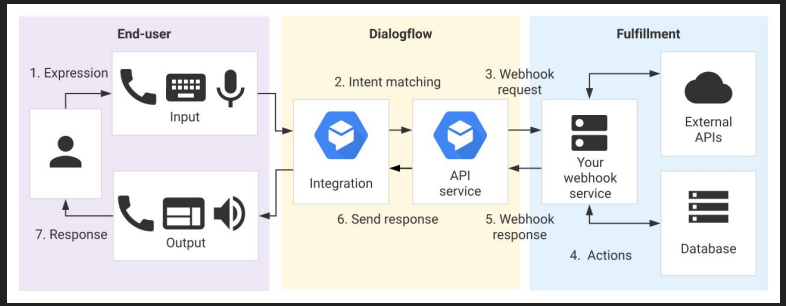
Cloud Infrastructure for Portable AI Milestone

Nathan Dolbir, Kennedy Fairey, Anirudh Oruganti

Our Mission

- To prove individual developers can create powerful ML systems on portable devices using cloud computing
- This has been done by large corporations, but not on an individual and small development basis

Alexa, the cloud ML industry standard.

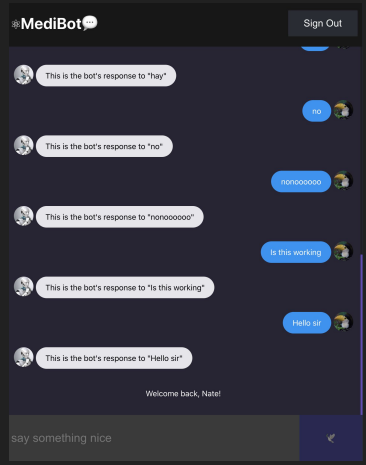


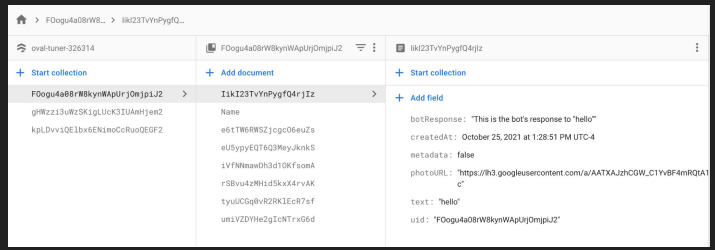
Dialogflow abstracts ML into an API rather than provide a way to implement your own models

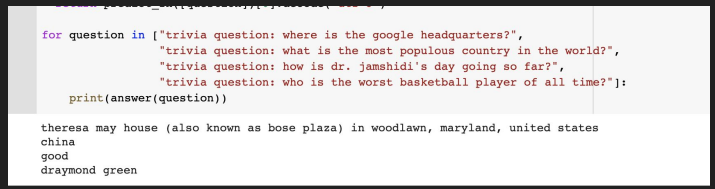
Project Status

Infrastructure

- End to end data pipeline
- Cloud Models -> Firebase -> Frontend UI







Progress

- Our first goal is to implement our end to end infrastructure with an already functioning model (T5 Transformer Model)
- This model answers any type of trivia question
- Model has been deployed in GCP
- UI is functional and has been deployed
- Need to completely connect backend to frontend

Project Status

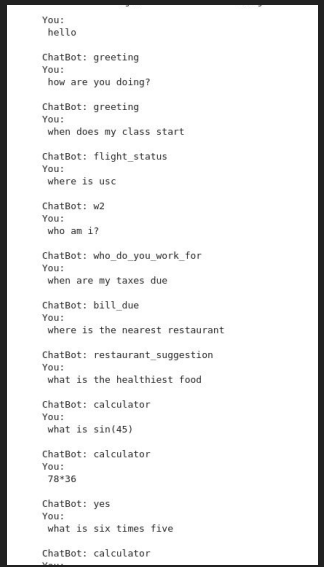
Chatbot Models

- Term Frequency - Inverse Document Frequency (TF-IDF)
- Text to Text Transfer Transformer (T5)

```
for question in ["trivia question: where is the google headquarters?",
                 "trivia question: what is the most populous country in the world?",
                 "trivia question: how is dr. jamshidi's day going so far?",
                 "trivia question: who is the worst basketball player of all time?"]:
    print(answer(question))
```

```
theresa may house (also known as bose plaza) in woodlawn, maryland, united states
china
good
draymond green
```

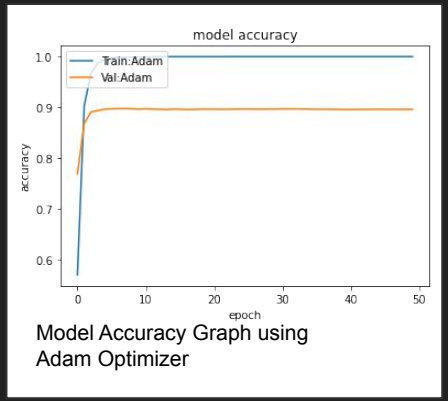
T5 Model using Colab Notebook, stored in Cloud



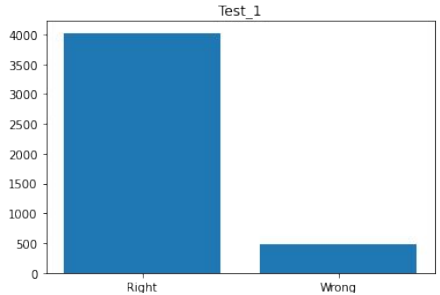
TF-IDF for sequence classification

About the TF-IDF Model

- Built using Sklearn and Keras Api.
- Uses Bag-of-Words approach for vectors.
- Uses “Tanh” Activation Layer.
- Uses Adam Optimizer



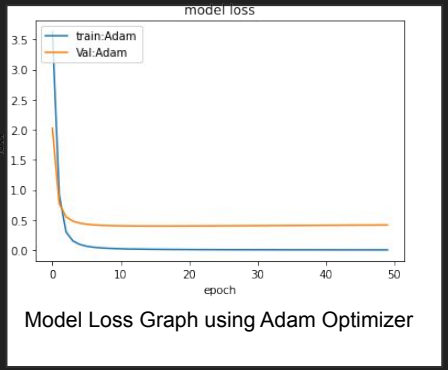
Model Accuracy Graph using Adam Optimizer



Number Correct and Incorrect Intent Prediction of 4500 Sentence Data Set

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_6 (Dense)              (None, 3658)              13384622
_________________________________________________________________
dense_7 (Dense)              (None, 1219)              4460321
_________________________________________________________________
layer2 (Dense)               (None, 150)               183000
=================================================================
Total params: 18,027,943
Trainable params: 18,027,943
Non-trainable params: 0
_________________________________________________________________
```

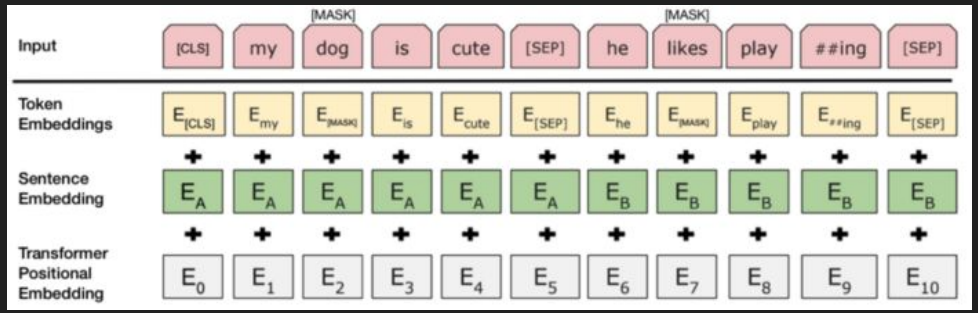
Model Summary



Model Loss Graph using Adam Optimizer

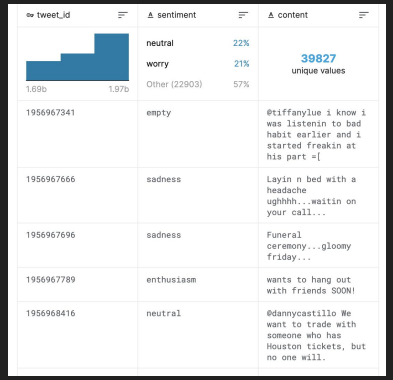
Bert Model

- BERT (Bidirectional Encoder Representations from Transformers) is state-of-the-art model made by Google that specializes in NLP tasks such as q&a based on text, classification tasks like sentimental analysis, etc.
- Text is processed bidirectional unlike RNN using a single directional sequence either from left to right or right to left, it encodes entire text at once.
- State-of-Art model results can be easily achievable via fine tuning pre-trained bert this model.



Models Plan

- Sentiment classifier (TF-BERT for Sequence Classification)
  - Simple BERT with different dataset
- Conversational Response (TF Transformer)
  - Sequence to sequence Transformer
  - Given classification matrix and blank sentence, fill sentence with dataset words assuming the classification



tweet_id	sentiment	content
1.69b – 1.97b	neutral 22% worry 21% Other (22903) 57%	39827 unique values
1956967341	empty	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[
1956967666	sadness	Layin n bed with a headache ughhhh...waitin on your call...
1956967696	sadness	Funeral ceremony...gloomy friday...
1956967789	enthusiasm	wants to hang out with friends SOON!
1956968416	neutral	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.

Dataset for sentiment classification

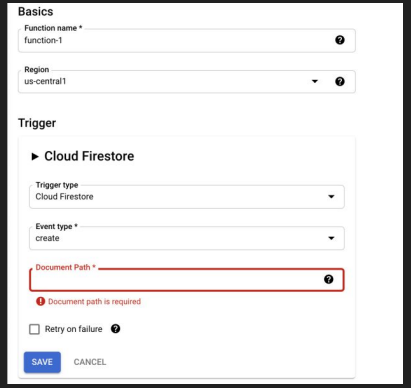

```
▾ t_34351a42-4b9d-47e2-99f2-f32efb789a9c:
    article_url: "https://www.washingtonpost.com/goingoutguide/movies/mission-impossible--fallout-is-big-
     y.html"
    config: "A"
  ▾ content: [] 22 items
    ▾ 0:
        message: "hi how are you doing over there? "
        agent: "agent_1"
        sentiment: "Neutral"
      ▸ knowledge_source: [] 1 item
        turn_rating: "Passable"
    ▾ 1:
        message: "Fine. Are you a movie fan?"
        agent: "agent_2"
        sentiment: "Curious to dive deeper"
      ▸ knowledge_source: [] 7 items
        turn_rating: "Excellent"
    ▾ 2:
        message: "yes, i am a movie fan. I think that there are many fine movies done by Tom Cruise."
        agent: "agent_1"
        sentiment: "Neutral"
      ▸ knowledge_source: [] 1 item
        turn_rating: "Good"
```

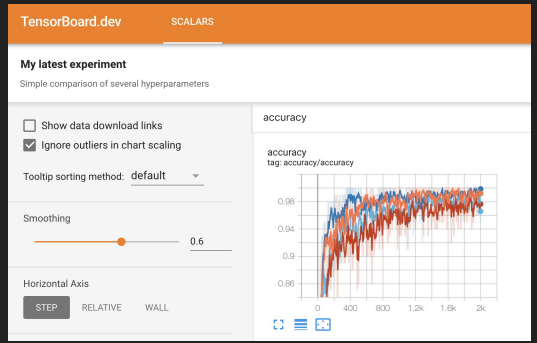
Progress

- Our next goal is to seemingly combine both of our preliminary models into a singular functioning model
- Ideally it will answer any trivia question and also be conversational
- Once the UI is connected to the T5 model and our chatbot is correctly predicting what to say, then we can integrate our new datasets and re-train

Project Tribulations and Goals

- Create Cloud Functions
- Create BERT Classifier and Transformer Seq2Seq models, integrate and deploy
  - Heavy workload on DB-cloud involved
- Polish UI, test on phone
- TensorBoard Model Evaluation
- Cloud Analytics for Infrastructure





References

https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html