Natural Language Processing
Application of NLP

MedApp : An automated medical diagnosis system

Submitted by :
2019UCO1505 Nalin Semwal
2019UCO1514 Abhishek Jha
2019UCO1516 Kartik Goyal
2019UCO1580 Harshit Gupta

Submitted to : Anisha Gupta

Overview

The main aim of this project is to provide an automated medical diagnosis system which could serve as an initial endpoint for a person experiencing health problems.

Domain of NLP : Multiclass Text Classification
Language : Python

Working

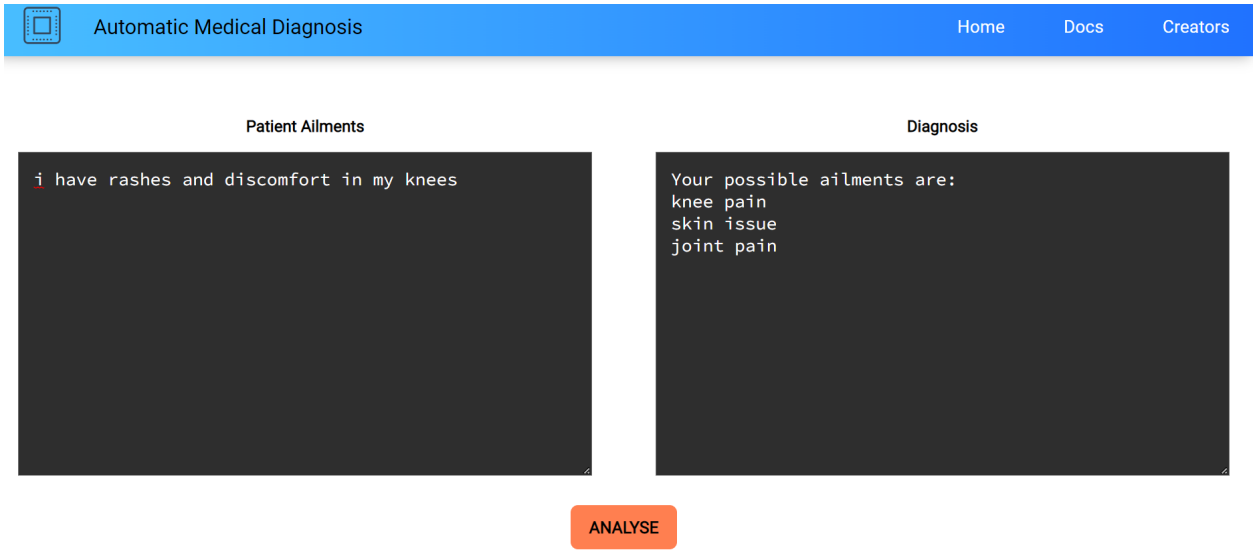
Patient can describe his symptoms in the text box. When diagnosis is called using the button, the result can be seen alongside. Based on a text based query of the user, our model provides a list of ailments which are the most probable to be experienced by the user.

Website link : https://ephemeral-dusk-0ef28c.netlify.app/

Github repository : https://github.com/kartik-3513/amd

Try these sentences :

- I have redness in my eyes and a strep throat
- I have weak knees
- I have weak knees that hurt a lot



Dataset and Preprocessing

The dataset used for the application consisted of a **list of text queries associated with their respective ailments**.

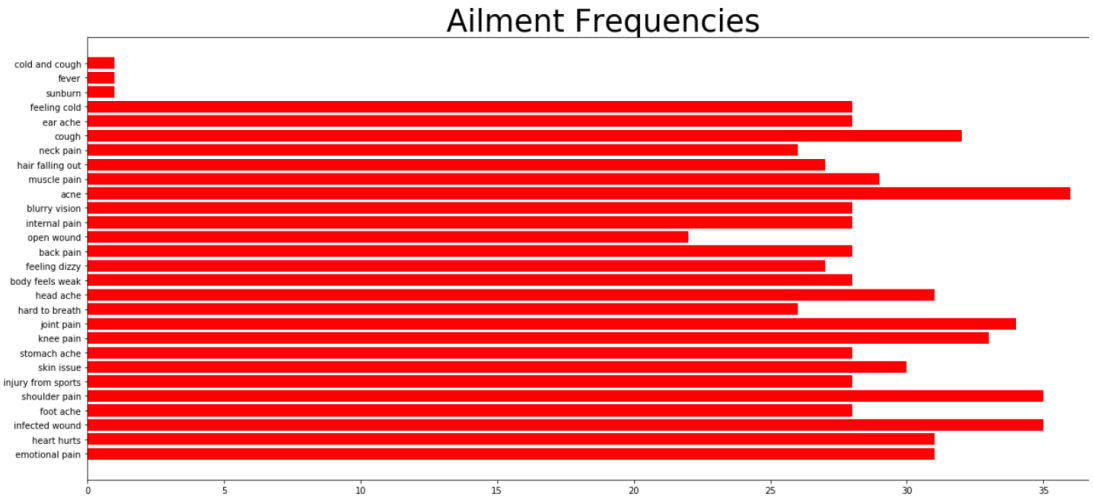
In [40]:
```
data
```

Out[40]:

	Phrase	Prompt
0	When I remember her I feel down	Emotional pain
1	there is too much pain when i move my arm	Heart hurts
2	My son had his lip pierced and it is swollen a...	Infected wound
3	My muscles in my lower back are aching	Infected wound
4	i have muscle pain that my back\nI Have Muscle...	Foot ache

Preprocessing

1. Duplicate entries from the data were dropped
2. Punctuations were eliminated
3. Abbreviations such as what's, shouldn't etc. were expanded with a custom dictionary
4. Each string was converted to lowercase.
5. SnowballSetmmer from nltk library was used to break down the words to their respective stems.
6. The query column was dropped and the stemmed phrases were used as the basis for our models

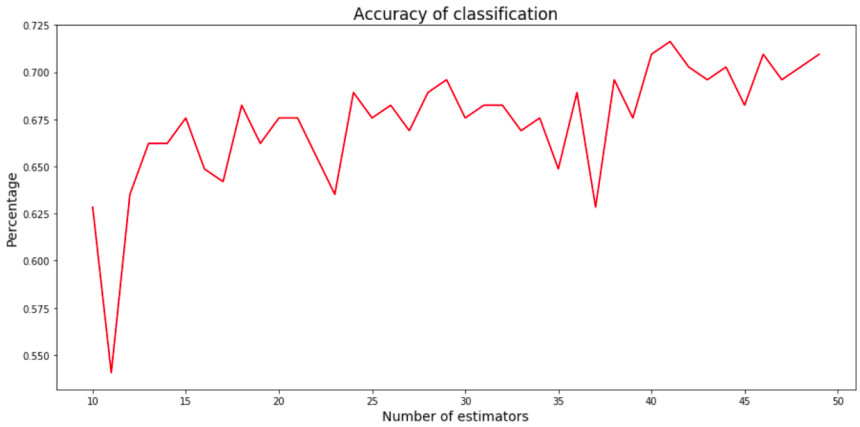


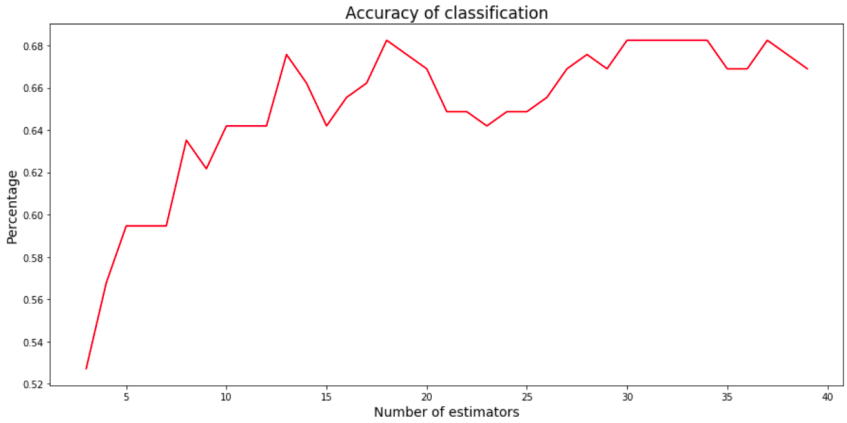
However, to increase the size of the data, we tried to **collect queries through google forms to make our data richer**.

Encoding and training

We used the TfIdfVectorizer from skearn to vectorize the dataset across its vocabulary. The **ngram range used was [1,3]** so that very large phrases are not included in the vocabulary. After the preprocessing of our data, the vocabulary stood at ~ 5000 elements across which a query was transformed.

After that, KNN and RandomForest models were trained on the data. The below results were seen with the model training size of 80% with figure 1 showing results of the RandomForestModel and figure 2, the KNN mode





Results and Inference

The RandomForestClassifier model with 42 estimators was used with ~ 71% accurate results. These were some results of the model on the held out testing data where 3 most probable ailments of the model are shown along with the original ailment -

```
Query : you will not believ me but this infect wound on my hand is from a paper cut i did not take serious
Original : infected wound
infected wound,open wound,skin issue,
Query : these red spot on my cheek are new what is it
Original : skin issue
skin issue,infected wound,acne,
Query : i have a skin rash after eat an ice-cream
Original : skin issue
skin issue,open wound,muscle pain,
Query : i have a sharp pain in my abdomen
Original : internal pain
internal pain,muscle pain,back pain,
Query : i feel a strang and power pain insid my rib cage
Original : heart hurts
internal pain,heart hurts,back pain,
```