INFORMATION RETRIEVAL USING TENSORFLOW AND ELASTIC SEARCH

INFORMATION RETRIEVAL ON MEDICAL QUESTION AND ANSWER DATA SET

VIVEK KASHYAP

Table of Contents

Context

This project is intended to build a question answering model where in user can ask a specific question and model will run through the data repository and come back with relevant answers. This model is not built with intent of having a conversational logic where there will be back and forth and follow up questions coming from user based on the previous responses.

The model is not tuned and perfected to state of art accuracy, but it is built to apply and demonstrate the concepts which have been learnt as part of Deep Learning.

For this project I had taken a set of questions and answers from healthcare domain. Model will use this question and answer data set as a base to retrieve and find closest match for a user query. Though this model was developed using medical data set this model can be easily generalized to other domain if there are predefined question and answers set available in that domain.

The next few sections have details about how the model was developed from collecting the data, building the algorithm and outcome seen. Towards the end I have also outlined other technique which can be applied in future to increase the accuracy of this model.

Data Collection

Data used for building the model was taken from publicly available dataset: medical-question-answer-data

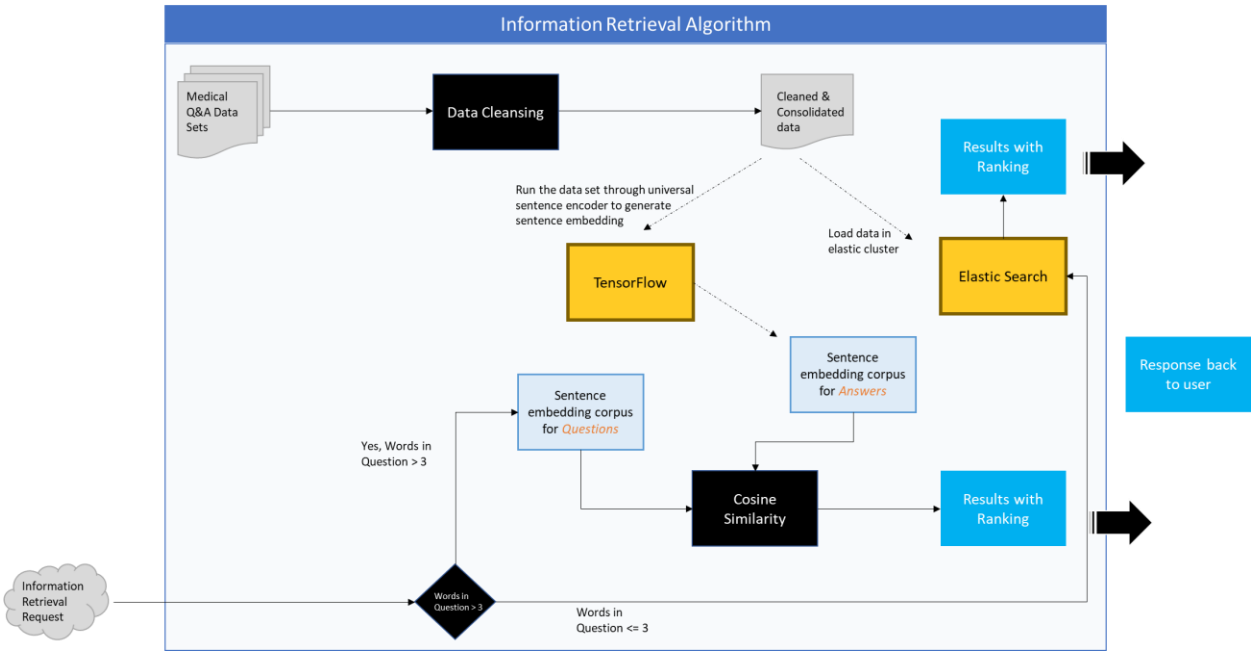
Source: https://github.com/LasseRegin/medical-question-answer-data

The source consists of data collected from

- eHealth Forum
- HealthTap
- iCliniq
- Question Doctors
- WebMD

Information Retrieval Model: Approach & Algorithm

Model Approach & Design:



Key components of the model are:

- **Data cleansing and consolidation**:

  Based data on which questions or queries would be executed can be in one single file or source or it can be in multiple files. For ease of use we cleaned and consolidated the data in one single location.

  Below are data samples across multiple files in the source. The structure of data is similar but there are some differences.

Source 1:


```
"answer": "the normal level of vitamin b12 in the blood is 200-1000ng/l. the levels of methylmalonic acid are generally high during anemia due to a lack of vitamin b12 or folic acid. methylmalonic acid can also be checked with a urine test. most hmos cover diagnostic services when medically necessary and prescribed by a participating physician. check with your insurance provider for more detailed information about your coverage.",
"question": "following my discution with afriend who is doctor out of california he told me i may suffer from b12 deficiency. latley my hmos doctor send ne to b12 test that seems fine(200). my friend strongly sugested that i try to get the mm test wich may be the only test that relevant for b12 deficiency. does someone knows if the hmos are covering this test ingeneral? thanks yarden",
"url": "http://ehealthforum.com/health/topic76835.html",
```


Source 2:


```
"answer": "hi well only possibility that you did not explore is having multiple sclerosis so you will need a spinal tab did you have any muscle weakness? is the squint in one eye? is the headache in one side? can you upload the brain mri? \u2026.",
"answer_author": "Dr Ahmed Fawzy",
"question": "i had a perfect vision until i suddenly developed a double vision when i was on a europe trip doctors cannot diagnose what\u2019s wrong with me?",
"question_text": "I had a perfect vision until I suddenly developed a double vision when I was on a Europe trip, doctors cannot diagnose what\u2019s wrong with me?",
```


Source 3:


```
"question": "i got a minor cut from a old rusty knife i am up to date on my tetanus shot but should i still go get it checked?",
"short_answer": "\nCut\n",
"answer": "signs and symptoms to look out for would be worsening redness swelling pain in the area of cut. also if you develop a fever that would suggest infection. should you see any of these recommend seeing a physician.",
```


Consolidation:

All the data was present in ***json*** file format. Data was consolidated in single file with structure

```
{
    "answer": "answer ......",
    "short_answer": "short answer ......",
    "question": "question.....",
    "question_text": "question text ......",
```

```
for file in os.listdir( path ):
    if file.endswith( ".json" ):
        json_array = json.load( open( path + file ) )
        for item in json_array:
            if ('answer' not in item):
                item['answer'] = 'None'
            if ('short_answer' not in item):
                item['short_answer'] = 'None'
            if ('question' not in item):
                item['question'] = 'None'
            if ('question_text' not in item):
                item['question_text'] = 'None'
```

answer and short_answer text is combined together and send to cleansing function

Cleansing:

After consolidation data was cleansed by building a cleansing function (shown below)

```
def clean_sent(s):
    s_ = s.lower()
    table = str.maketrans('', '', string.punctuation)
    #s_= [w.translate(table) for w in s_]
    s_ = re.sub(r"i'm", "i am", s_)
    s_ = re.sub(r"it's", "it is", s_)
    s_ = re.sub(r"he's", "he is", s_)
    s_ = re.sub(r"she's", "she is", s_)
    s_ = re.sub(r"that's", "that is", s_)
    s_ = re.sub(r"what's", "what is", s_)
    s_ = re.sub(r"where's", "where is", s_)
    s_ = re.sub(r"\'ll", " will", s_)
    s_ = re.sub(r"\'ve", " have", s_)
    s_ = re.sub(r"\'re", " are", s_)
    s_ = re.sub(r"\'d", " would", s_)
    s_ = re.sub(r"won't", "will not", s_)
    s_ = re.sub(r"can't", "cannot", s_)
    s_ = re.sub(r"[-()\"#/@;:<>{}+=~|.?,]", " ", s_)
    s_ = re.sub(r'[^\x00-\x7F]+', ' ', s_)
    s_ = re.sub(r'[(\n+)+]', ' ', s_)
    s_.encode('ascii', errors='ignore').strip().decode('ascii')
    return s_
```

- **Accepting the user question**:

Accept the user question. If number of words is more than three then retrieve information based on Universal Sentence Encoder otherwise retrieve based on elasticsearch

```
parser=argparse.ArgumentParser()
parser.add_argument("-q",help="this is my question?")
args= parser.parse_args()
str = args.q
str = str.replace('?','')
str = str.replace('?','')
if (len(str.split()) > 3):
    # Retrieve based on Universal Sentence Encoder
else:
    # Retrieve based on ElasticSearch
```

- **Find Sentence similarity using Tensor flow based Universal Sentence Encoder**:

Question and Sentence embedding is generated using tensorflow_hub module

```
module_url="https://tfhub.dev/google/universal-sentence-encoder/4")
model = hub.load( module_url )
def embed(self,input):
    return self.model(input)
```

Embedding for the sentences present in answer data set

Similarity between question and answer is calculated using cosine similarity

```
cosine_similarities = tf.reduce_sum(tf.multiply(sts_encode1, sts_encode2), axis=1)
clip_cosine_similarities = tf.clip_by_value(cosine_similarities, -1.0, 1.0)
scores = 1.0 - tf.acos(clip_cosine_similarities) / math.pi
scores_list = scores.numpy().tolist()
top_matches = sorted( range(len(scores_list) ), key=lambda i: scores_list[i], reverse=True )[:match]
```

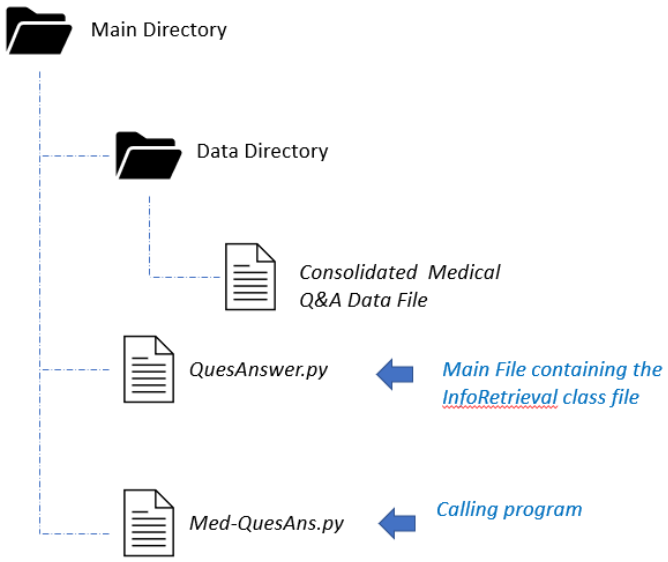
- **Elastic Search**:

```
res= elastic.search(index='cb',doc_type='medicalqa',body={
        'query':{
            'match':{
                "answer":"What are symptoms of Parkinson?"
            }
        }
    })
```

- **Modularizing and dockerizing the code base**:

  Code base is available in github location: https://github.com/vikashya/Information-Retrieval

  Structure of project will be



  Code is also available in docker image (docker link to be published)

Output & Result:

Here are some of the samples for Information Retrieval request tried on the model with responses received. As this model was designed to retrieve information from either Universal sentence encoder based search or Elastic search depending on length of the question, I have summarized the results also in those categories.

Find Sentence similarity using Tensor flow based Universal Sentence Encoder:

#	Question	Top Three Responses
1	i have been taking methadose and oxycotin illeagely for over a year now. i am now trying to quit. the withdrawls are the worst thing i have ever experienced. so i keep going back to get more. i have no insurance and i dont have the money to go to the e. r. is there any way i can ease the pain so i can get my life back in order. please help i dont know what to do.	1. **Score:0.728** : 'none it takes me to get my withdrawls if i run out a couple days and then i throw up 24 7 dirarreaha severe cannot eat or hold down water for me it is worse than anything i happen to know because i keep running out early as i have built up high tolerance to them for my medical issue and ii have two torn discs and need my meds but have ran out too many times for over a week at a time i cannot help you but you either go cold turkey like i have to or get help from a dr for suboxone sl strips they are wonderful for withdrawls but they too have worse withdrawl tendency than opiates i am willing to take anything to not have my bout with withdrawls good luck 2. **Score:0.6945**: "none my best advice to speak to your healthcare provider the one that prescribes the strips to you he or she can help you transition off the suboxone if on the other hand you mean you obtain the strips illicitly and now want to stop using them then i advise you seek addiction counseling sure you can taper off the suboxone dose on your own successfully but you need to address the underlying reasons why you began using suboxone to begin with any addiction involves 'environmental' and emotional factors as well as the actual physical addiction suboxone addiction is becoming a real problem if this is the situation with you i am sorry you have found yourself in this predicament but there is help available i commend you for getting off this substance there is no shame in seeking help wishing you well! the best way the only one that works is to gradually reduce the dose around 25% every 10 days never make a sudden stop or you

		will feel all the terrible effect of withdrawal! you can read in details the taper technique here link good luck and keep in mind even if you feel well continue to reduce the dose gradually takes you time" 3. **Score:0.688**: in a way people can develop what is very much like an addiction to certain foods and carbs are notorious you get through it like any other addict you decide it is time to quit you get rid of it all you tough out the withdrawal symptoms even though it stinks and once you are sober you don't go back
2	the stool color is black (not loose) - is the black color of stool due to any of the above medicines? being holiday can not consult doctor instantly	1. **Score:0.729:** ' see dr many possible causes of different stool color really dark is too vague bleeding from stomach or duodenum may produce stool as black as tar with a peculiar odor a health pro can easily check your stool for blood most common reason for different stool colors is different foods ' 2. **Score:0.726:** 'none dehydration definitely causes a change in stool consistency but not usually color water is actually stored in the colon not the bladder and can be used by the body when needed during times of dehydration water is pulled from the colon and will result in dry hard stools constipation stool color changes can be due to many things for instance if you took pepto bismol your stool would turn black from the bismuth the types of foods that you may eat may also change stool color knowing the color of the stool can give a clinician a clue as to the reason ' 3. **Score:0.724:** " stool color stool color is not important unless there is no color at all or it is bloody melena which is black sticky foul smelly stool signifies blood as well and is important otherwise green brown orange or shades thereof don't mean anything "
3	is it best to start on sinemet or hold off	1. **Score:0.643:** ' yes yes you should unless you want to get another bypass sooner rather than later ' 2. **Score:0.639:** ' should not be an issue ' 3. **Score:0.637:** ' what is the question with symptoms like this i think it is better to see u pghysician '
4	what is the proper way to apply sunscreen to the face?	1. **Score:0.708;** 'none applying sunscreen is a very important part of your daily skin care regimen especially if you are exposed to prolonged sunlight choosing your spf level should be based on how long you plan to be outdoors you may find that you normally turn red after being exposed to the sun for twenty minutes you will then mulitply that number with the spf factor to understand how long you will potentially be protected from the sun for example 20 minutes redness appears x 15 sunscreen with spf 15 protection from the sun for 300 minutes you will need to reapply the sunscreen after 300 minutes expire…….. 2. **Score:0.706:** 'none always use a broad spectrum sunscreen which blocks uva and uvb rays it should be an spf of at least 30 i am a big fan of physical sunscreens that contain zinc oxide and titanium dioxide put your sunscreen on last apply any anti aging products before you put on your sunscreen pick the appropriate sunscreen for your skin type if you are oily choose a serum and if you have dry skin choose one which contains moisturizing ingredients ' 3. **Score:0.705:** none well it sounds like you are on the right track in your line of work you certainly want to avoid too much sun exposure which can lead to skin cancer a couple of suggestions first remember that sunscreen needs to be applied at least 30 minutes before you go into the sun in order for it to be effective also read the label on your sunscreen and make sure you reapply it at the suggested intervals many sunscreens have to be re applied every 2 hours if you are getting sweaty you will need to reapply it even more often than that also you might consider looking for a sunscreen that contains zinc oxide which is a total sun blocker….

Elastic search:

#	Question	Top Three Responses
1	methadose illegaly	1. **Score: 12.20103:** noodles can be wheat rice mung bean soy etc. wheat based noodles may be treated with lye to alter the texture or borax (illegaly) to whiten and improve texture as well. rice based noodles typically have only starch or rice flour and water as ingredients.
2	black stool	1. **Score: 17.745642:** medicines that contain iron like tonoferon usually causes black stool. 2. **Score: 16.509777:** eating red meat usually won't cause black stools. but it's possible that eating large amounts of very rare meat might contain enough blood to cause black stools. as blood travels through the gi tract it turns stool black(melena). 3. **Score: 16.214962:** the most serious cause of black stools is bleeding from somewhere in the gi tract. bleeding in the stomach esophagus or intestines can turn into a black color by the time it leaves the body as stool. this can be caused by bleeding ulcers stomach inflammation broken blood vessels in the esophagus called varices or tumors. this type of bleeding can be life-threatening. the bleeding can also come from a nosebleed or from the mouth if blood is swallowed. black stools can also be caused by supplements such as iron or medicines such as pepto-bismol……..
3	start sinemet	1. **Score: 18.057598:** sinemet is a combination of levodopa and carbidopa. this medicine is used for treating parkinson disease. the choice of medicine(s) and dosage is strongly individual. you can consult your neurologist on whether or not to start sinemet or wait and hold off. 2. **Score: 16.005308:** sinemet cr is a version of sinemet that works for a longer period of time as it releases the drug more slowly into the body. 3. **Score: 13.701747:** see a sleep specialist consider drugs such as requip mirapex neupro (rotigotine) patch or sinemet or clonazapam.
4	apply sunscreen	1. **Score: 15.72783:** 1. avoid peak daylight hours 2. use sunscreen & re-apply often! 3. wear a cover-up over swim suit 4. stay under the umbrella.

		2. **Score: 15.638908:** perhaps you are allergic to one or more of the ingredients in the sunscreen. if there are any acids in the sunscreen this may burn or sting temporarily. if there is a break in your skin sunscreen may sting. my sunscreen has glycolic acid an alpha-hydroxy acid and it stings for a few seconds when i apply it after i shave. the combination of having shaved mixed with the acid is what stings. 3. **Score: 15.375744:** always use a broad- spectrum sunscreen which blocks uva and uvb rays it should be an spf of at least 30. i'm a big fan of physical sunscreens that contain zinc oxide and titanium dioxide. put your sunscreen on last apply any anti-aging products before you put on your sunscreen. pick the appropriate sunscreen for your skin type if you're oily choose a serum and if you have dry skin choose one which contains moisturizing ingredients.

For Universal sentence encoder based search it's seen most of times where it comes back with relevant search but there are times where it does not (Ex: Sample Question #3) . There is potential of tuning algorithm and get more relevant output which I have explained at high level in next section.

Potential Improvement and Fine Tuning

There are various potential improvements which can be further applied on the model to enhance it further:

- **Label and Retrieve data sets**: Data set which has been used for building InformationRetrievel model didn't had labels for entire set. This is typical challenge with most of the retrieval problems. We can start with a pool of pre-defined tags based on domain or industry which we can use for labelling the data set. InformationRetrievel algorithm can work in conjunction with labels and InformationRetrievel algorithm.
- **AnnoyIndex based InformationRetrievel**: Build sentence embedding for Question and Answer and finding best match using AnnoyIndex
- **Alternative algorithm** in place of Universal Sentence Encoder: Models like BertForQuestionAnswering from huggingface can be used to improve the accuracy of the output.