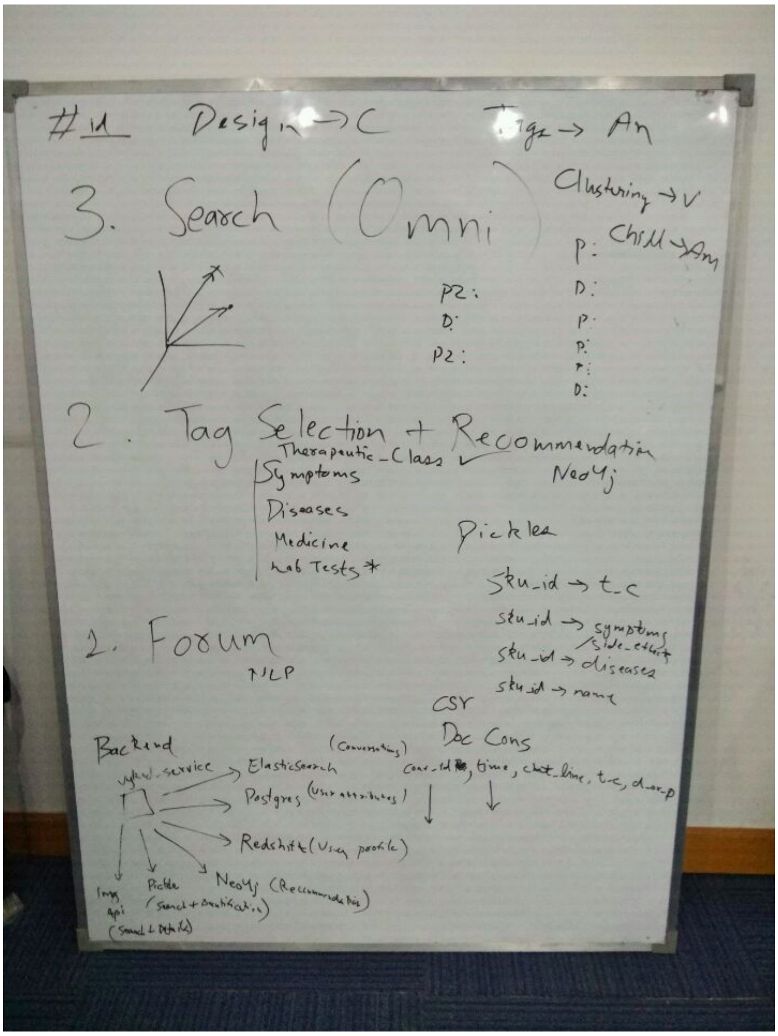
1MG FORUMS NLP

It was the beginning of the month of May when it was announced that a hackathon was going to be conducted in 1mg office. In the hackathon, we were supposed to build a product within two days and give a presentation at the end of it. Our team was composed of a backend developer,~~frontend developer~~ designer, an analyst and an intern who had recently joined our team and was thus renamed to three and a half men.

When we were discussing about the idea of the product to be made, we recognised the need of an online discussion forum such as Reddit and Quora that would be specific to healthcare sector. Hence, we narrowed our research to this area only, and finally came up this initial workflow -



1. We took out the data from Redshift such that the data-frame consists of features such as conversation-id, doctor-id and messages for each conversation.
2. Next, we assigned the tags related to diseases and symptoms to each conversation id, based on exact substring matching. For onboarding each user, we automatically assigned tags based on the user's past purchase history with option present of manually adding tags as well.

3. We also developed a search engine using Elasticsearch where a user can search any specific query present throughout our conversation data.
4. Also, related questions are generated for each conversation. This was done as follows -

(a) Conversation data was cleaned such that no characters were allowed and all the letters were converted uniformly to lower case.
(b) Doctor names, patient names, phone numbers, cities and states were anonymised by using the dictionary of common Indian names present.
(c) For recommendations, our goal was to convert all the conversations into a vector. First, we trained a word2vec model on filtered dataset (removing stopwords and words with no meaning) with downsampling all the frequent words and specifying the minimum word count. This model was able to capture the semantic meaning of each word and to relate it with other similar words since the corpus size was large enough to give reliable results. Word2Vec gives a vector to each word. With our goal in mind, we did a K-means clustering of the word vectors such as each cluster contains words with same semantics. Once getting a bag of clusters, we converted each conversation into a vector. In the last step, we trained a KD tree with 50,000 conversation vectors and found out five nearest neighbors to each conversation, hence giving recommendations.

Words having similar meanings to ‘pimple’ in our vocabulary:

```
('acne', 0.8569689393043518)
('blackhead', 0.7512112259864807)
('pore', 0.749397337436676)
('mole', 0.7424284219741821)
('scar', 0.728550910949707)
('whitehead', 0.7283890247344971)
('pimpels', 0.7152302861213684)
('blemish', 0.7115480899810791)
('pigmentation', 0.7094229459762573)
('face', 0.6988154053688049)
```

Words similar in meaning to “cough”:

```
('coughing', 0.8645079135894775)
('caugh', 0.8304303884506226)
('wheezing', 0.709398627281189)
('sneezing', 0.6942190527915955)
('mucus', 0.6811821460723877)
('caught', 0.6625418663024902)
('bulgum', 0.6577844023704529)
('phlegm', 0.6536963582038879)
('throat', 0.6477687358856201)
```

Similar to ‘sex’:

```
('intercourse', 0.7633767127990723)
('satisfy', 0.6999667286872864)
('sexual', 0.6906836628913879)
('errection', 0.6772133111953735)
('satisfied', 0.6384482383728027)
('ejaculated', 0.6273531913757324)
('erection', 0.6227284073829651)
('intercorse', 0.6179051399230957)
```

These vectors can also prove to be beneficial while creating a Symptom Checker to understand user queries by using these vectors to find synonyms.

It took us around 48 hours to build this from scratch to the frontend part involving rigorous and persistent hard work from all the team members. The hackathon ended on May 12, 2017 followed by a presentation by each team. On the same evening,with a mix of curiosity and nervousness, we began to wait for the result. To our surprise, we came out to be on top and this was the proud moment to remember -

After the hackathon, Gaurav Sir suggested us some necessary changes and asked us to get it ready for production. And this was where the real challenge began…
The main thing that we were lacking was good quality of conversations. The shortcomings we faced in our initial product were -

1. Removing all the characters were removing some important punctuations as well hence smothering the meaning of the conversation.
2. Lot of conversations contained the use of foul language (hindi mostly) and words which carried no meanings & having a large length.
3. We encountered some conversations having no significant meaning and low sentence length such as hi, hello, thank you, welcome etc.
4. Dataset also contained duplicate conversations and duplicate messages in a conversation.
5. Names of patients, doctors , phone numbers , names of cities & states of the users were not anonymised completely.
6. Some words are unintentionally anonymised such as ‘gaya’. (gaya is a hindi word for ‘gone’ as well as a city in Bihar)
7. The conversation data appeared way too chatty.

To counter these problems, we categorised our work into three stages that were -

1. Cleaning conversation data :

(a) This involved base filtering on the data such that only patient - doctor (Q & A) format conversations were allowed in order to regulate the chat thread appearance of the forums.
(b) Duplicate conversations and duplicate messages in a conversation were removed.
(c) Quality check was put up on the conversations such that conversations having 75% of words in proper english and not having lengthy meaningless words are only allowed.
(d) Short messages and messages containing specific words on doctor’s side such as ‘report’ (in ‘please share your report’) are removed.
(e) Conversations having question marks from the doctor’s side were removed in order to maintain the Q & A format.
(f) Final check was put on conversations having only single message either from user or doctor’s end.

2. Anonymising conversations :

(a) Patient’s names are anonymised (replaced with “PATIENT”) using dictionary containing names corresponding to each conversation id only so that words with different meanings are not anonymised unnecessarily.
(b) Doctor’s names, cities and states are left untouched.
(c) SKU’s and OTC’s names are searched for in the conversation from the entire list of 1mg’s products and links have been provided so that the user can be directed to the 1mg’s website page corresponding to that OTC or SKU.

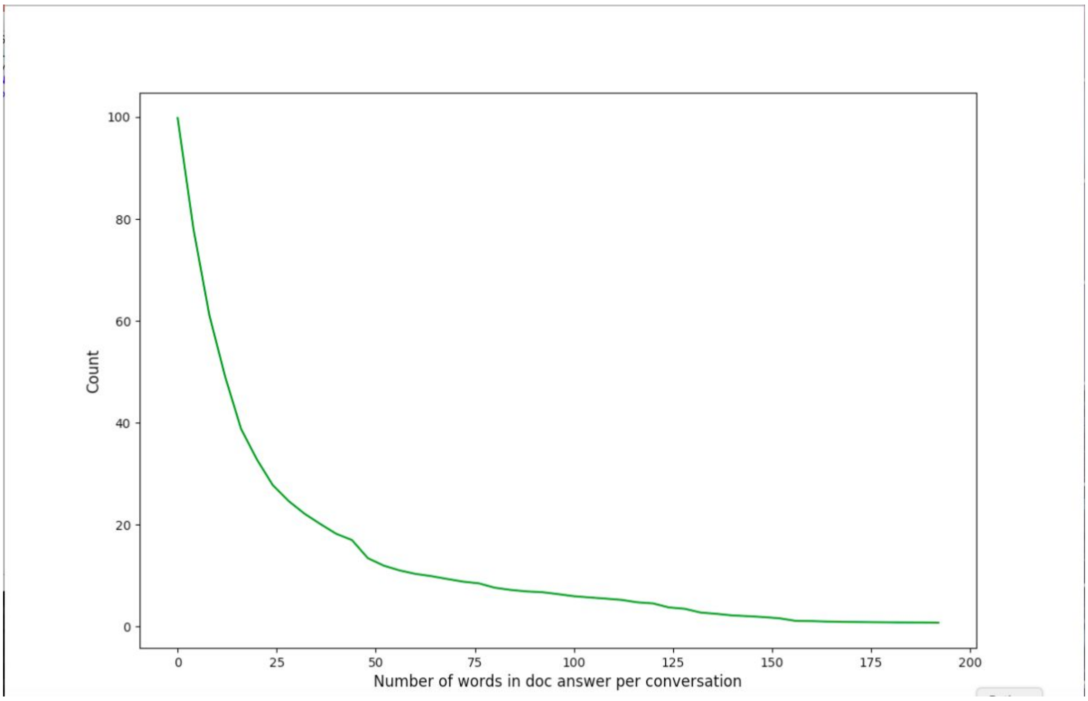
3. NLP part :

(a) The goal is to find the recommended conversations.

(b) This has been kept the same as the first model except that some parameters have been tuned due to the change in the conversation data and corpus available.

The problem we still faced is the short length of doctor's answers and lack of quality parameter so that conversations can be sorted.

Distribution of word counts of doctor's chunks each conversation of a random sample:



In order to overcome these shortcomings, we tried these methods -

1. **Patient - Doctor - Patient - Doctor pattern in a conversation :**

- In order to obtain better quality and to capture more conversation's messages, we tried to establish a conversation in which user's question will be the chunk consisting of user's starting messages and doctor's answer will comprise of the largest chunk of messages replied to the user in continuation thus tried to capture conversational data other than the single Q & A format.
- We also tried to make a conversation consisting of largest chunks of both doctor's and patient's conversations.

The design made for Version 0 asked for the Q & A format only, so this idea has not been applied as of yet.

2. Quality Score for each conversation :

After applying all these methods and despite all the efforts, we were not be able to get quality conversations having proper doctor's answers.

The reasons for diminished quality of doctor's answers were as follows -

(a) Many answers were generic responses such as "Consult a physician", "Kindly upload picture of the affected area" etc.

(b) Some doctors were involved in sending duplicate answers (CTRL + C and CTRL + V technique) to similar as well as different problems of the patients. Examples of duplicate answers are as follows -

> 'u can call vinod raina on 9XXXXXX or visit www.sexologistdoctors.com'
>
> 'Healthy weight gain is important, so focus on diet and exercise both-eat dates milkshake, nut, paneer, eggs, non-veg, fruits and fruit milkshake and home-cooked meals do some weight training, with proper exercise you can also include some protein supplements'

In order to overcome these issues, we gave score_1 and score_2 for each message based on TF-IDF method. These are explained as follows -

1. **Score_1**

- In this approach , we trained the TF-IDF (Term Frequency - Inverse Document Frequency) Vectoriser at a global level i.e. including all the words from the doctor's as well as user's messages.
- The model returns a list of tuples comprising of a word and its weight, highest when word occurs many times within a small number of messages (thus lending high discriminating power to those messages) and lowest when the term occurs in virtually all messages.
- Scores are given to each message by counting the number of words having a weight higher than the threshold & dividing it by the total number of words in a message.

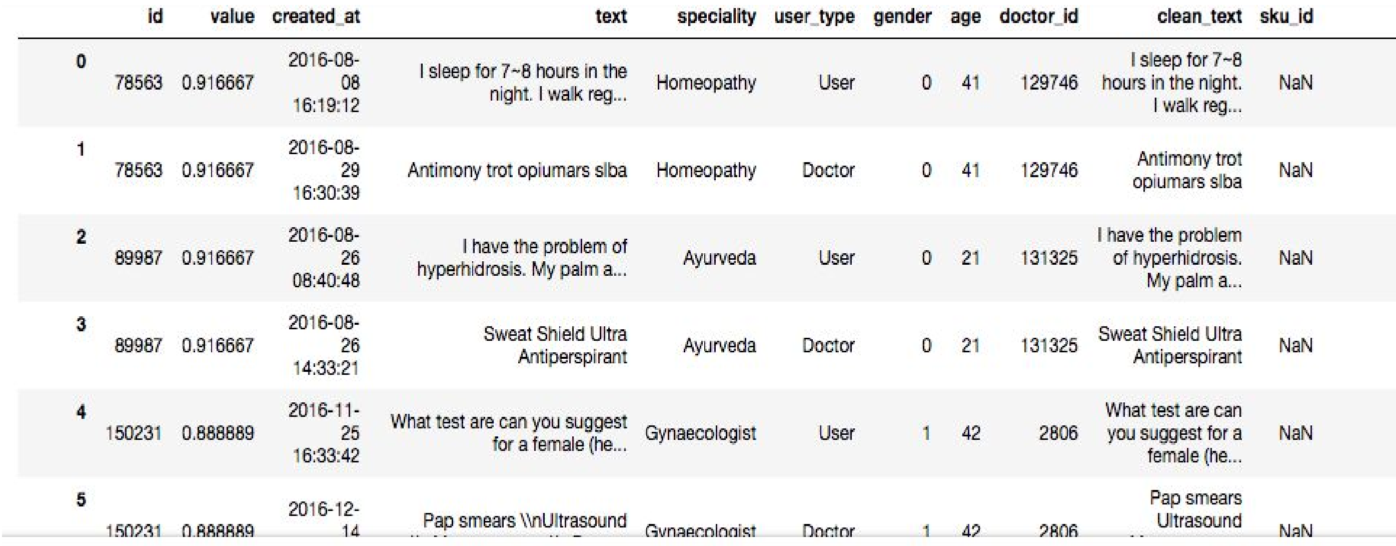
2. **Score_2**

- In this approach, TF-IDF Vectoriser was trained at each doctor's id level in order to detect the duplicate answers in a conversation.
- Similarly, model returns a list of tuples of word and weight & scores are given to each message.
- The difference being in this method is that the duplicate answers (frequent words present in message) will be getting a low score at each doctor's level.

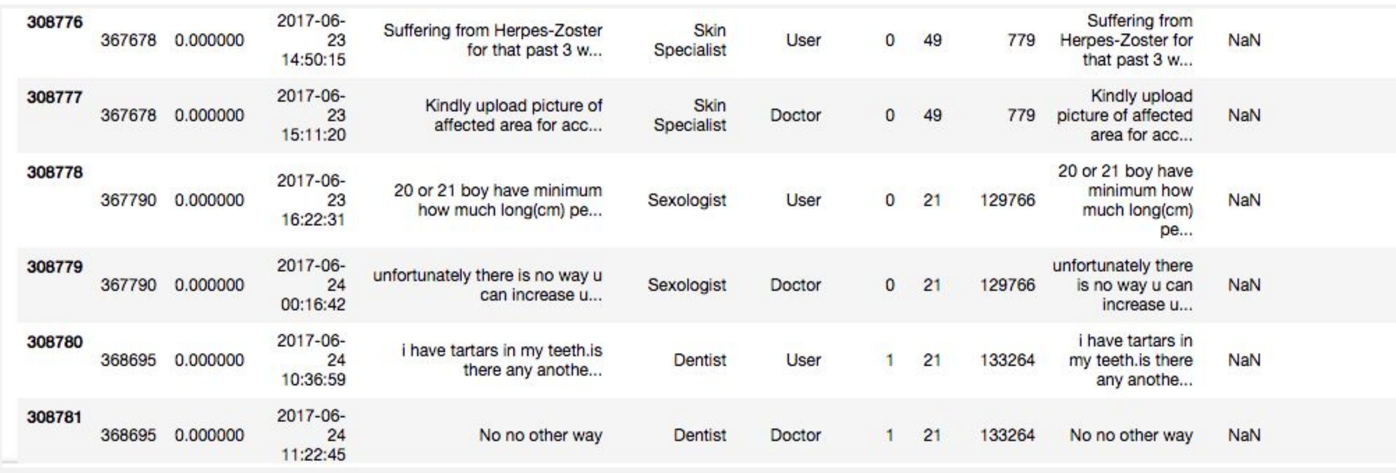
The final score per message is the weighted average of score_1 and score_2 with more weight given to the former.

For each conversation, a rating is given which is the average of the final scores of doctor's messages.

When sorted by these ratings per conversation in a decreasing order, these are the conversations coming on top and the bottom respectively (value being the rating) -



	id	value	created_at	text	speciality	user_type	gender	age	doctor_id	clean_text	sku_id
0	78563	0.916667	2016-08-08 16:19:12	I sleep for 7~8 hours in the night. I walk reg...	Homeopathy	User	0	41	129746	I sleep for 7~8 hours in the night. I walk reg...	NaN
1	78563	0.916667	2016-08-29 16:30:39	Antimony trot opiumars slba	Homeopathy	Doctor	0	41	129746	Antimony trot opiumars slba	NaN
2	89987	0.916667	2016-08-26 08:40:48	I have the problem of hyperhidrosis. My palm a...	Ayurveda	User	0	21	131325	I have the problem of hyperhidrosis. My palm a...	NaN
3	89987	0.916667	2016-08-26 14:33:21	Sweat Shield Ultra Antiperspirant	Ayurveda	Doctor	0	21	131325	Sweat Shield Ultra Antiperspirant	NaN
4	150231	0.888889	2016-11-25 16:33:42	What test are can you suggest for a female (he...	Gynaecologist	User	1	42	2806	What test are can you suggest for a female (he...	NaN
5	150231	0.888889	2016-12-14	Pap smears \\nUltrasound	Gynaecologist	Doctor	1	42	2806	Pap smears Ultrasound	NaN



308776	367678	0.000000	2017-06-23 14:50:15	Suffering from Herpes-Zoster for that past 3 w...	Skin Specialist	User	0	49	779	Suffering from Herpes-Zoster for that past 3 w...	NaN
308777	367678	0.000000	2017-06-23 15:11:20	Kindly upload picture of affected area for acc...	Skin Specialist	Doctor	0	49	779	Kindly upload picture of affected area for acc...	NaN
308778	367790	0.000000	2017-06-23 16:22:31	20 or 21 boy have minimum how much long(cm) pe...	Sexologist	User	0	21	129766	20 or 21 boy have minimum how much long(cm) pe...	NaN
308779	367790	0.000000	2017-06-24 00:16:42	unfortunately there is no way u can increase u...	Sexologist	Doctor	0	21	129766	unfortunately there is no way u can increase u...	NaN
308780	368695	0.000000	2017-06-24 10:36:59	i have tartars in my teeth.is there any anothe...	Dentist	User	1	21	133264	i have tartars in my teeth.is there any anothe...	NaN
308781	368695	0.000000	2017-06-24 11:22:45	No no other way	Dentist	Doctor	1	21	133264	No no other way	NaN

In the final step , a threshold value of the final rating is chosen and the conversations above the threshold are sent to the frontend. Moreover, the search results from ElasticSearch are also sorted on the basis of this score to show good quality results on the top.

Conclusion

We attempted to put the data generated at 1mg to good use to understand its usefulness and strengths. It was not an easy journey going from unstructured low-quality chats beneficial only to one user to Forums with significantly higher quality which can directly benefit several 1mg users immediately.