Project: Data Science Algorithms, Features Used , Model Deployment, and Client Side Design

Project Team Members- Avinash Ramesh, Charu Cheema, Cory Randolph
Project Team Name- Insight Finders

CMPE 256 - Fall, 2021
11/28/2021

9 Data Science Algorithms & Features Used

9.1 Algorithms & Features Overview

Our Question Answering Chatbot/System has two primary areas where Data Science Algorithms and Features are used: Wikipedia Features and NLP Model (see High-Level Architecture diagram for reference). The Wikipedia Features are focused on cleaning up the users free-form text input in a way that provides the most relevant wikipedia search results and context, from there the results/contexts are transformed to create suitable inputs for the next phase. The second area of the system focuses on extracting the most relevant answer to the question by using a pre-trained state-of-the-art NLP Question Answering Model.

9.2 Initial Data Science

During the exploration and development of this project we tried several different methods for cleaning and transforming the Wikipedia search results that were returned by the API. For example, we applied tokenization, pattern matching, and removal of stop words to the user's initial input and to the Wikipedia returned context in order to minimize the amount of noisy data that could enter into the NLP model.

As for the NLP model, we explored various models like BERT, ALBERT, and RoBERTa and found that the large version of RoBERTa was able to better Wikipedia contexts that were multiple paragraphs long and while providing succinct answer to the question based on the context.

9.3 Detailed Data Science Algorithms & Features Used

After settling on the primary Data Science Algorithms & Features for both the Wikipedia Features and the NLP Model, we transformed our code into production code by creating separate python modules organized by functional needs so that the code of our main function would be clean, easy to understand, and easy to maintain and modify.

While the High-Level Architecture diagram provides the main logic of our system, the highlights and summary of our code and modules below will demonstrate the details of the algorithms and features used.

1. Load Spacy Universal Sentence Encoder(Google's Universal Sentence Encoder)
    a. Pre-trained English sentence encoder
    b. Converts english language in to numerical vectors for NLP models
2. Collect and Process Users Input
    a. Receive the raw text input question from the user
    b. Apply tokenization and error handling to the users input to better flag relevant parts of speech
3. Retrieve Wikipedia Results

    a. Pass the transformed user input into the Wikipedia API to get several search results and contexts (body text of a Wikipedia page)
    b. Process and clean several different page results and append them to a master document
4. Match the best Wikipedia contexts to the query
    a. Apply content extraction and phrase matching
    b. Calculate similarity based on encoded sentence vectors
    c. Sort and rank based on highest similarity and return top N results (N = 10 in demo)
5. Store sorted results
    a. Convert data into a Pandas Dataframe and apply “data wrangling” to remove special characters and other common free form text issues
    b. Merge all similar contexts into one Dataframe
6. Send cleaned and rank’s contexts to the NLP model
    a. Choose RoBERTa model since it provided the best results
    b. Instantiate the RoBERTa model (Model is Cached in memory to speed up performance in production system)
7. Collect NLP model results
    a. Pair the top NLP answers with the context and other Wikipedia info for as the final results
8. Return answers, image and context based on initial user question
    a. Final results and most relevant information (answer, Image, and context) are returned in a format that is easily processed by the user interface

10 Client Side Design

Rather than just leaving our project in a Notebook where a user would have to run the code, we chose to build a clean user interface for the user to type in their question and then hit submit to retrieve results.

We intended to use the Streamlit Python library to create a simple web interface for our NLP project. Streamlit is a Python open-source framework for creating web apps for Machine Learning and Data Science.

In our client-side design, the user would type in a search query in a web application. After a brief waiting period the user receives ranked results with an image, Wikipedia link and text for reference.

Deployment URL: *https://cmpe256-q4uake3apq-uc.a.run.app*

USER QUERY:

Wiki QA System

Type your Query

who is Mark Zuckerberg?

Get Answers!

RESULTS:

FACEBOOK FOUNDER

Chan Zuckerberg Initiative

wiki: https://en.wikipedia.org/wiki/Chan_Zuckerberg_Initiative

The Chan Zuckerberg Initiative CZI is an organization established and owned by Facebook founder Mark Zuckerberg and his wife Priscilla Chan with an investment of 99 percent of the couple s wealth from their Facebook shares over their lifetime The CZI is set up as a limited liability company LLC and is an example of philanthrocapitalism CZI has been deemed likely to be one of the most well funded philanthropies in human history Its creation was announced on 1 December 2015 for the birth of their daughter Maxima Chan Zuckerberg Priscilla Chan has said that her background as a child of immigrant refugees and experience as a teacher and pediatrician for vulnerable children influences how she approaches the philanthropy s work in science education immigration reform housing criminal justice and other local issues The Chan Zuckerberg Initiative s main areas of work include Science Education and Justice and Opportunity which focuses on promoting housing affordability criminal justice reform and immigration reform The mission of the Chan Zuckerberg Initiative is to build a more inclusive just and healthy future for everyone and to advance human potential and promote equality in areas such as health education scientific research and energy In 2017 the Chan Zuckerberg Initiative pre leased a 102 079 square foot portion of the new Broadway Station development in downtown Redwood City California where it is headquartered

FACEBOOK FOUNDER

wiki: https://en.wikipedia.org/wiki/Priscilla_Chan

Personal life Chan married Facebook founder Mark Zuckerberg on May 19 2012 the day after the site s IPO Chan and Zuckerberg announced the birth of their daughter Maxima Chan Zuckerberg on December 1 2015 On August 28 2017 Chan gave birth to their second daughter whom they named August According to a Facebook post by Zuckerberg Chan is a Buddhist

FOUNDER AND CEO

wiki: https://en.wikipedia.org/wiki/Mark_Zuckerberg

On October 1 2012 Zuckerberg visited Russian Prime Minister Dmitry Medvedev in Moscow to stimulate social media innovation in Russia and to boost Facebook s position in the Russian market Russia s communications minister tweeted that Prime Minister Dmitry Medvedev urged the social media giant s founder to abandon plans to lure away Russian programmers and instead consider opening a research center in Moscow In 2012 Facebook had roughly 9 million users in Russia while domestic clone VK had around 34 million Rebecca Van Dyck Facebook s head of consumer marketing said that 85 million American Facebook users were exposed to the first day of the Home promotional campaign on April 6 2013 On August 19 2013 The Washington Post reported that Zuckerberg s Facebook profile was hacked by an unemployed web developer At the 2013 TechCrunch Disrupt conference held in September Zuckerberg stated that he is working towards registering the 5 billion people who were not connected to the Internet as of the conference on Facebook Zuckerberg then explained that this is intertwined with the aim of the Internet org project whereby Facebook with the support of other technology companies seeks to increase the number of people connected to the internet Zuckerberg was the keynote speaker at the 2014 Mobile World Congress MWC held in Barcelona Spain in March 2014 which was attended by 75 000 delegates Various media sources highlighted the connection between Facebook s focus on mobile technology and Zuckerberg s speech stating that mobile represents the future of the company Zuckerberg s speech expands upon the goal that he raised at the TechCrunch conference in September 2013 whereby he is working towards expanding Internet coverage into developing countries Alongside other American technology figures like Jeff Bezos and Tim Cook Zuckerberg hosted visiting Chinese politician Lu Wei known as the Internet czar for his influence in the enforcement of China s online policy at Facebook s headquarters on December 8 2014 The meeting occurred after Zuckerberg participated in a Q A session at Tsinghua University in Beijing China on October 23 2014 where he attempted to converse in Mandarin Chinese although Facebook is banned in China Zuckerberg is highly regarded among the people and was at the university to help fuel the nation s burgeoning entrepreneur sector Zuckerberg fielded questions during a live Q A session at the company s headquarters in Menlo Park on December 11 2014 The founder and CEO explained that he does not believe Facebook is a waste of time because it facilitates social engagement and participating in a public session was so that he could learn how to better serve the community Zuckerberg receives a one dollar salary as CEO of Facebook In June 2016 Business Insider named Zuckerberg one of the Top 10 Business Visionaries Creating Value for the World along with Elon Musk and Sal Khan due to the fact that he and his wife pledged to give away 99 of their wealth which is estimated at 55 0 billion In January 2019 Zuckerberg laid plans to integrate an end to end encrypted system for three major social media platforms including Facebook Instagram and WhatsApp On August 14 2020 Facebook integrated the chat systems for Instagram and Messenger on both iOS and Android devices The update encouraged cross communication between Instagram and Facebook users

AN AMERICAN MEDIA MAGNATE INTERNET ENTREPRENEUR AND PHILANTHROPIST

wiki: https://en.wikipedia.org/wiki/Mark_Zuckerberg

Mark Elliot Zuckerberg born 1984 05 14 May 14 1984 is an American media magnate internet entrepreneur and philanthropist He is known for co founding Meta Platforms Inc formerly named Facebook Inc and serves as its chairman chief executive officer and controlling shareholder He also is a co founder of the solar sail spacecraft development project Breakthrough Starshot and serves as one of its board members Zuckerberg attended Harvard University where he launched the Facebook social networking service from his dormitory room in February 2004 with his roommates Eduardo Saverin Andrew McCollum Dustin Moskovitz and Chris Hughes Originally launched to select college campuses the site expanded rapidly and eventually beyond colleges reaching one billion users by 2012 Zuckerberg took the company public in May 2012 with majority shares In 2007 at age 23 he became the world s youngest self made billionaire As of November 2021 Zuckerberg s net worth was 126 billion making him the 7th richest person in the world Since 2008 Time magazine has named Zuckerberg among the 100 most influential people in the world as a part of its Person of the Year award which he was recognized with in 2010 In December 2016 Zuckerberg was ranked 10th on Forbes list of The World s Most Powerful People

ENTREPRENEUR

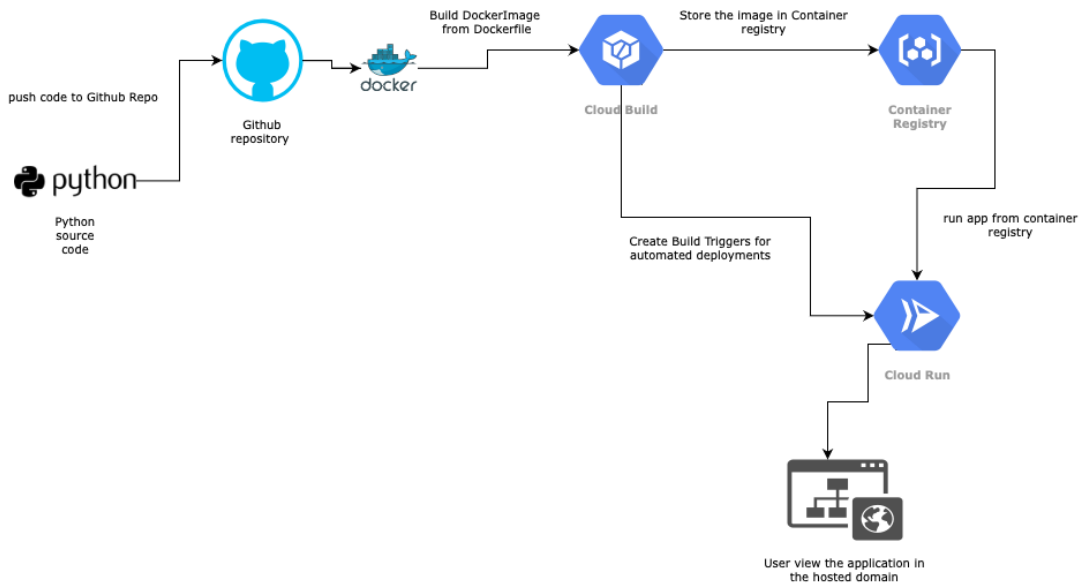
wiki: https://en.wikipedia.org/wiki/Mark_Zuckerberg

A movie based on Zuckerberg and the founding years of Facebook The Social Network was released on October 1 2010 starring Jesse Eisenberg as Zuckerberg After Zuckerberg was told about the film he responded I just wished that nobody made a movie of me while I was still alive Also after the film s script was leaked on the Internet and it was apparent that the film would not portray Zuckerberg in a wholly positive light he stated that he wanted to establish himself as a good guy The film is based on the book The Accidental Billionaires by Ben Mezrich which the book s publicist once described as big juicy fun rather than reportage The film s screenwriter Aaron Sorkin told New York magazine I don t want my fidelity to be the truth I want it to be storytelling adding What is the big deal about accuracy purely for accuracy s sake and can we not have the true be the enemy of the good Upon winning the Golden Globe Award for Best Picture on January 16 2011 producer Scott Rudin thanked Facebook and Zuckerberg for his willingness to allow us to use his life and work as a metaphor through which to tell a story about communication and the way we relate to each other Sorkin who won for Best Screenplay retracted some of the impressions given in his script I wanted to say to Mark Zuckerberg tonight if you re watching Rooney Mara s character makes a prediction at the beginning of the movie She was wrong You turned out to be a great entrepreneur a visionary and an incredible altruist On January 29 2011 Zuckerberg made a surprise guest appearance on Saturday Night Live which was hosted by Jesse Eisenberg They both said it was the first time they had met Eisenberg asked Zuckerberg who had been critical of his portrayal by the film what he thought of the movie Zuckerberg replied It was interesting In a subsequent interview about their meeting Eisenberg explained that he was nervous to meet him because I had spent now a year and a half thinking about him He added Mark has been so gracious about something that s really so uncomfortable The fact that he would do SNL and make fun of the situation is so sweet and so generous It s the best possible way to handle something that I think could otherwise be very uncomfortable

11 Model Deployment

To minimize the complexity and time to run this project all from within a local or cloud hosted notebook we packaged the whole project into a Docker image and hosted it on Google Cloud Platform.

Architecture Diagram:



Github Repository:

Our entire source code, dependencies, and Dockerfile have been uploaded to the Github repository.

Github Repo URL: *https://github.com/coryroyce/wiki_based_nlp_chat_bot*

Docker File:

Dockerfile contains the list of commands sent to the docker engine to build the image.

GCP Cloud Build:

GCP Cloud Build enables us to build the container image, store the built image in the Container Registry, and then deploy the image to Cloud Run.

GCP Container Registry:

Container Registry is a location for managing Docker images, performing vulnerability analysis, and deciding who has access to what using fine-grained access control. When you commit code to Cloud Source Repositories, GitHub, or Bitbucket, we can automatically build and push images to a private registry.

Deploy Docker Image on GCP Cloud Run:

The final step of the process is to deploy the GCP docker image on Cloud Run. Cloud Run helps to develop and deploy highly scalable containerized applications on a fully managed serverless platform. We can also automate the deployment of your NLP application to Cloud Run by creating Cloud Build triggers.

User Access:

Once the preceding steps have been completed, the user can use our NLP web application to retrieve results for his query.

12 References

- https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2749028.pdf
- https://chatbotslife.com/chatbot-qa-implementing-chatbot-solution-the-chatc-group-49acef43aaa9
- https://medium.com/@christophberns/using-crisp-dm-to-predict-car-prices-f15eb5b14025
- http://web.stanford.edu/class/cs224n/project/default-final-project-handout.pdf
- https://web.cse.ohio-state.edu/~bair.41/616/Project/Example_Document/Req_Doc_Example.html
- https://towardsdatascience.com/bert-to-the-rescue-17671379687f