Smart India Hackathon, Software Edition - 2019

By: Niharika Shrivastava

Indian Institute of Information Technology, Allahabad

Abstract

This paper describes the most significant computer science project I have worked on till date. It was a **product prototype** built during the **Smart India Hackathon, Software Edition - 2019**, in Hyderabad, India. The event was conducted from 2 March, 2019 - 4 March, 2019 stretched over 36 hours of hacking time. It was a team event consisting of six members. The problem statements provided were **current, real-world problems** that had no solutions. Being the **winner team** for the proposed problem statement, our solution is being actively scaled to fix the actual problem at hand.

Problem Statement

To implement intelligent natural language search for all R&D data of Dr. Reddy's Lab

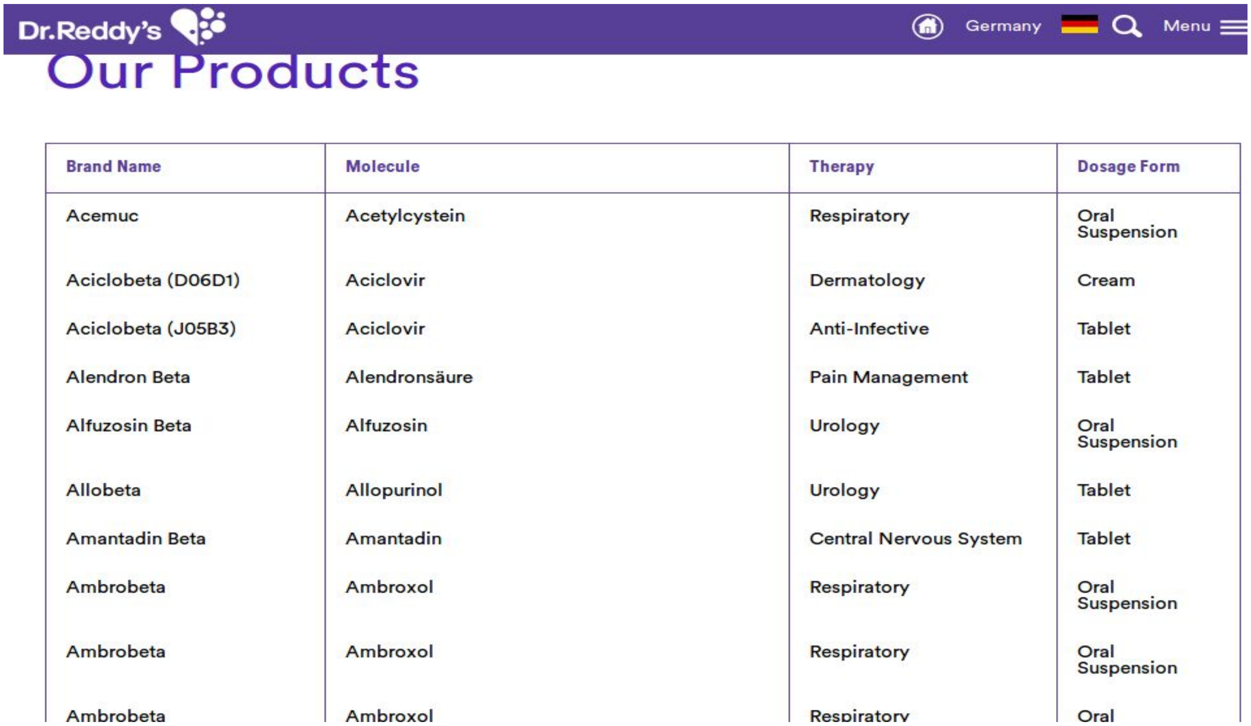
We place all our R&D reports, findings, etc. in the form of PPT slides in a shared folder. Is there a way to automatically build knowledge from these and help us with insights for any given search term?

Proposed Solution

Key technical challenges

1. No robust centralized hub: All the R&D data is either in the form of PPT slides or PDFs in a shared folder. This folder can be easily corrupted or manipulated (anyone can add/delete material) that will cause huge loss of valuable information.

2. Unstructured data: None of the R&D data is categorized therapy wise, creating a highly cumbersome dump of data.
   Eg: None of the products under *Respiratory* are grouped together.



Our Products

Brand Name	Molecule	Therapy	Dosage Form
Acemuc	Acetylcystein	Respiratory	Oral Suspension
Aciclobeta (D06D1)	Aciclovir	Dermatology	Cream
Aciclobeta (J05B3)	Aciclovir	Anti-Infective	Tablet
Alendron Beta	Alendronsäure	Pain Management	Tablet
Alfuzosin Beta	Alfuzosin	Urology	Oral Suspension
Allobeta	Allopurinol	Urology	Tablet
Amantadin Beta	Amantadin	Central Nervous System	Tablet
Ambrobeta	Ambroxol	Respiratory	Oral Suspension
Ambrobeta	Ambroxol	Respiratory	Oral Suspension
Ambrobeta	Ambroxol	Respiratory	Oral

3. No research-progress tracking: The researchers working on a particular molecule are unable to identify if a project has already been worked upon, or is being worked upon by someone else currently.

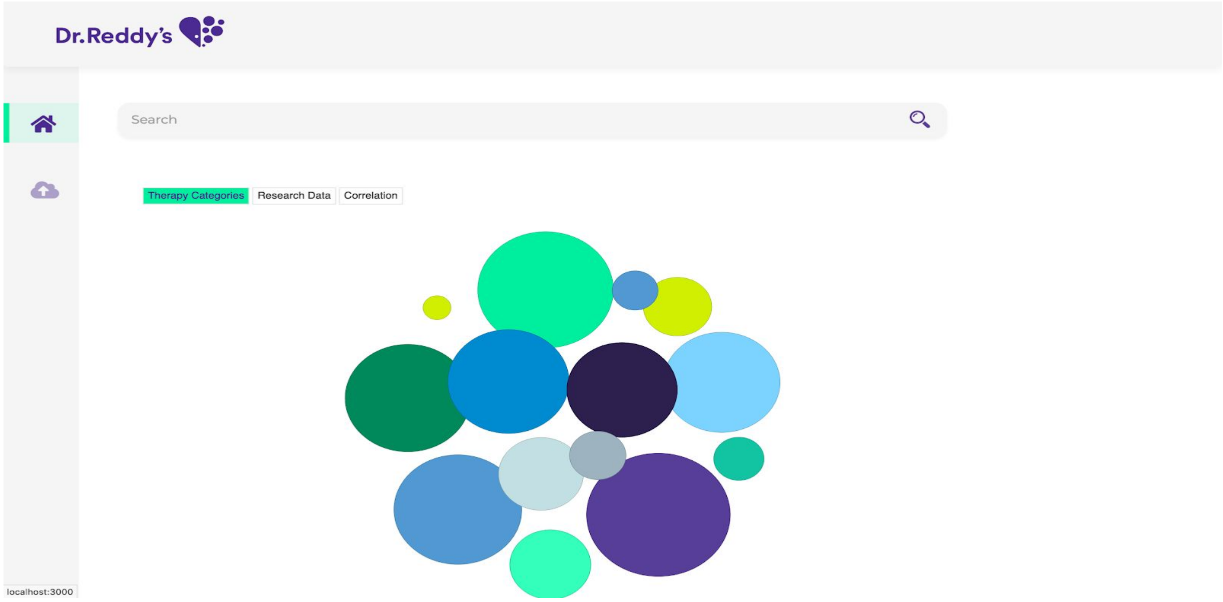
4. Lack of visual statistics: Loss of knowledge by not correlating important statistical features like amount of research made on each therapy per year.

5. No relational structure between data: There is no inplace query system that can help find insights for given search terms. Researchers have to scan through the entries to find the relevant data.

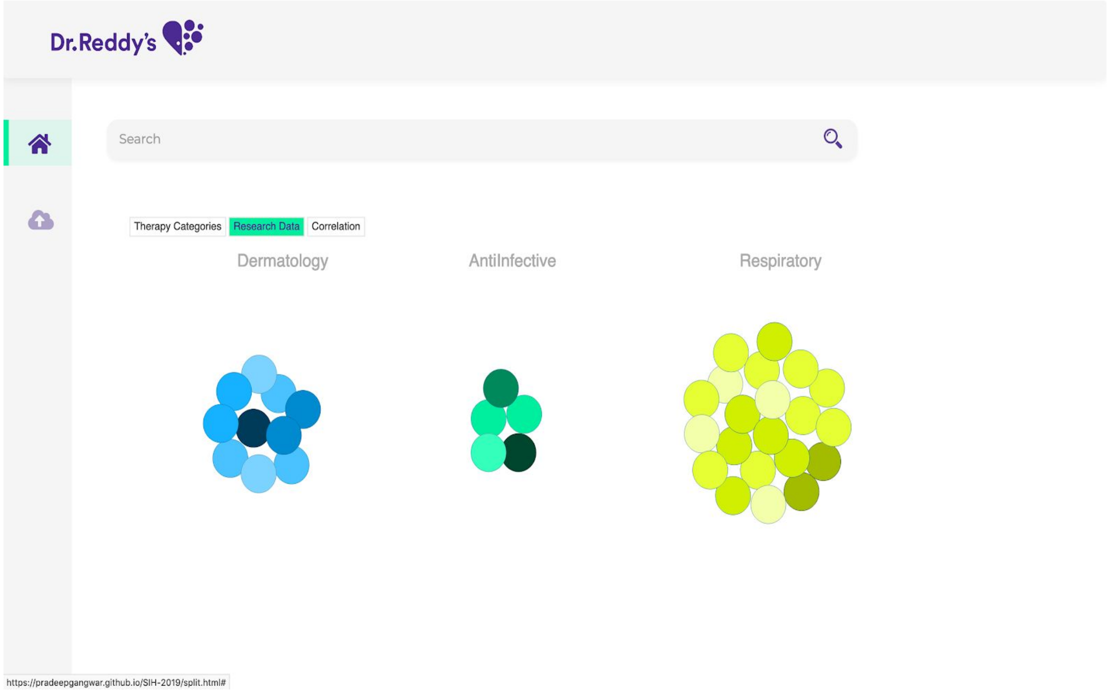
Product features

We built a user-friendly dashboard that will be a consolidation of all the R&D data along with a centralized knowledge base. Its features include:-

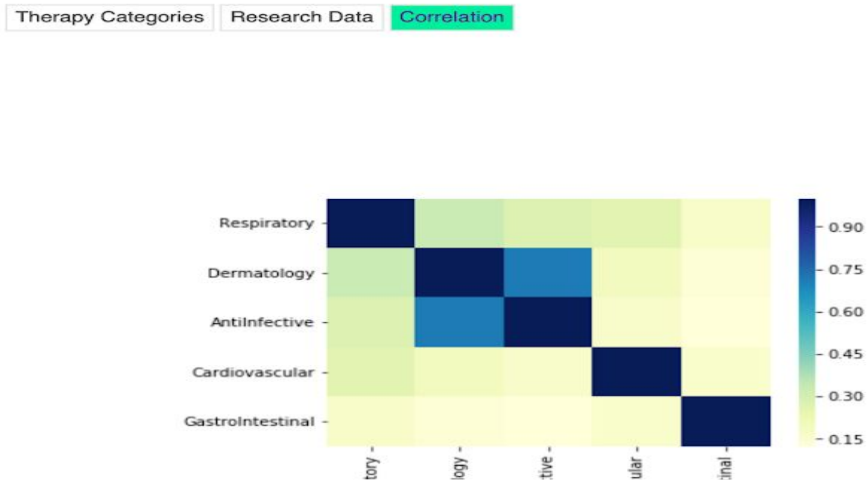
1. **Data visualization of the entire data**:
    a. *Therapy categories:* If a user wants to explore DR. Reddy's archives not specific to any topic, they see the broad therapy categories that have available R&D data. Each bubble represents a category, and the size of the bubble is proportional to the amount of research done in the area. Hovering onto the bubbles gives exact percentages and stats. Clicking a bubble lands onto all the archives pertaining to the therapy category.



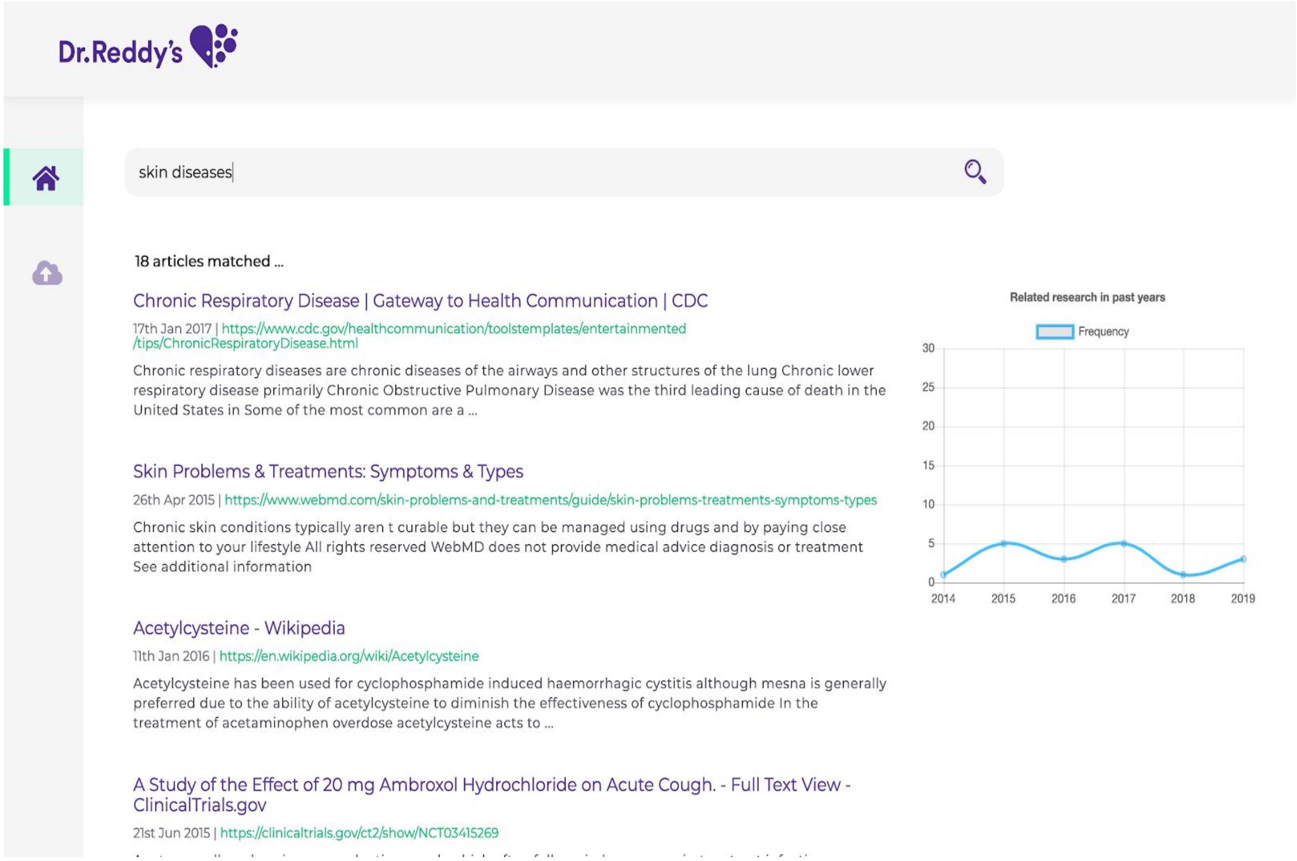
b. *Research Data:* If a user is looking for specific information that has already been worked upon, they see the R&D data segregated according to the different therapy categories. The number of bubbles in a category is proportional to the amount of R&D articles. Each bubble in a category corresponds to a research PPT or PDF. Clicking onto the bubble will direct you to the targeted PDF/PPT.



c. *Correlation:* A correlation matrix is generated between every therapy category specifying how closely the research in one therapy is related to another.



2. **Intelligent answers for specific/general queries:** Any query written (like a Google search) will provide results of specific R&D data containing information about them.



18 articles matched ...

Chronic Respiratory Disease | Gateway to Health Communication | CDC

17th Jan 2017 | https://www.cdc.gov/healthcommunication/toolstemplates/entertainmented/tips/ChronicRespiratoryDisease.html

Chronic respiratory diseases are chronic diseases of the airways and other structures of the lung Chronic lower respiratory disease primarily Chronic Obstructive Pulmonary Disease was the third leading cause of death in the United States in Some of the most common are a ...

Skin Problems & Treatments: Symptoms & Types

26th Apr 2015 | https://www.webmd.com/skin-problems-and-treatments/guide/skin-problems-treatments-symptoms-types

Chronic skin conditions typically aren t curable but they can be managed using drugs and by paying close attention to your lifestyle All rights reserved WebMD does not provide medical advice diagnosis or treatment See additional information

Acetylcysteine - Wikipedia

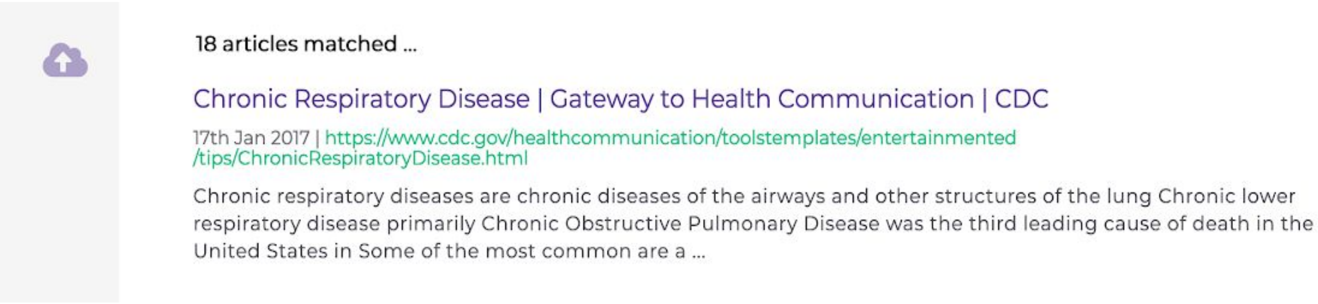
11th Jan 2016 | https://en.wikipedia.org/wiki/Acetylcysteine

Acetylcysteine has been used for cyclophosphamide induced haemorrhagic cystitis although mesna is generally preferred due to the ability of acetylcysteine to diminish the effectiveness of cyclophosphamide In the treatment of acetaminophen overdose acetylcysteine acts to ...

A Study of the Effect of 20 mg Ambroxol Hydrochloride on Acute Cough. - Full Text View - ClinicalTrials.gov

21st Jun 2015 | https://clinicaltrials.gov/ct2/show/NCT03415269

3. **Summarizing specific R&D data for easy exploration:** Every PPT is summarized automatically for quick scanning of what the article contains. We don't need to open every PPT to find what we're looking for.



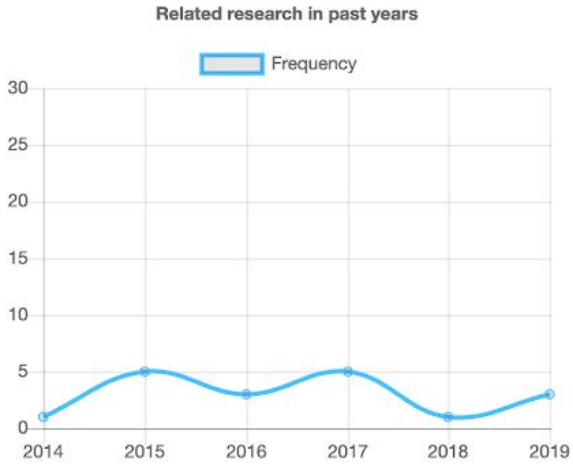
18 articles matched ...

Chronic Respiratory Disease | Gateway to Health Communication | CDC

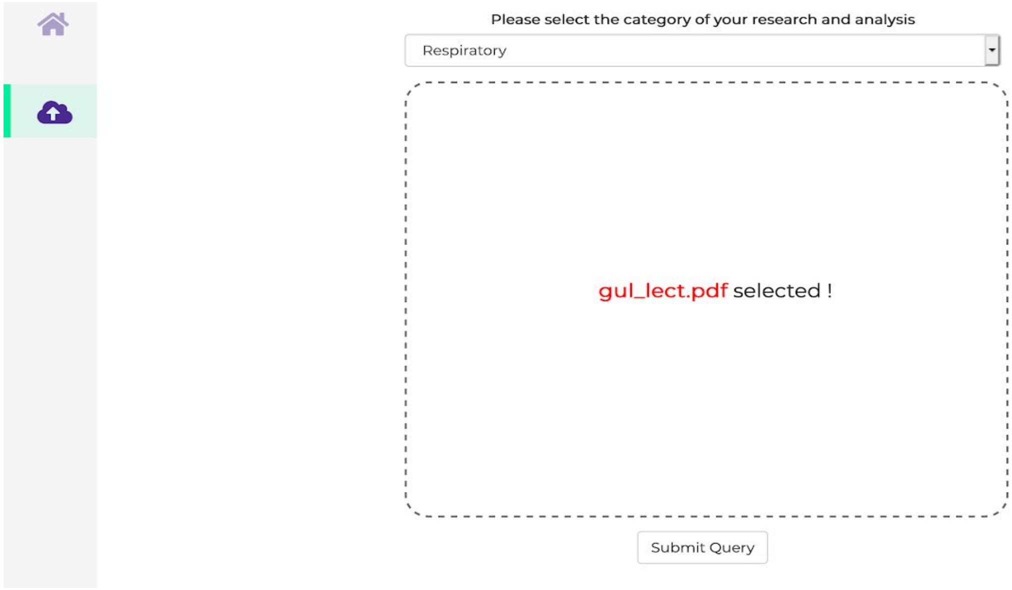
17th Jan 2017 | https://www.cdc.gov/healthcommunication/toolstemplates/entertainmented/tips/ChronicRespiratoryDisease.html

Chronic respiratory diseases are chronic diseases of the airways and other structures of the lung Chronic lower respiratory disease primarily Chronic Obstructive Pulmonary Disease was the third leading cause of death in the United States in Some of the most common are a ...

4. **Research progress tracking enabled:** With every query, a graph is generated simultaneously depicting the amount of research work available for that topic per year.



5. **Uploading a new file in PPT, PDF or any format:** To add any new R&D data, a simple upload is enabled that stores the work in a central database, runs all the algorithms for knowledge building, and updates the visualizations and stats on the dashboard.



Please select the category of your research and analysis

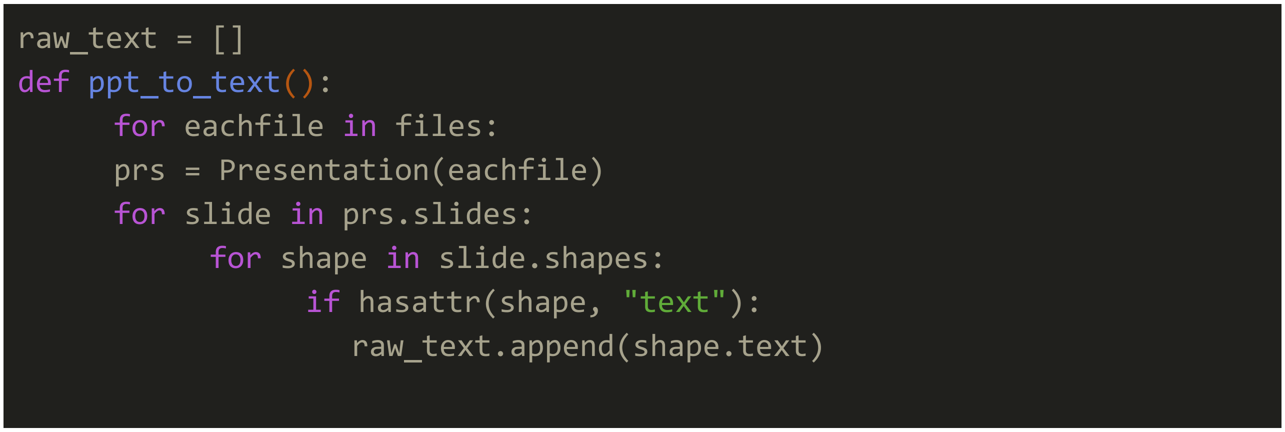
Respiratory

gul_lect.pdf selected !

Individual Role

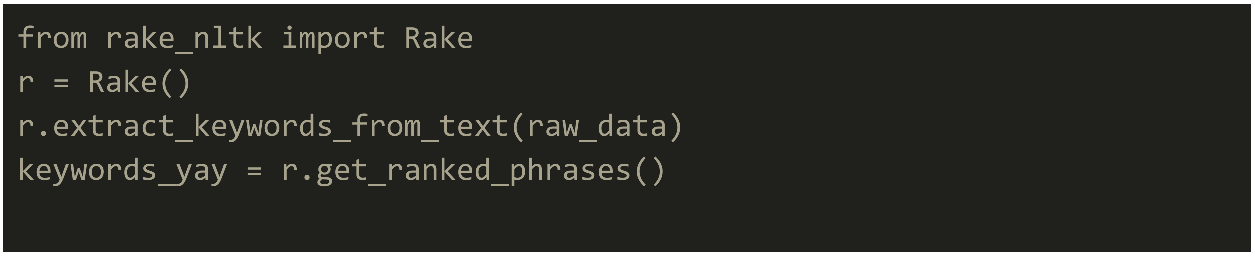
1. I was responsible for extracting data from the PPTs / PDFs and shape them in a manner to enable query search and correlation.

    a. Extract data in raw text:



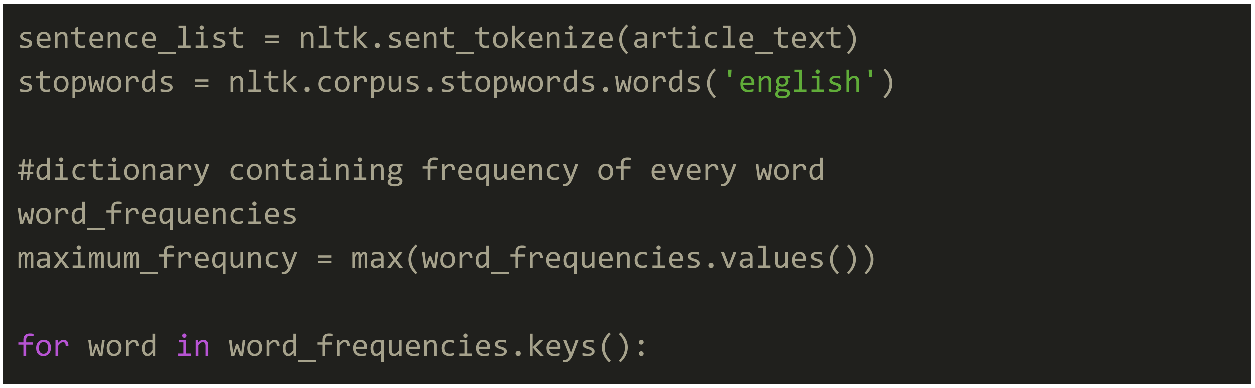
```
raw_text = []
def ppt_to_text():
     for eachfile in files:
     prs = Presentation(eachfile)
     for slide in prs.slides:
           for shape in slide.shapes:
                if hasattr(shape, "text"):
                   raw_text.append(shape.text)
```

    b. Form keywords: Used ***nltk*** to form upto 3-gram keyphrases. Removed stopwords and special characters.



```
from rake_nltk import Rake
r = Rake()
r.extract_keywords_from_text(raw_data)
keywords_yay = r.get_ranked_phrases()
```

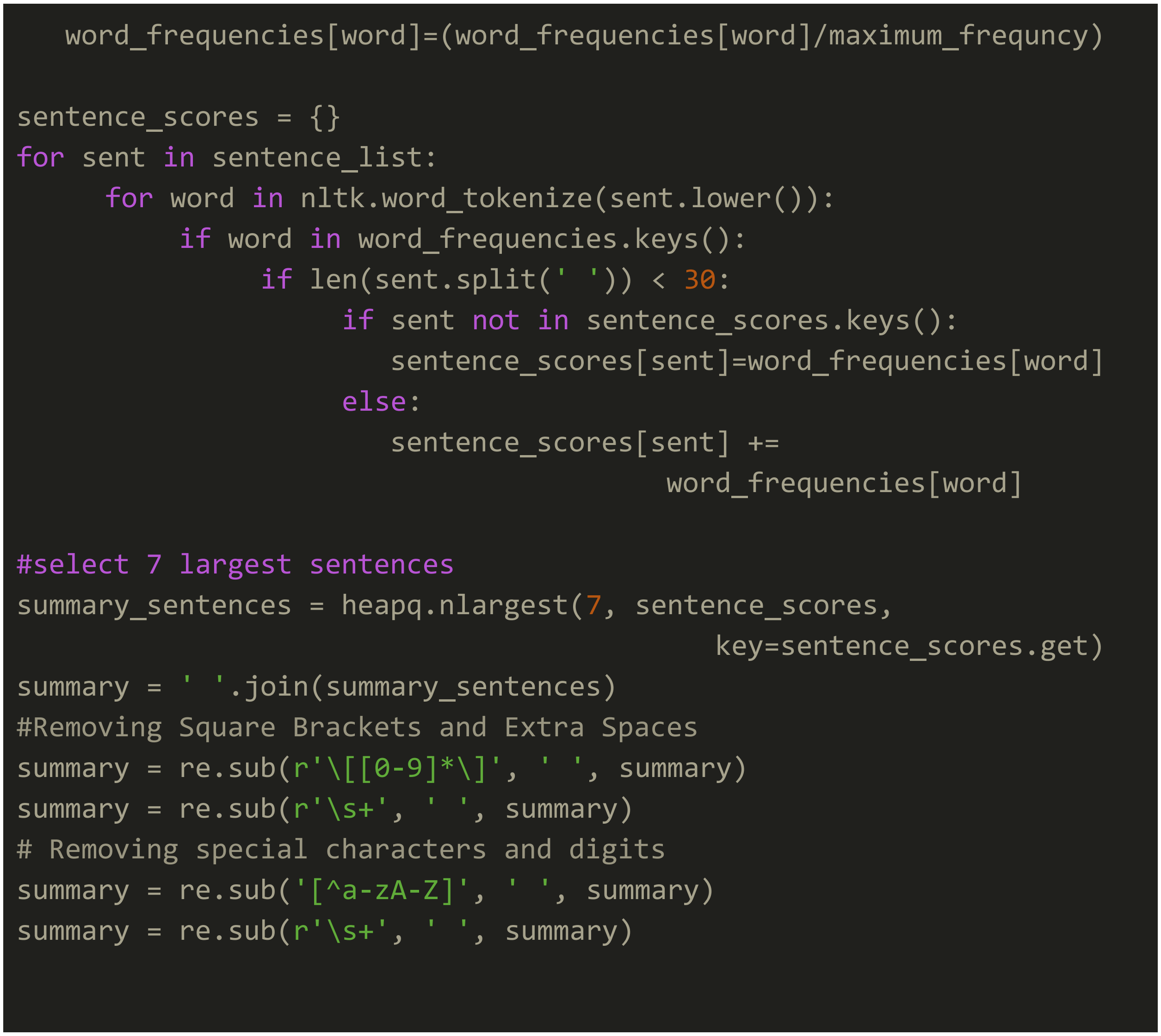
    c. Summarize PDFs/PPTs: By calculating the **TF-IDF value** of each word.



```
sentence_list = nltk.sent_tokenize(article_text)
stopwords = nltk.corpus.stopwords.words('english')

#dictionary containing frequency of every word
word_frequencies
maximum_frequncy = max(word_frequencies.values())

for word in word_frequencies.keys():
```

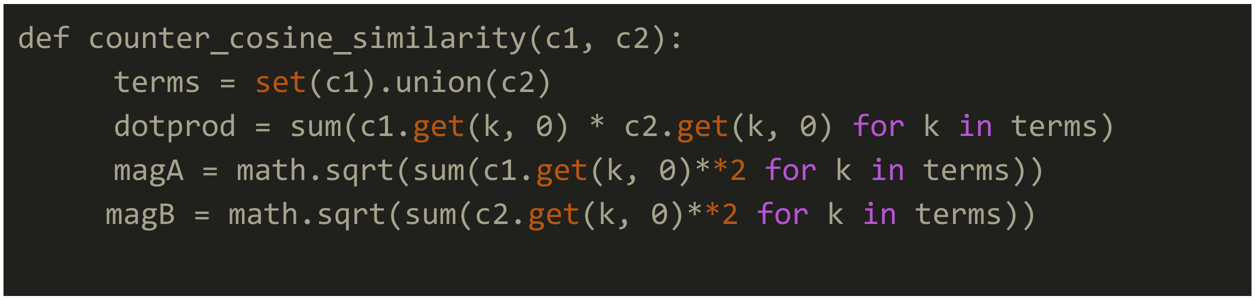


```
   word_frequencies[word]=(word_frequencies[word]/maximum_frequncy)

sentence_scores = {}
for sent in sentence_list:
      for word in nltk.word_tokenize(sent.lower()):
          if word in word_frequencies.keys():
               if len(sent.split(' ')) < 30:
                    if sent not in sentence_scores.keys():
                       sentence_scores[sent]=word_frequencies[word]
                    else:
                       sentence_scores[sent] +=
                                        word_frequencies[word]

#select 7 largest sentences
summary_sentences = heapq.nlargest(7, sentence_scores,
                                           key=sentence_scores.get)
summary = ' '.join(summary_sentences)
#Removing Square Brackets and Extra Spaces
summary = re.sub(r'\[[0-9]*\]', ' ', summary)
summary = re.sub(r'\s+', ' ', summary)
# Removing special characters and digits
summary = re.sub('[^a-zA-Z]', ' ', summary)
summary = re.sub(r'\s+', ' ', summary)
```

d. <u>Find correlation between articles:</u> Used **Natural Language Processing** to establish cosine similarity between keyphrases, thereby deciding degree of closeness of articles.



```
def counter_cosine_similarity(c1, c2):
      terms = set(c1).union(c2)
      dotprod = sum(c1.get(k, 0) * c2.get(k, 0) for k in terms)
      magA = math.sqrt(sum(c1.get(k, 0)**2 for k in terms))
     magB = math.sqrt(sum(c2.get(k, 0)**2 for k in terms))
```

2. There were 4 sets of evaluations. I gave all the 4 presentations and discussed prospective goals and changes to steer our project in the right direction.

Results

We were awarded the **first prize and a cash award of 75,000 INR.** The prototype is being scaled currently to suit Dr. Reddy's Lab needs.

Experience

I learnt how to work in a team and deliver effective solutions in a constrained time frame. This experience made me realize that solutions can come from any sphere of the society. It was a positive experience overall.

References

https://www.drreddys.com/germany/our-products/ - Dr. Reddy's lab products

https://www.drreddys.com/india/portfolio/therapy-areas/ Dr. Reddy's therapy areas