Mining Healthcare Forums

Akash Nigam †
anigam6@asu.edu

Mayur Padaval †
mpadaval@asu.edu

Nitika Garg †
nnitika@asu.edu

Sandeep Nadella †
vnadell1@asu.edu

Shubham Gupta †
sgupt198@asu.edu

Vinay Matcha †
vmatcha@asu.edu

† *Graduate Student*, Computer Science (MCS), Arizona State University

***Abstract*—Healthcare forums contain tons of information about diseases, treatments, and drugs especially from the personal experience of people. But that information is largely in unstructured form i.e. in the natural language as people recount their experiences.**

The lack of information in a structured form regarding disease-symptom-drug relation is one of the major drawbacks when it comes to drawing useful inferences and insight from data. This is the area which we are trying to address by mining healthcare websites. If large amounts of data on the medical information of various patients are analyzed and adequately interpreted, it can be used to identify the medical trends. This, in turn, can improve the quality of the pharmaceutical industry's decision-making process.

Keywords—Healthcare Mining web data mining, Web scraping, Web Mining, MetaMap, Ontology, Healthcare forums, UMLS

I. INTRODUCTION

The aim of this project is to build an information extraction system that can turn unstructured medical healthcare data from user posts of multiple sources (in WebMD.com[6], Drugs.com[7], Patient.info[8]) into structured information and build a parametric search interface for a category (a disease/symptom/drug). Using the search interface, users will be able to get the different attributes of input disease, possible diseases for given symptoms, or the usage of the input drug, etc. Also, the search output data attributes will be ranked based on their occurrence frequency on data sources, to help users in differentiating the more common ones to the less frequent ones. The attributes of a disease include symptoms, the category of disease, treatment.

II. PROBLEM STATEMENT

Our work is inspired by the following existing problems:

A. Unstructured Health-related experience data

Healthcare forums allow a user to write their healthcare experience without following any particular format. It allows users to recount and express their experience in a more natural way hence increase the usability of the platform and makes it human readable. But this flexibility reduces the ease of extracting knowledge out of that data using machines as natural language is largely unstructured.

B. Experiences spread across Multiple Sources

There are multiple forums on the web today that extracting and combining knowledge from data on a large scale can be very helpful in finding trends.

C. No Disease - Symptom search tools for data mined from user posts

Due to the lack of structure, there is no easy way to get the aggregated data from all the content from multiple sources in one place.

Following is the solution Proposed:

i. Unstructured Data ➔ Structured Data

Our goal is to mine data from these forums and convert it into a structured form.

Unstructured Data	**Structured Data**
"I saw an orthopedist because of pain and swelling in both knees 2 years ago. Arthritis Org said I need a total knee replacements. I am 75 and do not want to undergo surgery so I curtailed hiking which had been causing pain, I lost some weight. They don't hurt any worse since starting this exercise."	*Disease*: Arthritis *Symptoms*: Pain NOS Adverse Event, Swelling, Weight decreased, *Treatment*: Knee Replacement Arthroplasty, Exercise

Table 1. Structured data retrieved from Unstructured data

ii. Search Interface to query Top diseases and symptoms

By using the data collected from the healthcare forums, it can be aggregated in database and a search interface can be created to query top symptom, treatments for a disease and vice versa.

Top Diseases for a Symptom	**Top Symptoms of a Disease**	**Top Treatment for a Disease**
Symptom: Sore Throat **580** people reported Sore Throat. • **70%** had Epiglottitis • **64%** had Influenza	Disease: Malaria **980** people reported Malaria. • **80%** reported Fever • **70%** reported Headache	Disease: Influenza **1193** people reported influenza • **93%** took Rapivab • **80%** took Relenza

Table 2. Search output

III. Related Work

A variety of state-of-the-art methods have been employed for information extraction in humongous data present in medical and health domain. As clinical notes which make up the most significant component of the Electronic health records are mostly unorganized

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. SympGraph: A framework for mining clinical notes through symptom relation graph

One of the excellent works is SympGraph [1] mining framework which can be used to model and analyze the symptom relationships using clinical notes. SympGraph is a type of graph in which nodes are symptoms of different diseases and health problems and edges are the relationship between those symptoms. This SympGraph framework can be used to expand some given symptoms by utilizing the created SympGraph.

B. MedEx: A medication information extraction system for clinical narratives

Another such admirable work is MedEx [2] which is a natural language processing system that uses clinical notes to extract medication information. It can be used to reduce medication errors when a patient is transferred from one care setting to another. It can be also be used in EMR based clinical research which requires detailed medical information.

IV. System Architecture and Algorithm

Our project is divided into four major parts.

A. Web Scraping

B. Annotating and extracting medical concepts from Text

C. Ontology Creation

D. Search Interface

A. Web Scraping

We are scraping multiple healthcare websites. On these websites, the user posts are categorized based on the disease family. We are extracting following details from each user post -

- Disease name
- Post link
- Post heading
- Post content
- Post tags

Post tags and disease name can be useful for the user posts which does not produce any semantic type for the disease after passing it through MetaMap annotator. If this is the case, we append the disease name and post tags into the post heading and rerun the process.

For scraping, we are using Python's Scrapy[3] which is a fast and efficient framework that have tools to manage every stage of a web crawl, like Request Manage, Selectors and Pipelines. It uses XPath and CSS selector to identify elements needed on the web page.

Following is the screenshot for the generated CSV

disease	postHeading	postContent	postTags
Diabetes	Diabetes	Here recently I've had an increase of urination...Especially at night. I seem to be thirsty more and have dry m	Symptoms
Diabetes	Foot Ulcer	I have had a foot ulcer for about 3 months now. The foot doctor keeps cutting it away but it doesn't seem to	Type 2 Diabetes
Diabetes	I have low back pains so I have reduced my	I have low back pains so I have reduced my exercising scared of surgery I have noticed that I am really havin	g circulations problems is th
Diabetes	my type 1 wife keeps nodding off	hey guys, so my wife was diagnosed with type 1 about 6 years ago (due to cancer radiation damaged her pa	Type 1 Diabetes , Symptoms
Diabetes	Diabetic insulin	Can I take insulin after my meal??	Insulin , Type 2 Diabetes
Arthritis	Second hip replacement	Is it common to need a hip replacement on the left hip after having the right replaced	Hip Replacement
Asthma	Restrictive Airway or prelude to Asthma?	About two years ago I had a severe chest cold, complete with chills and fever, that lasted about 2 weeks. Ever	since then my throat feels
Sexual Healt	Possible Genital Herpes? Am going absolut	Ok so let me tell you a story. Apologies for the length Thursday 14th March I met a lady that id been talking	Genital Herpes
Asthma	Chronic Day Time Cough ONLY	My son is 11 years old .Since three months he is coughing persistently ONLY in day time and during night wh	Living & Managing , Triggers
Asthma	Do I have asthma?	I am a 13 yr old female who has trouble breathing in or anywhere around smoke, humid air, dust, and pollen	Living & Managing , Triggers
Asthma	Problems breathing	I've never been diagnosed with asthma or breathing issues. but I do have difficulties breathing and catching	Treatment , Diagnosis
Asthma	Does cold air help someone with asthma?	Knew a girl that would put her head in the freezer if she didn't have her inhaler. Does this really help?	Diagnosis
Asthma	Grand daughter woke up in the night.havin	My 11 year old grand daughter woke up with heavy chest.strugling to breath.she didn't have a fever wasn't c	Diagnosis
Osteoporosis	hyperparathyroidism not mentioned as cau	In your recent article and slide show on osteoporosis, you failed to mention hyperparathyroid disease that is a readily surgically correctabl	
Asthma	Another 'do I have asthma?' - spirometry n	33 y/o female ex smoker. I took the spirometry test for four times and all showed I have breathing issues. I c	Diagnosis , Tests
Osteoporosis	Upset about the diagnosis of osteoporosis	I just turned 72. Until last year, my bone density tests revealed osteopenia. Last year for the first time it revi	Bone Density
Lupus	Pseudogout	Ifm a 61 yr old women that has been experiencing Pseudogout since 2006. It started in my ankles, then my elbows . but I was not diagnos	
Lupus	Lupus and low Immunoglobulin A levels (Ig	I was diagnosed with SLE 6 years go and am 23 years old. I also have celiac disease. I recently found out that	Diagnosis
HIV/AIDS	Is it HIV symptoms?	Blood in semen TNTC in urine	

Figure 1. sample csv generated by scrapper

B. Annotating and extracting medical concepts from Text

After scraping the data from healthcare forums, we have the user comments in a CSV file. Now, we need to identify diseases, symptoms and treatments from the post. There are many different terminologies for a single disease and symptom. Different people use different terms to describe a disease. So, we need to have the knowledge of a medical domain expert who identify these different terminologies as a single disease, symptom or treatment.

UMLS: UMLS is a medical Thesaurus developed by NLM (National Library of Medicine). It consists of multiple standardized set of medical terminologies and relation among them.

It has three knowledge sources.

Metathesaurus	**Semantic Network**	**SPECIALIST Lexicon & Tools**
1 million+ **biomedical concepts** from over 100 sources	135 broad **categories** and 54 **relationships** between categories	**lexical information** and **programs** for language processing

Table 3. Components of UMLS

Metathesaurus handles synonyms and assigns unique IDs. Each term is called a "concept" and it also maintains hierarchical relations among concepts.

Semantic network categorizes concepts among diseases, symptoms, treatment, anatomy etc.

Specialist Lexicon tools handle the NLP part and spelling variants etc.

MetaMap: Now, we have this expert knowledge from UMLS, we need a tool that can use this knowledge to extract and annotate data. MetaMap is one such tool that annotates and extract medical concepts from text using UMLS. It is also developed by NLM. It is highly configurable in terms of

selecting UMLS resources and output etc. It uses NLP and computational-linguistic techniques

A sample text processed by MetaMap is as follows

Input Text	MetaMap Output
"I saw an orthopedist because of pain and swelling in both knees 2 years ago. Arthritis Org said I need total knee replacements. I am 75 and do not want to undergo surgery so I curtailed hiking which had been causing pain, I lost some weight. They don't hurt any worse since starting this exercise."	Pain NOS (Pain NOS Adverse Event) [Finding] SWELLING (Swelling) [Finding] knees (Knee) [Body Part, Organ, or Organ Component] total knee replacements (Knee Replacement Arthroplasty (procedure)) [Therapeutic or Preventive Procedure] lost weight (Weight decreased) [Finding] ARTHRITIS (Arthritis) [Disease or Syndrome] Exercise (Exercise Pain Management) [Therapeutic or Preventive Procedure]

Table 4. Output of MetaMap for the given input text

MetaMap Annotator Implementation: This project uses MetaMap annotator Java API for annotating the web data scrapped by the scraper project. The mmserver should be setup and running prior to running this project. The location of the CSV file containing the scraped data, MetaMap options should be given in the class file.

MetaMap's Java API and mmserver files can be downloaded from MetaMap's official website.

MetaMap's Java API is highly configurable. We can provide a list of words to ignore, words to include etc. We can also configure the sources it refers from UMLS.

We can also configure the output i.e. which semantic types to retrieve in the output.

For our project, we extracted the following semantic types

1. Diseases i.e. dsyn | Disease or Syndrome,
2. Symptoms i.e. sosy | Sign or Symptom,
3. Treatments i.e. topp | Therapeutic or Preventive Procedure,
4. Drugs i.e. clnd | Clinical Drug and
5. Body part i.e. bpoc | Body Part, Organ, or Organ Component

MetaMap annotates data post by post and write the result into CSV file. To handle posts without a disease name, we got the concept name from the post category under which user posted. The posts which do not contain any useful information will automatically be ignored by MetaMap as nothing would be annotated under given semantic types.

C. *Ontology Creation*

Ontology describes the various entities of our information extraction system and their underlying relationships. The various entities are disease, symptom, Treatment & anatomy.

- We have created Disease, Symptom and Treatment ontology.
- Objects and relationships will be stored using tables in relational DB. We have used MySQL for out project.
- Relationships will consist of disease related with their symptom, treatment, anatomy. These will be used for giving search interface query output.
- The output of MetaMap is inserted into the ontology database. For this we are first loading the MetaMap output csv data into a table and then using SQL stored procedure and cursor to put that data in our ontology tables.
- Weight of each symptom, treatment, for a particular disease will be kept. This will allow to get k top frequent symptom, treatment for a disease and vice versa.

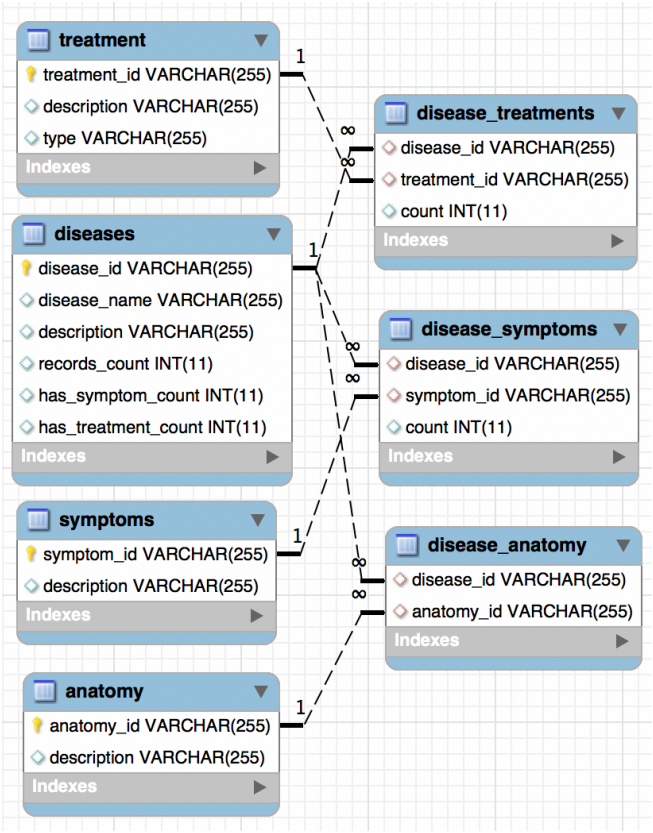


Figure 2. Ontology schema design

Below is the sample database figure showing an example 'Acne' disease and its related symptoms. Symptom count denotes the number of users reporting that symptom.

disease_id	disease_name	record_count	total_symptom_count	symptom_count	symptom_id	symptom_name
C0702166	Acne	553	105	15	C0030193	Pain
C0702166	Acne	553	105	11	C0015230	Exanthema
C0702166	Acne	553	105	7	C0033774	Pruritus
C0702166	Acne	553	105	6	C0027769	Nervousness
C0702166	Acne	553	105	5	C0221423	Illness (finding)
C0702166	Acne	553	105	4	C0018681	Headache
C0702166	Acne	553	105	4	C0026821	Muscle Cramp
C0702166	Acne	553	105	3	C0234233	Sore to touch
C0702166	Acne	553	105	3	C0085624	Burning sensation
C0702166	Acne	553	105	2	C1446787	Cramping sensation quality
C0702166	Acne	553	105	2	C0015672	Fatigue
C0702166	Acne	553	105	2	C0849370	blemishes
C0702166	Acne	553	105	2	C0557875	Tired
C0702166	Acne	553	105	2	C0312414	Menstrual spotting

Figure 3. SQL table representing disease-symptom relationship

D. Search Interface :

We have created web interface for symptom checker and disease checker. Symptom checker is for checking the diseases associated with symptoms, based on the symptom given by the user, the response will be the diseases that associated with the symptom and percentage of people that reported that symptom. This is like inverse document frequency, percentage of people reported the symptom for a disease will be the total number of people reported for that disease divided by the total number of people reported for all the diseases.

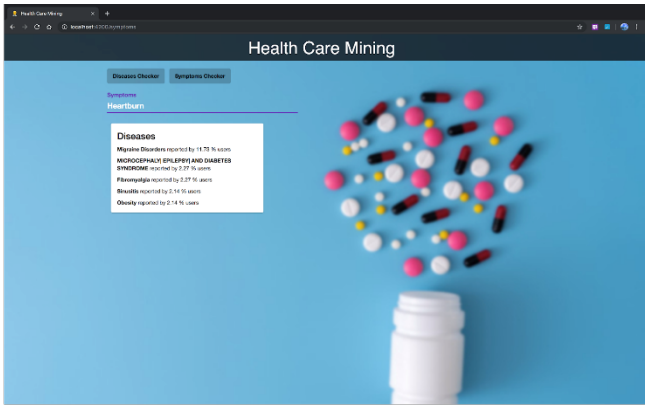


Figure 4. Search interface: Symptom Checker

Disease checker is for checking the symptoms and treatments associated with that disease, (i.e.) given a disease name by the user, the response will be the top five symptoms and treatments associated with that disease.

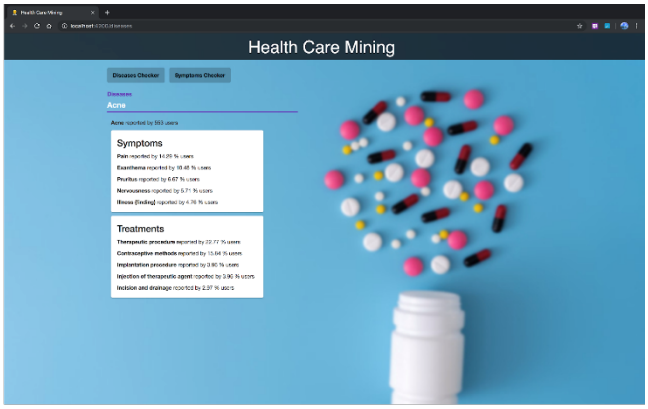


Figure 5. Search Interface: Disease Checker

We have used Angular with Spring Boot for building the search interface. In order to avoid the search calls for the diseases and symptoms which are not in database, we have used ajax query to provide options for diseases and symptoms which are present in database only. Users cannot select the disease or symptoms which are not in database. When the user enters the disease or symptom name in the search interface and submits the button, the REST API call happen and brings the response for the query from the database. We have used JPA hibernate queries for bringing information from the database for the REST API calls.

V. DATASETS (DESCRIPTIONS, SIZES AND PREPROCESSING STEPS)

The datasets for this project are the Discussion Forums from health care websites where data is posted in unstructured form by users. e.g.

- https://www.patient.info/
- https://www.drugs.com/
- https://www.webmd.com/

We have scraped around 35000 user posts from drugs.com and 10000 from webmd.com and 9000 from patient.info for all available medical conditions on these websites.

VI. DEPLOYMENT AND EVALUATION

The execution pipeline involves running the following modules in sequence

1. Scraper
2. MetaMap Annotator
3. Ontology Creation
4. Search Interface Backend Deployment
5. Search Interface Frontend Deployment

A. Scraper:

The scraper can be run on each site using the following commands from the */HealthCareMining/scraper* folder

$ scrapy crawl patient_info
$ scrapy crawl webmd
$ scrapy crawl drugs

The debug output of the scraper is as follows

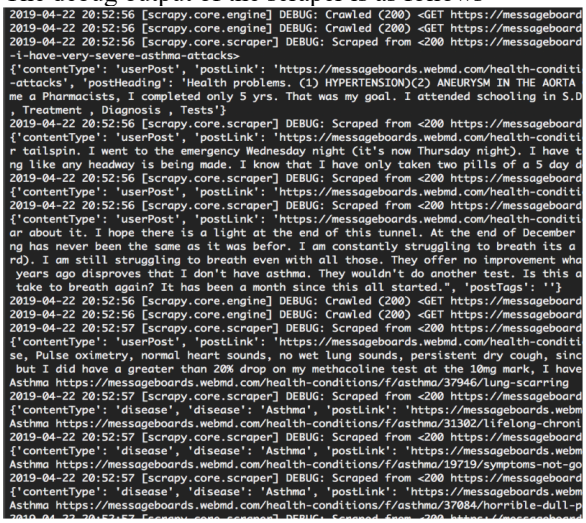
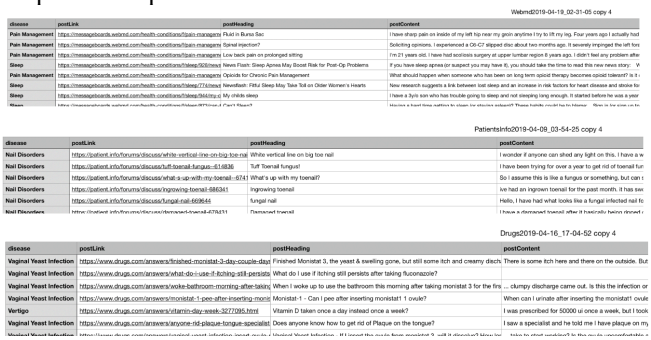


Figure 6. Debug output of the scrapper

As can be seen from the screenshot the scraper crawls through the various post links and extracts the post content from the HTML tags. The *contentType* represents the type of data being extracted. It is then used for mapping the post category name with the post content. Various content types like *postHeadings, postLink, postTags, postContent* are extracted.

The extracted data is written to a CSV file under */HealthCareMining/scraper/data* folder. The structure of the output of scraper if as follows



Webmd2019-04-19_02-31-05 copy 4

disease	postLink	postHeading	postContent
Pain Management	https://messageboards.webmd.com/health-conditions/pain-management	Fluid in Burss Sac	I have sharp pain on inside of my left hip near my groin anytime I try to lift my leg. Four years ago I actually had
Pain Management	https://messageboards.webmd.com/health-conditions/pain-management	Scleral injection?	Soliciting opinions. I experienced a C6-C7 slipped disc about two months ago. It severely impinged the left foo
Pain Management	https://messageboards.webmd.com/health-conditions/pain-management	Low back pain on prolonged sitting	I'm 21 years old. I have had scoliosis surgery at upper lumbar region 6 years ago. I didn't feel any problem afte
Sleep	https://messageboards.webmd.com/health-conditions/sleep/928/news	News Flash: Sleep Apnea May Boost Risk for Post-Op Problems	If you have sleep apnea (or suspect you may have it), you should take the time to read this new news story: V
Pain Management	https://messageboards.webmd.com/health-conditions/pain-management	Opioids for Chronic Pain Management	What should happen when someone who has been on long term opioid therapy becomes opioid tolerant? Is it s
Sleep	https://messageboards.webmd.com/health-conditions/sleep/774/news	Newsflash: FDA Sleep May Take Toll on Older Women's Hearts	New research suggests a link between lost sleep and an increase in risk factors for heart disease and stroke for
Sleep	https://messageboards.webmd.com/health-conditions/sleep/564/mg-c	My childs sleep	I have a 3yrs son who has trouble going to sleep and not sleeping long enough. It started before he was a year

PatientsInfo2019-04-09_03-54-25 copy 4

disease	postLink	postHeading	postContent
Nail Disorders	https://patient.info/forums/discuss/white-vertical-line-on-big-toe-nail	White vertical line on big toe nail	I wonder if anyone can shed any light on this. I have a w
Nail Disorders	https://patient.info/forums/discuss/tuft-toenail-fungus--614836	Tuft Toenail fungus!	I have been trying for over a year to get rid of toenail fun
Nail Disorders	https://patient.info/forums/discuss/what-s-up-with-my-toenail--6741	What's up with my toenail?	So I assume this is like a fungus or something, but can
Nail Disorders	https://patient.info/forums/discuss/ingrowing-toenail-686341	Ingrowing toenail	Ive had an ingrown toenail for the past month. It has sw
Nail Disorders	https://patient.info/forums/discuss/fungal-nail-669644	fungal nail	Hello, I have had what looks like a fungal infected nail fo

Drugs2019-04-16_17-04-52 copy 4

disease	postLink	postHeading	postContent
Vaginal Yeast Infection	https://www.drugs.com/answers/finished-monistat-3-day-couple-days	Finished Monistat 3, the yeast & swelling gone, but still some itch and creamy discha	There is some itch here and there on the outside. But
Vaginal Yeast Infection	https://www.drugs.com/answers/what-do-i-use-if-itching-still-persists	What do I use if itching still persists after taking fluconazole?	
Vaginal Yeast Infection	https://www.drugs.com/answers/woke-bathroom-morning-after-taking	When I woke up to use the bathroom this morning after taking monistat 3 for the firs	... clumpy discharge came out. Is this the infection or
Vaginal Yeast Infection	https://www.drugs.com/answers/monistat-1-pee-after-inserting-monis	Monistat-1 - Can I pee after inserting monistat1 1 ovule?	When can I urinate after inserting the monistat1 ovule
Vertigo	https://www.drugs.com/answers/vitamin-day-week-3277085.html	Vitamin D taken once a day instead once a week?	I was prescribed for 50000 ui once a week, but I took
Vaginal Yeast Infection	https://www.drugs.com/answers/anyone-rid-plaque-tongue-specialist	Does anyone know how to get rid of Plaque on the tongue?	I saw a specialist and he told me I have plaque on my

Figure 7. Sample CSV generated by scraping WebMD, patientinfo, drugs.com

B. *MetaMapAnnotator:*

These files will be fed as input to the next module in the pipeline which is MetaMap Annotator.

To run MetaMapAnnotator module the MetaMap server pipeline should be setup which includes MedPost/SKR POS Tagger server and Word Sense Disambiguation server to be setup and running. The detailed instructions for setup are available under README.md of the MetaMap Annotator module [4].

The debug output is as follows

```
Phrase: h.pylori test,
 Mappings:
Phrase: blood
 Mappings:
Phrase: and
 Mappings:
Phrase: biopsy test for celiac,
 Mappings:
Phrase: tested for pancreatitis
 Mappings:
  Filter Status: false
  Concept Id: C0030305
  Concept Name: PANCREATITIS
  Preferred Name: Pancreatitis
  Matched Words: [pancreatitis]
  Semantic Types: [dsyn]
Phrase: and
 Mappings:
Phrase: everyone
 Mappings:
Phrase: says
 Mappings:
Phrase: the tests
 Mappings:
Phrase: show
 Mappings:
Phrase: nothing.
```

Figure 8. MetaMap debug output

Each user post content is fed to the MetaMap server. The MetaMap server breaks the sentence to phrases and the annotates the lexical match word with the POS tags by feeding it to POS Tagger server. The output is then checked for UMLS concepts and the mappings are generated based on the mappings with highest candidate scores. Here [dsyn] represents disease or syndrome semantic class. The stop words are fed in the *IgnoredWords.csv* file under resources folder which will be filtered out. Helper functions for filtering concepts based on the POS tags are also added and were experimented with.

Words List
to
I
symptom
symptoms
dress
disease
diseases
disorder
disorders
let
hi

Figure 9. Stop word list

The concepts identified are written to output CSV file at */HealthCareMining/MetaMapAnnotator/resources* location. The output files are as follows

PostNumber	DiseaseId	DiseaseName	SymptomId	SymptomName	TreatmentId	TreatmentName	DrugId	DrugName	BodypartId	BodypartName
1	C0029408	Degenerative polyarthritis								
1			C0003862	Arthralgia						
1			C0917801	Sleeplessness						
5	C0029408	Degenerative polyarthritis								
5			C0030193	Pain						
5					C3687832	DRUG5				
6	C0343024	Cellulitis of leg								
6			C0231749	Knee pain						
6			C2220048	Bilateral knee pain						
8	C0029408	Degenerative polyarthritis								
8					C0741419	back surgery				
8									C1269004	Entire mitral valve
8									C0006104	Brain

Figure 10. Output of MetaMap

This file is fed as input to the SQL PROC under the Ontology module. The instructions are available under [5]. This module creates the ontology schema and inserts the data into the database. The frequency of occurrence of the concepts is calculated and stored in database by this module. The data from this database is fetched by the Search interface backend and served to the front end.

C. *Search Interface Backend Deployment*

Open the *searchInterface* module in the *healthcare-data-mining* as a Maven repository, so that all the dependencies in the pom.xml will be downloaded and configured. Start the MySQL server and change the database configuration in *healthcare-data mining/searchInterface/src/main/resources/application.properties*. Once the changes are made, run spring boot application file *searchInterface/src/main/java/com/swm/searchInterface/SearchInterfaceApplication.java*. Apache Tomcat server will be started under URL, http://localhost:8080/.

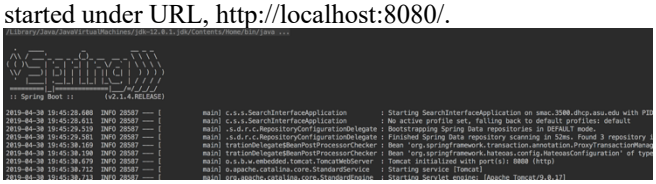


Figure 11. Debug output of back-end server-1

D. *Search Interface Frontend Deployment*

Go to the folder */healthcare-data-mining/searchInterface/src/main/resources/front-end/app-project* in terminal and run the command *ng serve* which will serve the frontend HTML, CSS, JS resources under

URL, http://localhost:4200/.

```
sanadell@smac ~/NetBeansProjects/HealthCareMining/searchInterface/src/main/resources/front-end/app-project (master)
$ ng serve
** Angular Live Development Server is listening on localhost:4200, open your browser on http://localhost:4200/ **

Date: 2019-05-01T02:48:08.061Z
Hash: 0cd572009a930815a8b2
Time: 13087ms
chunk {es2015-polyfills} es2015-polyfills.js, es2015-polyfills.js.map (es2015-polyfills) 284 kB [initial] [rendered]
chunk {main} main.js, main.js.map (main) 39.6 kB [initial] [rendered]
chunk {polyfills} polyfills.js, polyfills.js.map (polyfills) 236 kB [initial] [rendered]
chunk {runtime} runtime.js, runtime.js.map (runtime) 6.08 kB [entry] [rendered]
chunk {styles} styles.js, styles.js.map (styles) 180 kB [initial] [rendered]
chunk {vendor} vendor.js, vendor.js.map (vendor) 7.02 MB [initial] [rendered]
i [wdm]: Compiled successfully.
```

Figure 12. Debug output of back-end server-2

The search interface looks like this

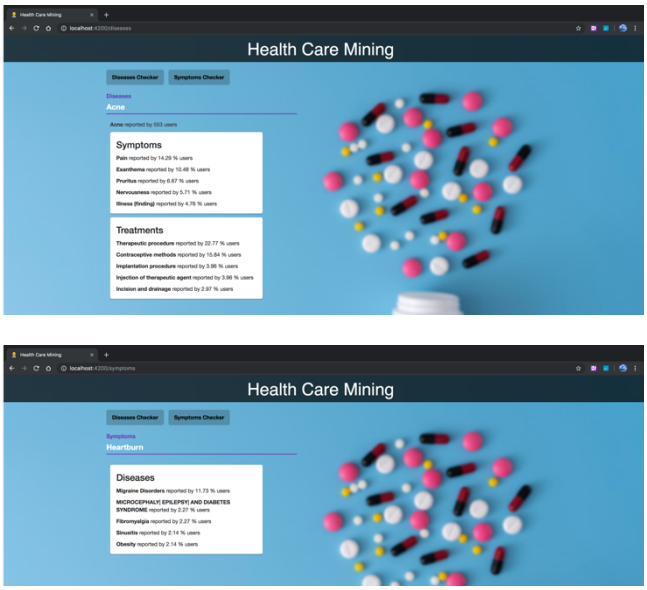


Figure 13. Search Interface

VII. Division of Work and Contributions

Task	Contributors
• Building Web Scraper for scraping unstructured data from multiple sources • Cleaning the user posts for html tags	Akash Nigam, Mayur Padaval
• Building Annotator pipeline using MetaMap • Building filters for stop words • Integration of MetaMap with rest of the modules	Sandeep Nadella
• Customizing MetaMap to filter required Semantic type from MetaMap output. • Handling posts with missing disease names.	Nitika Garg
• Designing the schema for disease, symptom, treatment ontology • Wring queries for schema/table creation.	Shubham Gupta
• Writing stored procedure for data insertion in Ontology tables	
• Setting up Spring and Angular projects • Handling JPA hibernate calls • Handling REST API calls	Vinay Matcha
• UI design	Vinay Matcha, Nitika Garg, Mayur Padaval, Sandeep Nadella
• Integration and Testing	All

VIII. Conclusions

Our main goal is to mine unstructured healthcare data from multiple sources and gain some insights. We extracted data from the discussion forum of three of the most popular healthcare websites. We used MetaMap annotator that finds relative semantic types for diseases, symptoms, and treatments for each scraped post. We designed and implemented the ontology that defines and stores the relationship between different symptoms, diseases, and treatment. Finally, we created a search interface that has disease checker and symptom checker. By using disease checker, the user can find the top 5 symptoms and treatments for the selected input disease. Similarly, the symptom checker returns the top 5 most probable diseases for the input symptom. We are also providing the percentage of people reporting the particular disease for the symptom and vice versa.

For the sake of simplicity, we assumed that user posts do not contain negative contextual information or have multiples diseases. Also, currently our application supports search operation on only one symptom. We can extend our application to handle multiple symptoms to further narrow down the search for the disease. The project source code is available in GitHub [9].

IX. References

1. Parikshit Sondhi, Jimeng Sun, Hanghang Tong, ChengXiang Zhai: SympGraph: a framework for mining clinical notes through symptom relation graphs. KDD 2012: 1167-1175
2. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, et al. (2010) MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 17: 19–24
3. https://docs.scrapy.org/en/latest/intro/overview.html
4. https://github.com/akashnigam/healthcare-data-mining/tree/master/MetaMapAnnotator
5. https://github.com/akashnigam/healthcare-data-mining/tree/master/ontology
6. https://messageboards.webmd.com
7. https://patient.info/forums
8. https://www.drugs.com/answers/conditions
9. https://github.com/akashnigam/healthcare-data-mining