

#11915010 Raghu Punnamraju
#11915043 Anmol More
#11915001 Sriganesh Balamurugan
#11915052 Kapil Bindal

Chatbot URL - <https://bot.dialogflow.com/7ece3edf-bff2-4527-ae49-c68209b471ee>

Telegram Bot username – @AAIAssignmentCorona_Bot

Table of Contents

I. Data Collection	2
Methodology	2
Sources Used	2
NLP Pipeline	2
Codebase explanation.....	2
Pre-processed Dataset description.....	3
Clean Knowledge Base.....	4
II. Version 1: Build a simple similarity-based FAQ bot	4
Command line model	4
Question similarity approach	4
Steps to run	4
Flask Deployable model.....	5
Bi directional similarity approach.....	5
III. Version 2: Build a Dialogflow based FAQ bot.....	7
Knowledge based corpus	7
Intent Action and Parameters	7
Realtime webhook.....	8
Web Based Bot.....	10
Telegram Bot	12
IV. Discussion and Ways to Improve	13
Personas.....	13
Ideal Conversation Flowchart	14
Ways to improve your bot	14
Improve user experience	14
Improve bot knowledge base.....	14
Multiple model deployment	15
Realtime Integration	15
References.....	15

I. Data Collection

Methodology

We focused on the following data collection principles – reliability, simplicity, contemporariness and automation & collected 400+ FAQs from International and government of India websites.

- **Reliability:** It is absolutely critical that the data is reliable and at no cost we can be the reason to spread false information. We limited our data sources for statistics to world meters and WHO. For health ad clinical information, we limited the data sources like WebMD, who, etc.
- **Simplicity:** While gathering the FAQs, we focused on simplicity as it is absolutely critical for the users to understand and relate the responses. Information must be understood and explained in simple English. This enabled us to reduce our processing needs later.
- **Contemporariness:** We made sure no data source is older than one week. As the statistics and care suggestions are changing on a daily basis, we did not want to respond with stale information.
- **Automation:** We gave automation the less priority as we wanted to focus on the quality and simplicity first. Our automation was limited to the data processing as opposed to data collection. We did the data collection manually.

Sources Used

1. European Centre for Disease Prevention and Control - <https://www.ecdc.europa.eu/en/covid-19/questions-answers>
2. Official Website Belgium - <https://www.info-coronavirus.be/en/faq/>
3. Govt of Karnataka Website - <https://www.mohfw.gov.in/pdf/FAQ.pdf>
4. WHO - <https://www.who.int/news-room/q-a-detail/q-a-coronaviruses>
5. Government of India website - <https://www.mohfw.gov.in/pdf/FAQ.pdf>

NLP Pipeline

As explained in diagrams below, we used two different methodologies to build pipeline as explained below

Codebase explanation

- *FAQ_codebase_1.py* – This contains code to combine multiple data sources collected and then clean and create json format dataset for Bert based model. Additionally a csv set is created to run different models based on similarity score. Sample JSON format output :

```
{
  'title': 'General',
  'paragraphs': [
    {
      'context': 'COVID 19 is the infectious disease caused by the most recently discovered coronavirus. This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019',
      'qas': [
        {
          'answer_start': 156,
          'text': 'Wuhan, China, in December 2019',
          'answer_start': 156,
          'text': 'Wuhan, China, in December 2019'
        },
        {
          'answer_start': 156,
          'text': 'Wuhan, China, in December 2019',
          'answer_start': 156,
          'text': 'Wuhan, China, in December 2019'
        }
      ],
      'question': 'When does corona disease started',
      'id': 'KWCAH7Wqdk757HhaoSQVGV',
      'answers': [
        {
          'answer_start': 0,
          'text': 'COVID 19 is the infectious disease caused by the most recently discovered coronavirus',
          'answer_start': 0,
          'text': 'COVID 19 is the infectious disease caused by the most recently discovered coronavirus'
        },
        {
          'answer_start': 0,
          'text': 'COVID 19 is the infectious disease caused by the most recently discovered coronavirus',
          'answer_start': 0,
          'text': 'COVID 19 is the infectious disease caused by the most recently discovered coronavirus'
        }
      ]
    }
  ]
}
```

- *FAQ_codebase_2.py* – This contains code to use BERT based model and return output based on user's query

```

(base) anmol@x86_64-apple-darwin13 assignment % python FAQ_codebase_2.py
2020-04-03 12:44:40.546705: I tensorflow/core/platform/cpu_feature_guard.cc:142] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2 FMA
2020-04-03 12:44:40.556872: I tensorflow/compiler/xla/service/service.cc:168] XLA service 0x7fdab4afadb0 initialized for platform Host (this does not guarantee that XLA will be used). Devices:
2020-04-03 12:44:40.556901: I tensorflow/compiler/xla/service/service.cc:176]   StreamExecutor device (0): Host, Default Version
Query : Can I travel ?

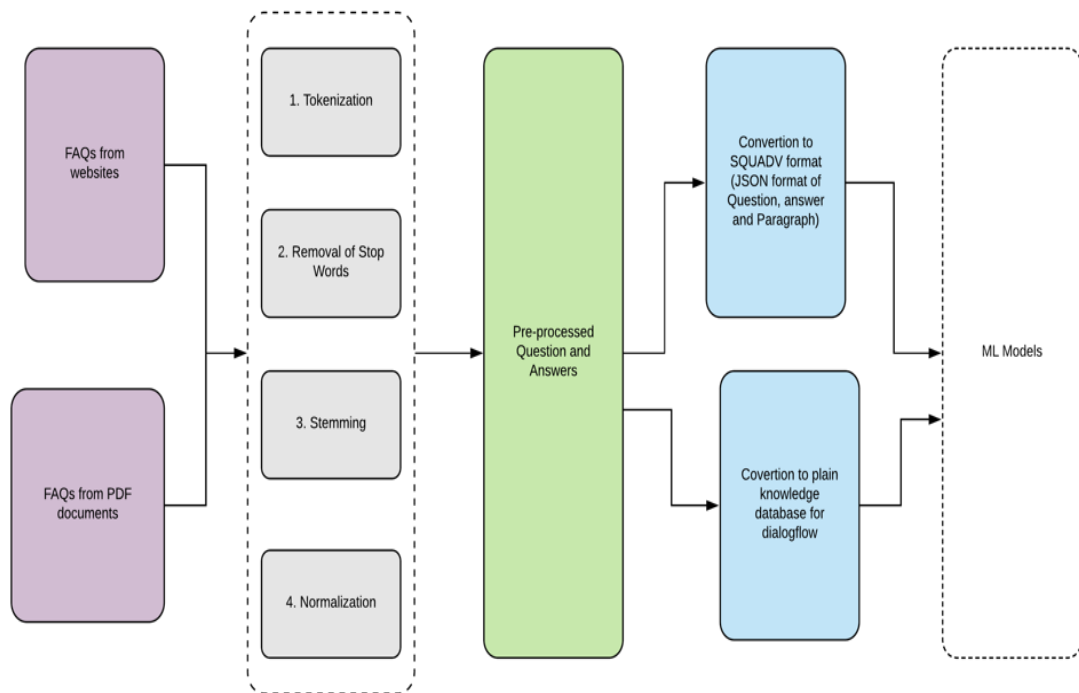
Answer from Bot: Any non-essential travel abroad is prohibited

Matched to Question : Is it still allowed to travel abroad?

Paragraph pickup from : Any non-essential travel abroad is prohibited. The following reasons in particular are considered to be essential reasons for travelling abroad:

(base) anmol@x86_64-apple-darwin13 assignment % █

```



Pre-processed Dataset description

Pre processed dataset is submitted as 'dataset_collected.csv' with following fields

- Title – Describes group of queries like 'General Information', 'Info from Government' etc.
- Question – Question that can be answered from a given paragraph
- Context – Paragraph collected
- Text – Answer to the question
- Source – Website, PDFs, etc
- Source URL – URL of data collection
- Organization – Organization from which data was collected
- Country – If related to specific country, region, otherwise International
- Last Retrieved Time – Time of data collection

Clean Knowledge Base

This file has been submitted as 'knowledge_base.csv'

II. Version 1: Build a simple similarity-based FAQ bot

Two different versions of program are built. Both can run on command line, but for smaller size and speed we have provided detailed steps for a simpler model.

Command line model

Though we have tried models, we are delivering similarity based approach as a command line model. As this is fast compared to bert models, which we tried as flask deployable model.

Question similarity approach

At first we create embedding with our own question corpus to find similar questions that have been manually annotated. Next whole question is converted to vector embeddings.

Now, for every query that user asks a vector to vector comparison is done using cosine similarity and most similar question is looked up. The answer to most similar question is sent as reply by bot.

Steps to run

Please use virtual env and install packages from requirements.txt if direct run of python_run.py fails

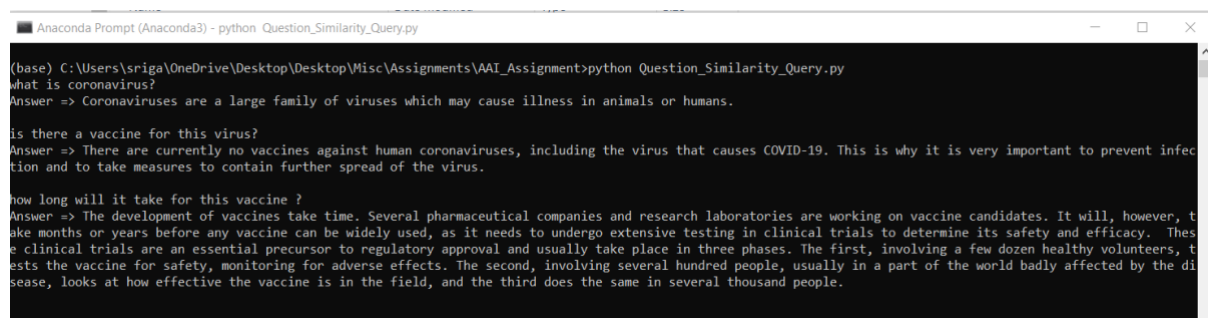
Please make sure data/knowledge_base.csv and models/*.pkl files are present in current directory

```
>python python_run.py
```

Detailed steps to build and run (Takes less than 60 seconds)

1. `conda create --name aaiassignment`
2. `conda activate aaiassignment`
3. `pip install -r requirements.txt`
4. `python Model_Codebase_1_Question_Similarity_Featurization.py`
5. `python python_run.py`

Screenshot from two different machines of running code



```
Anaconda Prompt (Anaconda3) - python Question_Similarity_Query.py

(base) C:\Users\sriga\OneDrive\Desktop\Desktop\Misc\Assignments\AAI_Assignment>python Question_Similarity_Query.py
what is coronavirus?
Answer => Coronaviruses are a large family of viruses which may cause illness in animals or humans.

is there a vaccine for this virus?
Answer => There are currently no vaccines against human coronaviruses, including the virus that causes COVID-19. This is why it is very important to prevent infection and to take measures to contain further spread of the virus.

how long will it take for this vaccine ?
Answer => The development of vaccines take time. Several pharmaceutical companies and research laboratories are working on vaccine candidates. It will, however, take months or years before any vaccine can be widely used, as it needs to undergo extensive testing in clinical trials to determine its safety and efficacy. These clinical trials are an essential precursor to regulatory approval and usually take place in three phases. The first, involving a few dozen healthy volunteers, tests the vaccine for safety, monitoring for adverse effects. The second, involving several hundred people, usually in a part of the world badly affected by the disease, looks at how effective the vaccine is in the field, and the third does the same in several thousand people.
```

```
(aaassignment) anmol@x86_64-apple-darwin13 AAI Assignment % python python_run.py
Welcome to Corona Chatbot !
Please enter your query, press ctrl +c to exit
What is corona
Answer => COVID-19 is the infectious disease caused by the most recently discovered coronavirus

Please enter your query, press ctrl +c to exit
How does it work
Answer => If someone experiences minor symptoms of illness (a cough (irritating dry cough), sore throat, shortness of breath and/or high temperature, feverish
ness, muscle ache), they will not be tested for the novel coronavirus (unless they are at risk because they have an underlying, pre-existing condition). This
person then remains voluntarily in so-called self-isolation at home, minimising their contact with other people, including those living in the same household.
This way they can avoid infecting others.

You can end the period of self-isolation yourself once you have been free of symptoms for 48 hours (see the page on 'Self-isolation and self-quarantine').

People with symptoms of illness for whom a doctor has decided a test is necessary: If the test is positive, the person remains isolated until 48 hours after t
hey have recovered, and at least ten days from the moment their symptoms of illness appeared (see the fact sheet on 'Self-isolation' further down in the D
ocuments tab). Because it is not clear when the risk of passing on the infection ends, even after this they should continue to follow the rules of hygiene and
how to behave.

Please enter your query, press ctrl +c to exit
```

Flask Deployable model

Bi directional similarity approach

Here we took a pretrained bert model of nlp, and retrained it using our dataset for multiple epochs. The final model is a Question-Paragraph based model, which looks for similarity in question and paragraph and returns the sub content from paragraph as output.

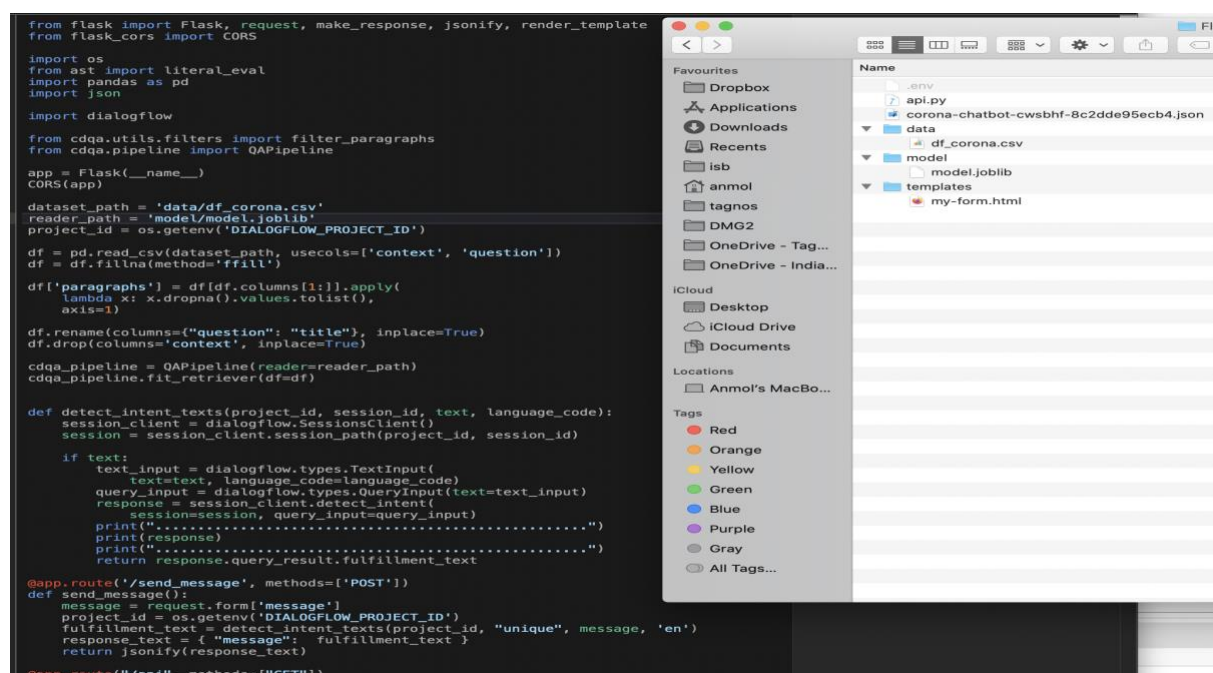
This relies on json format structure for data

Model : Due to heavy model size that couldn't fit even after compression, we are submitting a Onedrive URL (Please request for access if un accessible) – https://isbhydromoh-my.sharepoint.com/:f:/g/personal/anmol_more_cba2020s_isb_edu/Ei1RNOdbSvxDqK6EZm5Q4QIBrTPJvAfsP8FFbzXQSCbVhw?e=9JsJCV

Configuration required :

- This requires additional configuration with gcloud APIs to connect it to dialogflow and can't be run directly, hence providing running code output.
- Below folder structure can be used to run a flask server to serve requests on local machine.

Steps to run : python api.py



```
[(base) anmol@x86_64-apple-darwin13 Final Submission % python python_run.py
2020-04-03 15:00:21.477242: I tensorflow/core/platform/cpu_feature_guard.cc:142]
2020-04-03 15:00:21.497499: I tensorflow/compiler/xla/service/service.cc:168] XLA/
2020-04-03 15:00:21.497529: I tensorflow/compiler/xla/service/service.cc:176]
Welcome to Corona Chatbot ! How can I help you ?
Press enter twice to quit
I am worried about covid-19
Query : I am worried about covid-19
```

Reply from Bot: it is not clear or known if the spread of COVID-19 will decrease

Should I wear a mask to protect myself

Query : Should I wear a mask to protect myself

Reply from Bot: to avoid infecting other people

Can I go to my holiday home, for example in the Ardennes or the coast?

Query : Can I go to my holiday home, for example in the Ardennes or the coast?

Reply from Bot: Beauticians cannot offer their services at home

How can I avoid getting infected?

Query : How can I avoid getting infected?

Reply from Bot: by car

How long in incubation period for covid-19

Query : How long in incubation period for covid-19

Reply from Bot: 1- 14 days

What are the symptoms of coronavirus disease?

Query : What are the symptoms of coronavirus disease?

Reply from Bot: mild upper-respiratory tract illnesses

Where can I get tested?

Query : Where can I get tested?

Reply from Bot: here you will be advised if a medical assessment is necessary

Do schools remain open?

Query : Do schools remain open?

Reply from Bot: No, essential sectors can remain open

Where can i get help

Query : Where can i get help

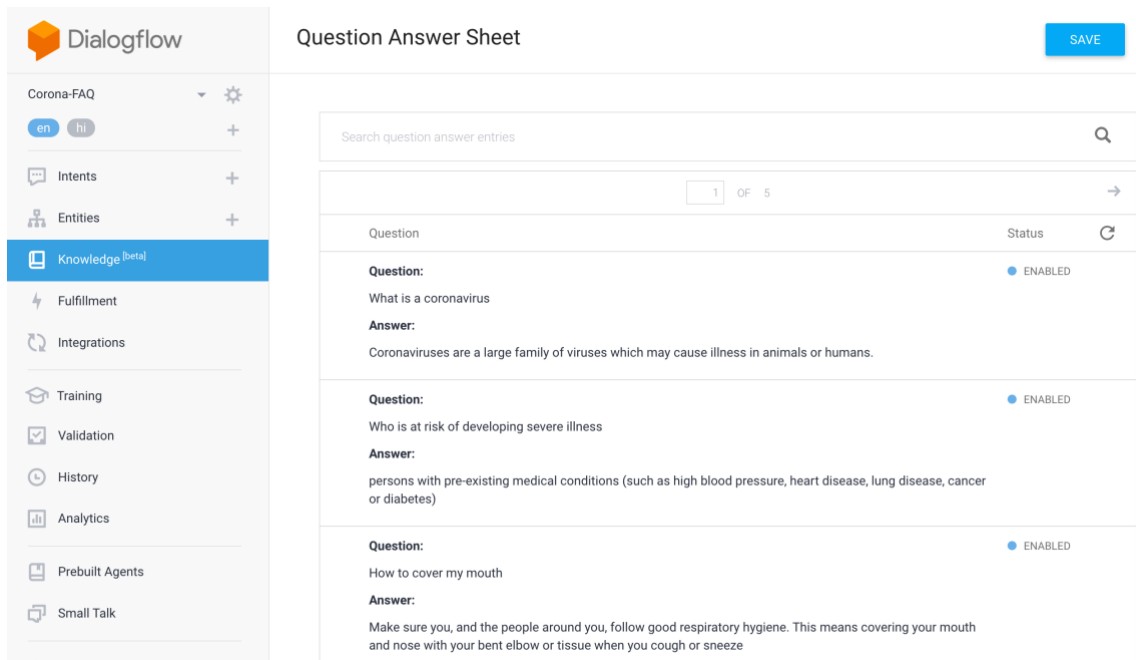
Reply from Bot: Avoiding contact with others and visits to medical facilities

III. Version 2: Build a Dialogflow based FAQ bot

In order to build dialog flow we used knowledge based corpus, custom intents, custom action and parameters and Realtime webhook to do backend call.

Knowledge based corpus

Question and Answer columns from extracted & processed dataset was used. This was fed to dialog flow. Screenshots below



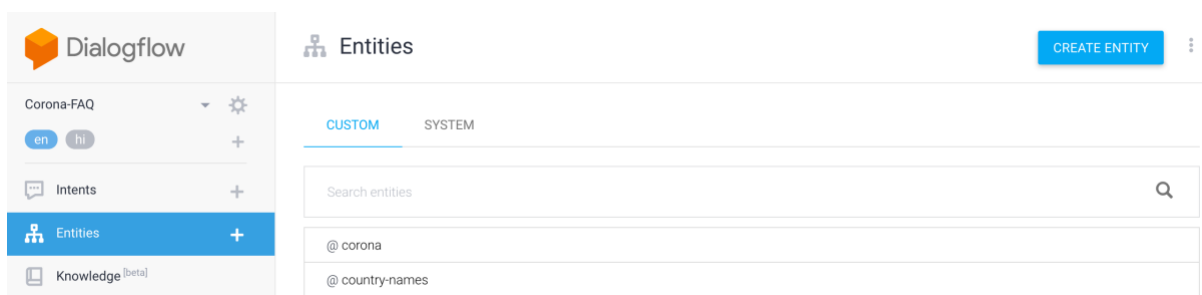
The screenshot shows the Dialogflow console interface for a project named 'Corona-FAQ'. The left sidebar contains navigation options: Intents, Entities, Knowledge (beta), Fulfillment, Integrations, Training, Validation, History, Analytics, Prebuilt Agents, and Small Talk. The 'Knowledge' section is selected, displaying a 'Question Answer Sheet'. At the top right of this section is a 'SAVE' button. Below the header is a search bar labeled 'Search question answer entries'. A table lists three entries, each with a 'Question', an 'Answer', and a status of 'ENABLED'.

Question	Answer	Status
Question: What is a coronavirus	Answer: Coronaviruses are a large family of viruses which may cause illness in animals or humans.	ENABLED
Question: Who is at risk of developing severe illness	Answer: persons with pre-existing medical conditions (such as high blood pressure, heart disease, lung disease, cancer or diabetes)	ENABLED
Question: How to cover my mouth	Answer: Make sure you, and the people around you, follow good respiratory hygiene. This means covering your mouth and nose with your bent elbow or tissue when you cough or sneeze	ENABLED

Intent Action and Parameters

Specific Intents and actions were defined to streamline the user query and return appropriate answers, as shown in screenshot below.

As shown in screenshot below, different users might use Corona, Covid, Covid-19 interchangeably and chatbot should be able to interpret them in similar fashion




The screenshot shows the Dialogflow console interface for the 'Entities' section of the 'Corona-FAQ' project. The left sidebar is the same as in the previous screenshot, with 'Entities' selected. The main area is titled 'Entities' and has a 'CREATE ENTITY' button at the top right. Below the title are tabs for 'CUSTOM' and 'SYSTEM'. The 'CUSTOM' tab is active, showing a search bar 'Search entities' and a list of two custom entities: '@ corona' and '@ country-names'.




corona








SAVE





☒ Define synonyms  ☐ Regexp entity  ☐ Allow automated expansion ☒ Fuzzy matching 



corona	covid	corona virus	covid-19	coronavirus
	coronaviruses	noble corona virus	Enter synonym	


 Dialogflow



Corona-FAQ  
en hi 



 Intents 
 Entities 
 Knowledge [beta]
 Fulfillment
 Integrations



 Training
 Validation
 History
 Analytics




 Prebuilt Agents
 Small Talk

 Docs

• how i catch corona  SAVE 


Contexts  

Events  





Training phrases  Search training phrases  

” Add user expression

” Can I catch corona from the source

Action and parameters 

Enter action name

REQUIRED 	PARAMETER NAME 	ENTITY 	VALUE	IS LIST 
<input type="checkbox"/>	source	@sys.any	\$source	<input type="checkbox"/>
<input type="checkbox"/>	corona	@corona	\$corona	<input type="checkbox"/>
<input type="checkbox"/>	Enter name	Enter entity	Enter value	<input type="checkbox"/>

Realtime webhook

Webhook based system was connected to locally deployed flask model which would server queries directly from backend in Realtime. This can provide realtime updated information from backend, instead of just relying on knowledge base. Screenshot below with details.

Steps to enable webhook

1. Run a flask server in localhost
2. Use ngrok to connect localhost to publicly available IP
3. Enable webhook for relevant Intents.
4. Enable webhook in fulfilment in Dialogflow

Flask server running on localhost

```
(base) anmol@x86_64-apple-darwin13 CoronaHelpDesk % python api.py
2020-04-03 13:52:56.890963: I tensorflow/core/platform/cpu_feature_guard.cc:142] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2 FMA
2020-04-03 13:52:56.905652: I tensorflow/compiler/xla/service/service.cc:168] XLA service 0x7fd3c9e3af30 initialized for platform Host (this does not guarantee that XLA will be used). Devices:
2020-04-03 13:52:56.905671: I tensorflow/compiler/xla/service/service.cc:176]   StreamExecutor device (0): Host, Default Version
* Tip: There are .env or .flaskenv files present. Do "pip install python-dotenv" to use them.
* Serving Flask app "api" (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: off
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
127.0.0.1 - - [03/Apr/2020 14:02:17] "GET / HTTP/1.1" 200 -
127.0.0.1 - - [03/Apr/2020 14:02:17] "GET /favicon.ico HTTP/1.1" 404 -
127.0.0.1 - - [03/Apr/2020 14:02:44] "POST / HTTP/1.1" 200 -
```

Ngrok connecting Dialogflow Webhook

```
ngrok by @inconshreveable

Session Status      online
Account             (Plan: Free)
Version             2.3.35
Region              United States (us)
Web Interface        http://127.0.0.1:4040
Forwarding           http://6c1c9423.ngrok.io -> http://localhost:5000
Forwarding           https://6c1c9423.ngrok.io -> http://localhost:5000

Connections         ttl    opn    rt1    rt5    p50    p90
                   0      0      0.00   0.00   0.00   0.00
```

Enable webhook fulfilment in Query

The screenshot shows the Dialogflow console interface. On the left is a sidebar with navigation options: Corona-FAQ, Intents, Entities, Knowledge, Fulfillment, Integrations, Training, Validation, History, Analytics, Prebuilt Agents, and Small Talk. The main area displays the configuration for the 'corona-spread-query' intent. The 'Responses' section is expanded, showing a list of 5 text responses for COVID-19 spread. The 'Fulfillment' section at the bottom shows the 'Enable webhook call for this intent' toggle is turned on.

Dialogflow

Corona-FAQ

en hi

Intents

Entities

Knowledge [beta]

Fulfillment

Integrations

Training

Validation

History

Analytics

Prebuilt Agents

Small Talk

corona-spread-query [!]

Responses [?]

DEFAULT TELEGRAM

Text or SSML Response [?]


- can spread from person to person through small droplets from the nose or mouth which are spread when a person with COVID-19 coughs
- People can catch COVID-19 if they breathe in droplets from a person with COVID-19 who coughs out or exhales droplets
- Studies to date suggest that the virus that causes COVID-19 is mainly transmitted through contact with respiratory droplets rather than through the air
- It is possible to catch COVID-19 from someone who has, for example, just a mild cough and does not feel ill
- Enter a text or SSML response variant

ADD RESPONSES

☒ Set this intent as end of conversation [?]

Fulfillment [?]

☒ Enable webhook call for this intent

 Dialogflow

Corona-FAQ

en hi

+


Intents


+


Entities

+


Knowledge [beta]

 Fulfillment


 Integrations

 Fulfillment


Webhook

ENABLED 

Your web service will receive a POST request from Dialogflow in the form of the response to a user query matched by intents with webhook enabled. Be sure that your web service meets all the [webhook requirements](#) specific to the API version enabled in this agent.

URL*	<input type="text" value="https://6c1c9423.ngrok.io"/>	
BASIC AUTH	<input type="text" value="Enter username"/>	<input type="text" value="Enter password"/>
HEADERS	<input type="text" value="Enter key"/>	<input type="text" value="Enter value"/>
<div> Add header</div>		

Chatbot URL - <https://bot.dialogflow.com/7ece3edf-bff2-4527-ac49-c68209b471ee>



Dialogflow

Corona-FAQ

en

hi

+

Intents

+

Entities

+

Knowledge [beta]

Fulfillment

Integrations

Training

Validation

History

Analytics

Prebuilt Agents

Small Talk

> Docs

Standard

Free

Upgrade

travel-query

SAVE

Contexts

Events

Training phrases

Search training phrases

” Add user expression

” should i travel now

” when should i stay at home

Action and parameters

Responses

DEFAULT

TELEGRAM


+


Text or SSML Response

1 Stay home if you feel unwell. If you have a fever, cough and difficulty breathing, seek medical attention

2 avoid traveling to places especially if you are an older person or have diabetes, heart or lung disease.

Screenshot : Web based demo dialogflow

**Corona-FAQ**

POWERED BY  Dialogflow

Hi

Hi! How are you doing? 🔊

I am doing good

Excellent. I'm here to help keep it that way. 🔊

I want to know about Corona virus

Ask something...

**Corona-FAQ**


POWERED BY  Dialogflow


develops difficulty breathing. Older people, and those with underlying medical problems like high blood pressure, heart problems or diabetes, are more likely to develop serious illness. People with fever, cough and difficulty breathing should seek medical attention. 🔊

Why is disease called coronavirus disease 2019, COVID - 2019?

Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the

Ask something...

**Corona-FAQ**

POWERED BY  Dialogflow


How likely am I to catch COVID-19?


The risk of catching COVID-19 from the feces of an infected person appears to be low 🔊

Are antibiotics effective in preventing or treating the COVID-19?

COVID-19 is the infectious disease caused by the most recently discovered coronavirus. This new virus and disease were unknown before the outbreak began

Ask something...

**Corona-FAQ**

POWERED BY  Dialogflow

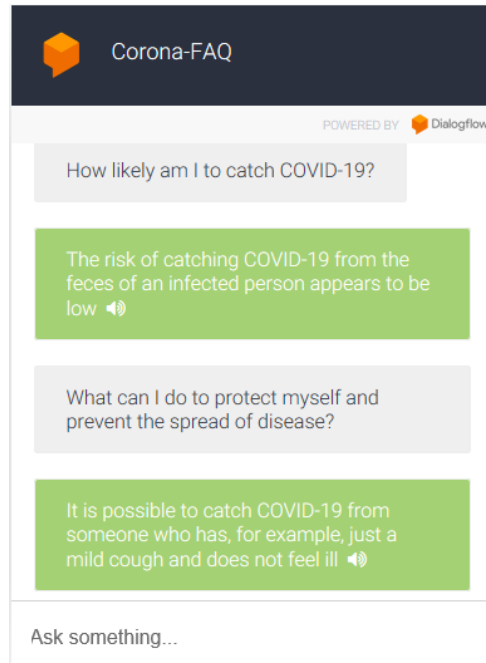
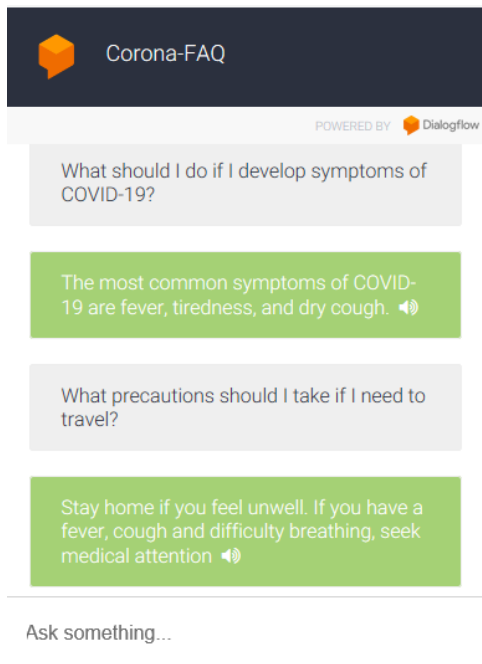
Can I catch covid-19 ?

It is possible to catch COVID-19 from someone who has, for example, just a mild cough and does not feel ill 🔊

How does COVID-19 spread?

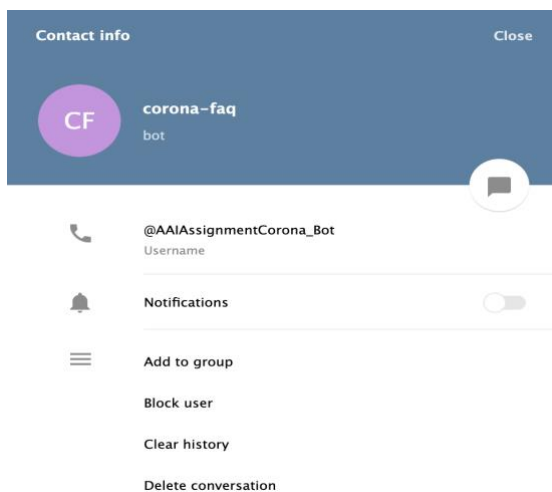
Studies to date suggest that the virus that causes COVID-19 is mainly transmitted through contact with respiratory droplets rather than through the air 🔊

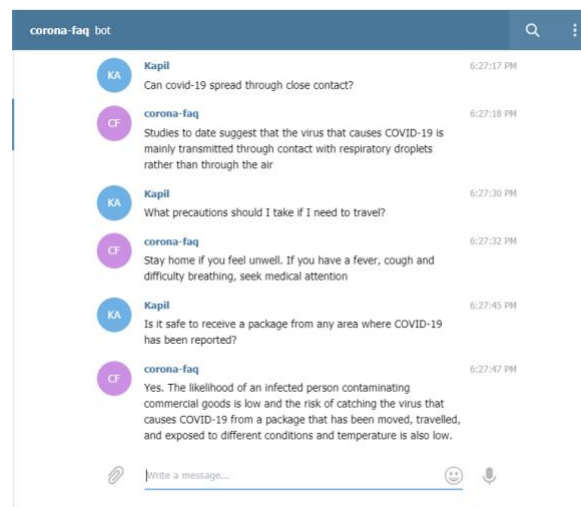
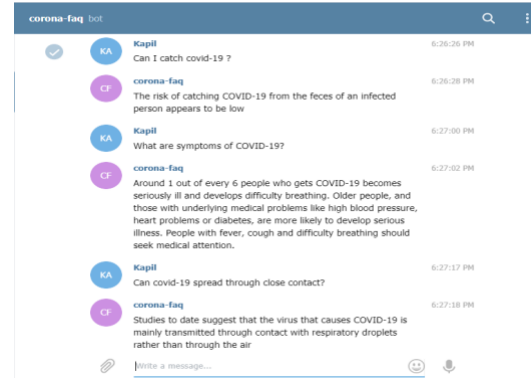
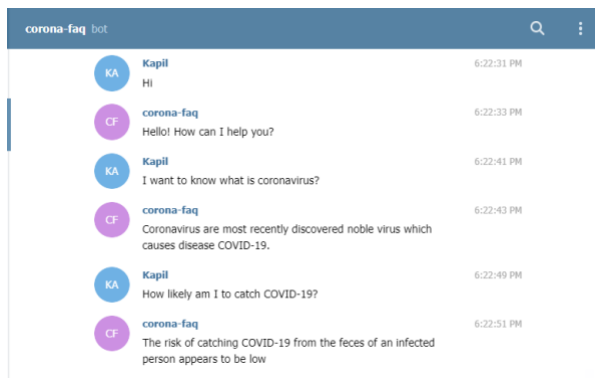
Ask something...



Telegram Bot

We created a telegram bot app and connected it to our dialogflow. It is accessible at @AAIAssignmentCorona_Bot.





IV. Discussion and Ways to Improve

Personas

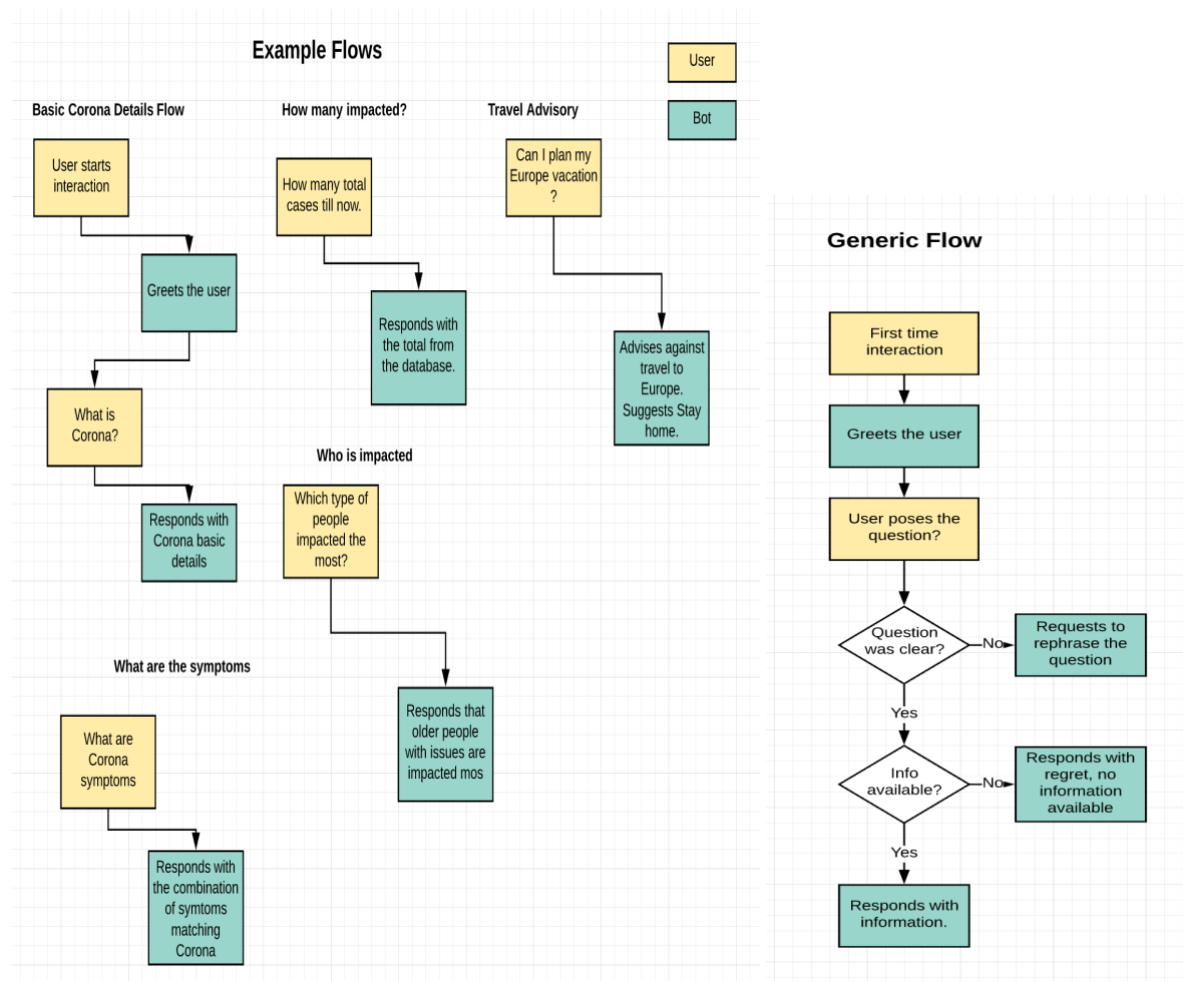
We could see the following personas using our information bot.

- **Potential corona patients:** These are the users who currently have some symptoms and are wondering if theirs match what corona has. Their sentiment could be nervous or concerned and they would like to know what are the complete set of symptoms that uniquely match corona. It is absolutely critical to give them the accurate and reliable information. While symptoms is their primary information sought, details like who to call, when to visit the doctor, tests required and self-isolation approaches would be highly relevant secondary information.
- **Concerned Parents:** Parents are concerned about the kids and they are eager to know how Corona spreads and how to avoid such circumstances. Their sentiment would be curious and proactively concerned. Their primary information sought is spread patterns, hotspots, foods and medications that could compound. Their secondary information sought would be corona symptoms, demographics of people who are impacted the most, how long corona survives on any surface and any precautions on meds and foods to be taken.
- **Curious Kids:** These are curious because they have their virtual classroom assignments that are driving them to do research and come back with some statistics and solutions. Their sentiment will be curious to learn. Their primary information sought would be the statistics like how many

countries are impacted, how many impacted in their state/country, the duration of incubation, how to stay safe. Their secondary information sought would be the travel moratorium details.

Ideal Conversation Flowchart

An ideal flow chart has been shown below, which greets the user and answers relevant queries



Ways to improve your bot

Improve user experience

- Create intuitive UI using jquery and other web development tools.
- Add follow-up intents to keep the conversation open, and answer a follow up query – Right now every query is considered independently
- Add different integrations with WhatsApp, Facebook Messenger or even a play store app.

Improve bot knowledge base

- Increase knowledge base to connect to Realtime data on counts, affected people, government efforts etc

Multiple model deployment

- We have focussed on similarity based models like Question similarity or bi directional similarity models like BERT, but we can try other models like RNN based deep learning models
- Also, tf-idf and similarity based simpler model could have been improved further.

Realtime Integration

- Cloud Deployable Model – Flask server can be deployed on cloud, so it can server 24x7, instead of local deployment

References

- <https://medium.com/analytics-vidhya/adapting-bert-question-answering-for-the-medical-domain-2085ada8ceb1>
- <https://medium.com/@aainabajaj39/domain-specific-excel-dataset-to-squadv1-1-dataset-format-using-python-2b99f827b8a2>
- <https://towardsdatascience.com/lets-build-an-intelligent-chatbot-7ea7f215ada6>
- <https://medium.com/datadriveninvestor/build-your-first-chatbot-in-10-lines-of-code-88c4f15e39c9>
- <https://towardsdatascience.com/nlp-building-a-question-answering-model-ed0529a68c54>
- <https://towardsdatascience.com/modern-question-answering-systems-explained-4d0913744097>
- <https://towardsdatascience.com/identifying-duplicate-questions-on-quora-top-12-on-kaggle-4c1cf93f1c30>