University of North Carolina at Charlotte
ITCS 5111 Intro to Natural Language Processing

QA Health Assistant for COVID 19

Department of Computing & Informatics
Final Project Presentation
Fall 2020

Instructor,
Dr. Samira Shaikh

Project Group - 22

- Muthu Priya Shanmugakani Velsamy
- Mohammed Hussain Musthaq Syed Nizam Babu

Research Question

- The Research topic for our project is Natural Language Generation.
- Build a Question Answering system which answers the general public questions regarding the COVID pandemic with better accuracy.
- The answers to the questions are retrieved from the collection of research papers related to COVID.

Data Source & Description

- Competition for Epidemic Question Answering (EPIC-QA) by the Text Analysis Conference (TAC) for creating the chatbot/QA model with best accuracy results.
- Link to the competition: https://bionlp.nlm.nih.gov/epic_qa/
- The CORD data for training the Question Answering model is provided in kaggle.
- Link to the dataset: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Prior Work

We reviewed implementations of various QA models from the following sites:

- Kaggle: BERT & BART QA
- Github: BERT SQuAD QA
- Medium: QA using BERT Pipeline
- Morioh: cdQA pipeline model
- Towardsdatascience: QA model in python
- Kaggle: BERT QA

Prior Work

For Text Summarization we reviewed the following resources:

- Medium: Text summarization on COVID data
- Towardsdatascience: Summarization using BART model
- Python: BERT Summarizer
- GeekforGeeks: Text Summarization in python
- KdNuggets: Automated Text Summarization
- GeekforGeeks: Extractive Text Summarization using Gensim

Our Approach

- Imported all the json files.
- Pre processed all the json files and extracted a clean dataframe.
- Implemented BERT (Bidirectional Encoder Representations from Transformers) model to build/train our QA system.
- Implemented Summarization model
- We used Exact match (accuracy) for testing the accuracy of the model.
- Implemented BLEU score and ROUGE score for the QA model evaluation.

Output / Result


```
{'paper_id': '000ed27575c028d3173a3fd59be053446454f985', 'metadata': {'title': 'COVID-1
9 and its Modes of Transmission', 'authors': [{'first': 'Rutu', 'middle': [], 'last':
'Karia', 'suffix': '', 'affiliation': {}, 'email': ''}, {'first': 'Ishita', 'middle':
[], 'last': 'Gupta', 'suffix': '', 'affiliation': {}, 'email': ''}, {'first': 'Harshwar
dhan', 'middle': [], 'last': 'Khandait', 'suffix': '', 'affiliation': {}, 'email': ''},
{'first': 'Ashima', 'middle': [], 'last': 'Yadav', 'suffix': '', 'affiliation': {}, 'em
ail': ''}, {'first': 'Anmol', 'middle': [], 'last': 'Yadav', 'suffix': '', 'affiliatio
n': {}, 'email': ''}]}, 'abstract': [{'text': 'The World Health Organization recognized
SARS-CoV-2 as a public health concern and declared it as a pandemic on March 11, 2020.
Over 12 million people have been affected across several countries since it was first r
ecognized. SARS-CoV-2 is thought to commonly spread via respiratory droplets formed whi
le talking, coughing, and sneezing of an infected patient. As several cases, with an ab
sence of travel history to the majorly affected areas were identified, a strong possibi
lity of community transmission could have been possible. Broadly, two modes of transmis
sion of COVID-19 exist-direct and indirect. The direct mode includes (1) transmission v
ia aerosols formed via surgical and dental procedures and/or in the form of respiratory
droplet nuclei; (2) other body fluids and secretions, for example, feces, saliva, urin
e, semen, and tears; and (3) mother-to-child. Indirect transmission may occur via (1) f
omites or surfaces (e.g., furniture and fixtures) present within the immediate environm
ent of an infected patient and (2) objects used on the infected person (e.g., stethosco
pe or thermometer). As many of these modes may be underestimated, it is necessary to em
phasize and illustrate them. The goal of this paper is to briefly review how SARS-CoV-2
```


Json file format of our data

Output / Result

	paper_id	abstract	body_text
0	a0d063dca746b135afe0451ce0b3bb1e06cf15ae	Background Brazil ranks second worldwide in to...	The COVID-19 pandemic has created an unprecede...
1	edb294108440787c9f074483fd3c953a83e53622		Die Corona-Pandemie ist eine Gefahr für die Ge...
2	a0bc6bc5b8547b98a2d77b81ca81cb18fa1b7ee9		To the editor, We read with great interest the...
3	6b9d9eb2e9f448a5d2b3646b37b16534211cb3ff	Coronavirus disease 2019 is a global pandemic ...	Forces beyond your control can take away every...
4	961458c62b1ac196cf312994ff02e5edbd6a1c6a		civil and state cohesion, prosperity and power...
...	...	...	...
995	d50f90c6b6d9441382b9d9032c1ded1fc12ca196	Stochasticity and spatial heterogeneity are of...	When the epidemiologists at a public health ag...
996	d18a705998ad871dad46aeabeeed0a20909c10df	Respiratory diseases are a major contributor t...	Establishment of a surveillance strategy in La...
997	9ff0fbcfa1e606dbd692b91c59f76e7f183958c2	The coronavirus disease (COVID-19) pandemic is...	The disease known as coronavirus caused by SAR...
998	313d6762ff0c7e18ed7af39482b04fbd2d280bc7	Glycosylation is a ubiquitous post-translation...	The increasing emergence of infectious disease...
999	0903dd0da2be2a7b492da5e2eba573c7f44fb23f	This essay discusses hope and optimism with re...	Much of current rhetoric in response to the gl...

1000 rows × 3 columns

Dataframe extracted from the json files (1000 files were used to reduce processing time and computational resources)

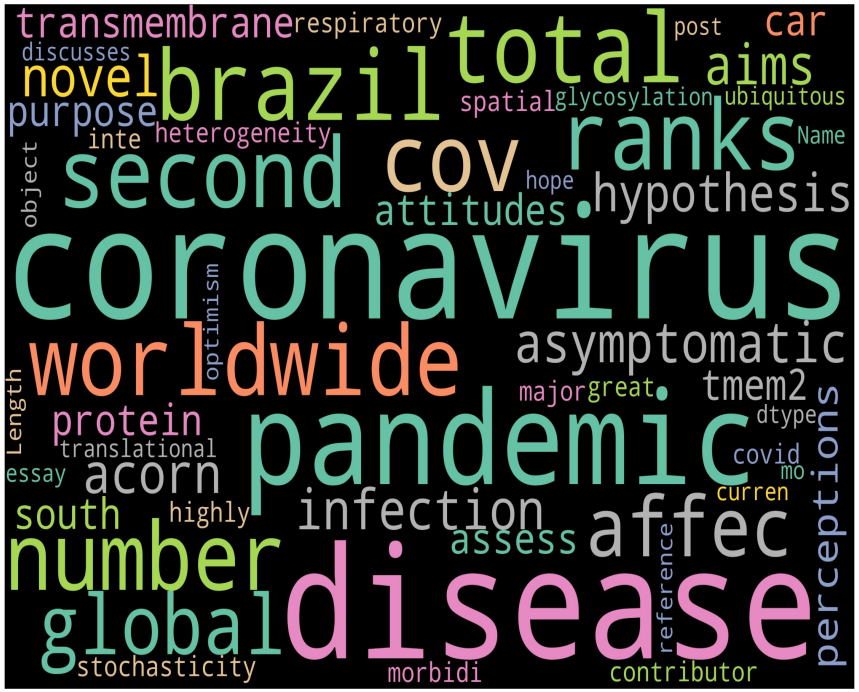
Output / Result

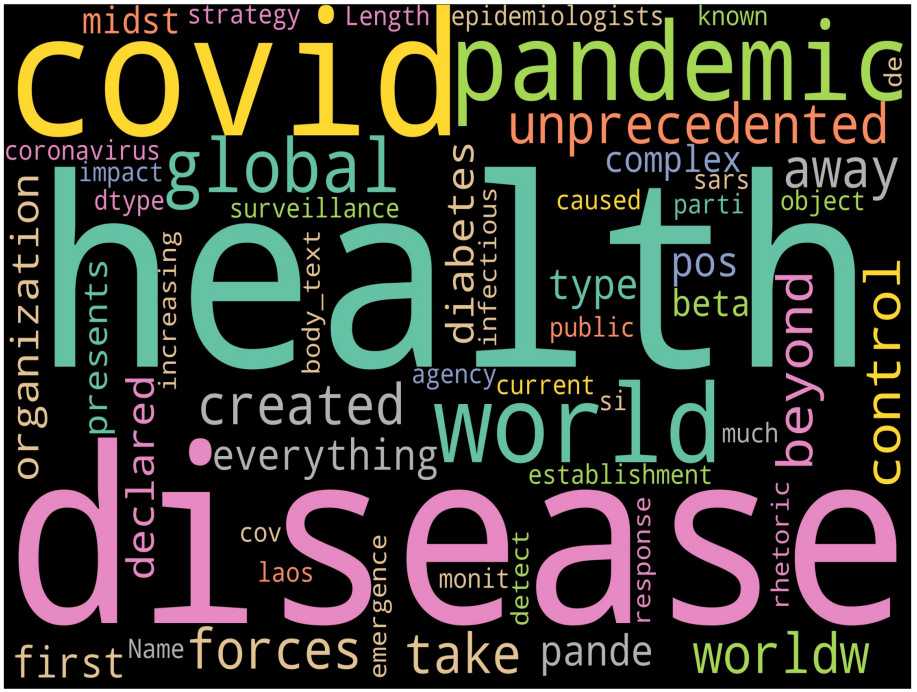
	paper_id	abstract	body_text	abstract_word_count	body_word_count	body_unique_words
0	a0d063dca746b135afe0451ce0b3bb1e06cf15ae	brazil ranks second worldwide total number cov...	covid 19 pandemic created unprecedented worldw...	372	4086	1298
1	6b9d9eb2e9f448a5d2b3646b37b16534211cb3ff	coronavirus disease 2019 global pandemic affec...	forces beyond control take away everything pos...	208	2841	1110
2	c79ce955bfc71ffe8159bca6bc81d783a86d8edf	asymptomatic novel coronavirus infection acorn...	world health organization first declared pande...	261	1089	509
3	be9bdbb4987a83ad38fb0b65018528055e13eab7	aims hypothesis transmembrane protein 27 tmem2...	type 2 diabetes complex disease presents beta ...	243	4445	1531
4	acfb6e59bf5f762b7a749c4bdc3613360fdc2160	purpose assess perceptions attitudes south car...	midst global covid 19 pandemic world health or...	225	7146	2382
...	...	...	...	...	...	...
702	d50f90c6b6d9441382b9d9032c1ded1fc12ca196	stochasticity spatial heterogeneity great inte...	epidemiologists public health agency detect si...	115	5837	1493
703	d18a705998ad871dad46aeabeeed0a20909c10df	respiratory diseases major contributor morbidi...	establishment surveillance strategy laos monit...	272	4935	1725
704	9ff0fbcfa1e606dbd692b91c59f76e7f183958c2	coronavirus disease covid 19 pandemic highly i...	disease known coronavirus caused sars cov 2 de...	359	2679	941
705	313d6762ff0c7e18ed7af39482b04fbd2d280bc7	glycosylation ubiquitous post translational mo...	increasing emergence infectious diseases parti...	124	9163	2603
706	0903dd0da2be2a7b492da5e2eba573c7f44fb23f	essay discusses hope optimism reference curren...	much current rhetoric response global impact c...	36	1872	817

707 rows × 6 columns

Final Dataframe after all the preprocessing is done

Output / Result





Most common words in the dataframe abstract and main text columns

Output / Result

```
query_sample = "How to prevent Corona ?"
relevant_sentence = df['abstract'].values
nlp(question = query_sample, context = relevant_sentence)
```

```
/opt/conda/lib/python3.7/site-packages/transformers/tokenization_utils_base.py:1374: FutureWarning: The `max_len` attribute has been deprecated and will be r
emoved in a future version, use `model_max_length` instead.
  FutureWarning,
```

```
{'score': 0.03531736135482788,
 'start': 70,
 'end': 131,
 'answer': 'Understanding possible socioeconomic ethnic health inequities'}
```

```
query_sample = "What is the Incubation period for COVID 19"
relevant_sentence = df['body_text']
predicted_answer = nlp(question = query_sample, context = relevant_sentence)
nlp(question = query_sample, context = relevant_sentence)
```

```
/opt/conda/lib/python3.7/site-packages/transformers/tokenization_utils_base.py:1374: FutureWarning: The `max_len` attribute has been deprecated and will be r
emoved in a future version, use `model_max_length` instead.
  FutureWarning,
/opt/conda/lib/python3.7/site-packages/transformers/tokenization_utils_base.py:1374: FutureWarning: The `max_len` attribute has been deprecated and will be r
emoved in a future version, use `model_max_length` instead.
  FutureWarning,
```

```
{'score': 0.2350313514471054,
 'start': 2659,
 'end': 2672,
 'answer': 'feb 27 4 2020'}
```

Output / Result

```
ques = ["What is COVID19"]
ans_pd,Summary_text_3 = ANS_Model(ques,len(id2abstract[:1000]))
Summary_text_3
```

```
Token indices sequence length is longer than the specified maximum sequence length for this model (535 > 512). Running this sequence through the model will result in indexing errors
Truncation was not explicitely activated but `max_length` is provided a specific value, please use `truncation=True` to explicitely truncate examples to max length. Defaulting to 'longest_first' truncation strategy. If you encode pai
rs of sequences (GLUE-style) with the tokenizer you can select this strategy more precisely by providing a specific strategy to `truncation`.
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-49-2679fe6438fc> in <module>
      1 ques = ["What is COVID19"]
----> 2 ans_pd,Summary_text_3 = ANS_Model(ques,len(id2abstract[:1000]))
      3 Summary_text_3

<ipython-input-47-f6e1e01b1898> in ANS_Model(ques, count)
     74     ans_pd['abstract_by_ans'] = extrac_list
     75     ans_pd = ans_pd.sort_values(by=['Confident'],ascending=False).reset_index(drop=True)
---> 76     Summary_text = Summary_Model(ans_pd,100,model)
     77     return ans_pd,Summary_text

<ipython-input-46-4533c594a6cc> in Summary_Model(pd, count, model)
     18             # Encode multiple sentences using tokenizer.batch_encode_plus
     19             # tokenizer.batch_encode_plus will generate a dictionary which contains the input_ids, token_type_ids and the attention_mask as list for each input sentence
---> 20             one_token = tokenizer.batch_encode_plus([tokens_a], max_length = 1024, return_tensors = 'pt')  # return_tensors = 'pt' If set, will return pyTorch objects instead of list of python integers.
     21             all_tokens.append(one_token)
     22

/opt/conda/lib/python3.7/site-packages/transformers/tokenization_utils_base.py in batch_encode_plus(self, batch_text_or_text_pairs, add_special_tokens, padding, truncation, max_length, stride, is_split_into_words, pad_to_multiple_of,
return_tensors, return_token_type_ids, return_attention_mask, return_overflowing_tokens, return_special_tokens_mask, return_offsets_mapping, return_length, verbose, **kwargs)
   2321             return_length=return_length,
   2322             verbose=verbose,
-> 2323             **kwargs,
   2324         )
   2325

/opt/conda/lib/python3.7/site-packages/transformers/tokenization_utils.py in _batch_encode_plus(self, batch_text_or_text_pairs, add_special_tokens, padding_strategy, truncation_strategy, max_length, stride, is_split_into_words, pad_t
o_multiple_of, return_tensors, return_token_type_ids, return_attention_mask, return_overflowing_tokens, return_special_tokens_mask, return_offsets_mapping, return_length, verbose, **kwargs)
    558                 ids, pair_ids = ids_or_pair_ids, None
    559             else:
--> 560                 ids, pair_ids = ids_or_pair_ids
    561
    562             first_ids = get_input_ids(ids)

ValueError: too many values to unpack (expected 2)
```

BERT QA & BART Summarization (Method-2)

Output / Result

```
ValueError: too many values to unpack (expected 2)
```

Search Stack Overflow

+ Code + Markdown

```
ques = ["What is range of incubation period for coronavirus SARS-CoV-2 COVID-19 in humans"]
ans_pd,Summary_text_4 = ANS_Model(ques,len(id2abstract[:100]))
Summary_text_4
```

```
"Can't find the answer. Try re-phrasing your question."
```

```
ques = ["What is known about transmission, incubation, and environmental stability for the 2019-nCoV","What are the case fatali
ans_pd,Summary_text_5 = ANS_Model(ques,len(id2abstract[:250]))
Summary_text_5
```

```
"Can't find the answer. Try re-phrasing your question."
```

BERT QA & BART Summarization (Method-2)

Output / Result

```
ques = ["What is SARS?"]
ans_pd,Summary_text_3 = ANS_Model(ques,len(id2abstract[:50]))
Summary_text_3
```

```
Your max_length is set to 142, but you input_length is only 103. You might consider decreasing max_length manually, e.g. summarizer
('...', max_length=50)
```

```
| seasonal influenza  vaccines lack efficacy against drifted or pandemic influenza strains.. coronaviruses (covs) are by far the larges
t group of known positive-sense rna viruses having an extensive range of natural hosts. in the past few decades, newly evolved coronavi
ruses have posed a global threat to public health.. bmj open publishes all reviews undertaken for accepted manuscripts..we report  a la
boratory-confirmed case of severe acute respiratory syndrome (sars) in a pregnant woman..
```

```
[{'summary_text': ' seasonal influenza  vaccines lack efficacy against drifted or pandemic influenza strains . coronaviruses (covs) are
by far the largest group of known positive-sense rna viruses having an extensive range of natural hosts . in the past few decades, newl
y evolved coronaviraluses have posed a global threat to public health .'}]
```

BERT QA & BERT Pipeline Summarization (Method-3)

Output / Result

[77]:
```
compare_bleu(expected_Summary_text_3, sum_3)
```

```
The BLEU score of accuracy is:  0.5542089483371553
```

```
/opt/conda/lib/python3.7/site-packages/nltk/translate/bleu_score.py:490: UserWarning:
Corpus/Sentence contains 0 counts of 2-gram overlaps.
BLEU scores might be undesirable; use SmoothingFunction().
  warnings.warn(_msg)
```

```
from rouge_score import rouge_scorer

scorer = rouge_scorer.RougeScorer(['rouge1', 'rougeL'], use_stemmer=True)
scores = scorer.score(expected_Summary_text_3, sum_3)
scores
```

0]
```
{'rouge1': Score(precision=0.16666666666666666, recall=0.08791208791208792, fmeasure=0.11510791366906475),
 'rougeL': Score(precision=0.0625, recall=0.03296703296703297, fmeasure=0.04316546762589928)}
```

BERT QA & BERT Pipeline Summarization (Method-3)

Output / Result

```
     31
     32             if number_ids < 512:
---> 33                 start_scores, end_scores = model_new(torch.tensor
     34             else:
     35                 start_scores, end_scores = model_new(torch.tensor
2]]).to(device))

/opt/conda/lib/python3.7/site-packages/torch/nn/modules/module.py in _cal
    725             result = self._slow_forward(*input, **kwargs)
    726         else:
--> 727             result = self.forward(*input, **kwargs)
    728         for hook in itertools.chain(
    729                 _global_forward_hooks.values(),

TypeError: forward() takes 2 positional arguments but 3 were given

Search Stack Overflow
```

BERT with Linear model & covid_bert_base QA model (Method-4)

Output / Result

```
        )
        (decoder): Linear(in_features=768, out_features=30522, bias=True)
      )
    )
  )
  (output_linear): Linear(in_features=30522, out_features=2, bias=True)
```

BERT with Linear model & covid_bert_base QA model (Method-4)

Team members Contributions

- *Muthu Priya*
  - Review prior works online (kaggle, stackoverflow, github)
  - Choose/Collect Dataset
  - EDA - Contraction words, NULL values, Word Cloud
  - Model - BERT QA Pipeline, Summarization model ; Tried cdQA model
  - Test - ROUGE score
  - Project Report
  - Presentation slides
- *Mohammed Hussain*
  - Review prior works online (kaggle, stackoverflow, github)
  - Import Data
  - EDA - Stop words, Lower case words, punctuations
  - Model - BERT QA model, Summarization model ; Tried SQuAD BERT
  - Test - BLEU score
  - Project Report
  - Presentation slides

Acknowledgement

- We sincerely thank our Professor *Dr. Samira Shaikh* and our TA *Erfan Al-Hossami* for their continued support & guidance throughout the course project.
- We also thank all the authors of the google content & solution providers, for their work which we have referred in our course project.

Thank You!

By,
Group - 22