực tiếp bằng câu lệnh > streamlit run app.py).

GitHub:

- https://github.com/Foxxy-HCMUS/e2eqa
- Mở app bằng các câu lệnh:
  > cd src
  > streamlit run app.py

References:

- https://challenge.zalo.ai/portal/e2e-question-answering
- https://github.com/Telegram-Zalo/zac2022-e2e-qa
- https://huggingface.co/
- https://streamlit.io/generative-ai