

Medical Chatbot Development: Utilizing Knowledge Graph & Machine Learning Techniques

01

Name of the student	Roll Number
Lakshmikanth Reddy B	CB.EN.U4CCE20011
Sohith Reddy CH	CB.EN.U4CCE20012
Sai Tarun D	CB.EN.U4CCE20014
Gowtham Kumar Y	CB.EN.U4CCE20073

19CCE495 : PROJECT PHASE 2

TEAM ID : ECE013

PROJECT ADVISOR : Dr. PRABHA G

Problem Statement

02

To construct a Chatbot that can answer user's queries based on the information available in Knowledge graph and provide sufficient information regarding medical issues.

Motivation

03

- Patients seek health information online through URLs.
- Chatbot simplifies information access, empowering users to learn about diseases.
- The project is driven by the key motivations which are Enhanced Information Retrieval, Contextual Understanding, Evidence-Based Decision Making, Empowering Patients.

Objectives

- To implement web scraping techniques to gather related medical data from various sources & data cleaning.
- Enhancing and testing the Knowledge graph and building a baseline structure.
- To develop a question-answering system with the developed knowledge graph in integration with the front-end developed website.
- To assess the effectiveness of the created chatbot using various metrics.

Novelty

05

The existing systems has limited interactivity with the chatbot. In our project we would like the user to have maximum understanding of the system using a user friendly website where they can directly chat with an interactive chatbot with easy accessibility and implemented performance metrics.

Literature Survey

06

Title of the paper	Year	Authors	Publication	Comments	Advantages	Drawbacks
A Novel web scraping approach using the additional information obtained from web pages	2020	Erdinc Uzun	IEEE Access	<ul style="list-style-type: none">• How to understand a complicated website.• Studies related to different web scraping techniques.• Technique used: UzunExt.	<ul style="list-style-type: none">• Improving time efficiency of data extraction.• Exceptional speed will increase.• Minimal code overhead.	<ul style="list-style-type: none">• Handling of Complex HTML Structures.• Vulnerability to HTML Changes.• Versatility Across Websites.
COVID-Scraper: An Open-Source Toolset for Automatically Scraping and Processing Global Multi-Scale Spatiotemporal COVID-19 Records	2021	HAI LAN , DEXUAN SHA ANUSHA SRIRENGAN ATH MALARVIZHI , YI LIU	IEEE Access	<ul style="list-style-type: none">• This paper mainly focuses on how to scrap data from different sources dealing with structured and unstructured data	<ul style="list-style-type: none">• Developed a tools for scraping which is very much flexible and scalable.	<ul style="list-style-type: none">• The data quality control and validation is not fully automated because accuracy is not guaranteed.

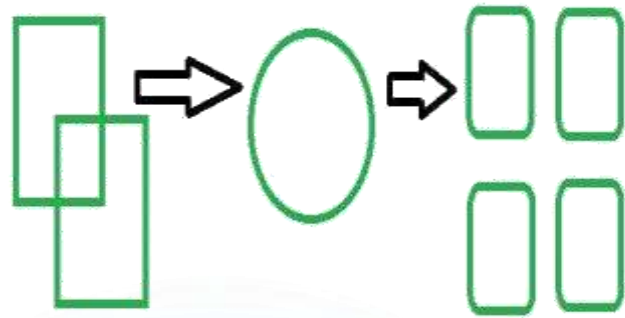
Title of the paper	Year	Authors	Publication	Comments	Advantages	Drawbacks
A Survey on Knowledge Graphs: Representation, Acquisition, and Applications	2021	Shaoxiong Ji , Shirui Pan, Pekka Marttinen , and Philip S. Yu	IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS	<ul style="list-style-type: none"> The paper's strength lies in its thorough coverage of recent advancements and emerging trends in knowledge graph research. 	<ul style="list-style-type: none"> Comprehensive Review of Knowledge Graph Research Structured Categorization and Taxonomies 	<ul style="list-style-type: none"> Complexity and Accessibility of Advanced Topics. Potential Lack of Specific Focus.
Research on medical question answering system based on knowledge graph	2021	Zhhixue Jiang, Chengying Chi, And Yunyun Ahan	IEEE Access	<ul style="list-style-type: none"> An detailed explanation & experiment based on knowledge graph & also query Q&A system. 	<ul style="list-style-type: none"> A Detailed Research on Knowledge graph and question answering system was given. 	<ul style="list-style-type: none"> The attribute from the data are very less only drug , symptom and disease names are considered so accuracy low.

Title of the paper	Year	Authors	Publication	Comments	Advantages	Drawbacks
Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study.	2023	Lino Murali a,c , G. Gopakumar b , Daleesha M. Viswanathan c , Prema Nedungadi	Elsevier Journal of Bio medical information's	<ul style="list-style-type: none">The use of medical knowledge graphs has the potential to improve healthcare outcomes by extracting new links and hidden patterns from health data sources.	<ul style="list-style-type: none">Different approaches of creating medical knowledge graph and their applications.	<ul style="list-style-type: none">Some EHR being physical & unstructured can be a challenging in collection of data.
Complex Knowledge Base Question Answering: A Survey	2022	Yunshi Lan, Gaole He , Jinhao Jiang, Jing Jiang, Wayne Xin Zhao	IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING	<ul style="list-style-type: none">Overview of recent methods in KBQA with a focus on solving complex questions.Key challenges that arise when answering such complex questions.Technique used: SP and IR	IR: <ul style="list-style-type: none">EfficiencyFlexibilityRobustness	IR: <ul style="list-style-type: none">Lack of Context
Design and Development of We-CDSS Using Flask Framework: Conducting Predictive and Prescriptive Analytics for Coronary Artery Disease	2022	DIVYASHREE N. , (Member, IEEE), AND NANDINI PRASAD K. S., (Senior Member, IEEE)	IEEE Access	<ul style="list-style-type: none">This paper is mainly taken into consideration to understand the concepts of Flask framework	<ul style="list-style-type: none">AccessibilitySecurity featuresUser - friendly	<ul style="list-style-type: none">Learning curve

Title of the paper	Year	Authors	Publication	Comments	Advantages	Drawbacks
Question Answering Over Knowledge Graphs: A Case Study in Tourism	2022	Sereh Aghaei, Elie Raad, And Anna Fesnel	IEEE Access	<ul style="list-style-type: none"> Aims to provide answers to natural language questions (NLQs) using KGs. This paper proposes a two-phase approach to QA-KG for small and medium-sized KGs. 	<ul style="list-style-type: none"> More scalable More accurate More flexible 	<ul style="list-style-type: none"> Time-consuming process Over fitting
Knowledge Graph for China's Genealogy	2021	Xindong Wu , Fellow, IEEE, Tingting Jiang Yi Zhu , and Chenyang Bu	IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING	<ul style="list-style-type: none"> This paper focuses on the development of a genealogical knowledge graph model called Huapu-Kg and utilize genealogical knowledge graphs. 	<ul style="list-style-type: none"> Structured data representation Query and analysis Integration of AI Feasibility validation 	<ul style="list-style-type: none"> Data quality integration Scalability Complexity Expertise requirements
Automated clinical knowledge graph generation framework for evidence based medicine	2023	Fakhare Alam a , Hamed Babaei Giglou b , Khalid Mahmood Malik a	Elsevier Journal Expert systems with applications	<ul style="list-style-type: none"> Proposes a topic specific, PICO enabled, and fully automated framework to curate information and create KG of different clinical domains. 	<ul style="list-style-type: none"> Building a knowledge graph according to requirements by using PICO framework are remarkable. 	<ul style="list-style-type: none"> The knowledge graph is built according to the 2 data sets which are considered so they are limited to an extent along with the relations of the KG.

Methodology

10



Website pages –
Unstructured data
(https://)

Web Scraping

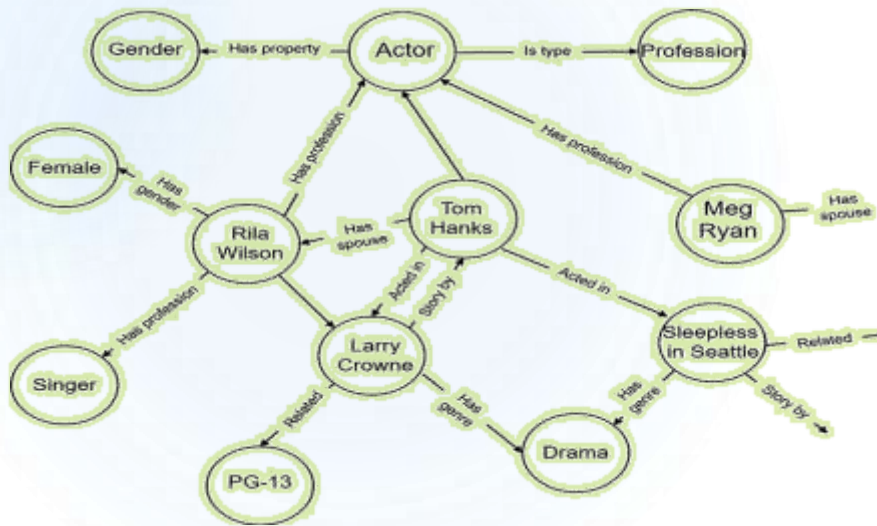
Structured data



Storage Format



Data Cleaning



Knowledge graph construction

The screenshot shows a web browser displaying a JSON response from a chatbot. The response is a JSON object with the following structure:

```
1 {
2   "best_matches": [
3     "Cancer",
4     "Tumor",
5     "Breast cancer"
6   ],
7   "intent": "description",
8   "output": "Cancer is the uncontrolled growth of abnormal cells in the body. Cancerous cells are also called malignant cells."
9 }
```

Chatbot



User Interface with chatbot

Data Scraping

- To gather the relevant data from various sources.
- It involves extracting the structured and unstructured information from trusted medical websites.
 1. Identifying data Sources
 2. Designing web scraping strategies
 3. Implementing data scraping
 4. Storing in MongoDB
- Website : <https://medlineplus.gov/healthtopics.html>

Data Cleaning

12

- It is the process of finding and removing errors.
- It ensures data consistency and quality.
 1. Understanding data set
 2. Identify redundancies and inconsistencies
 3. Data cleaning techniques
 - a) Removing irrelevant data
 - b) Remove duplicates
 - c) Remove special symbols
 - d) Remove additional spaces

Knowledge graph Construction

- It organize data into interconnected nodes representing entities and relationships.
- Nodes: diseases, symptoms, treatments, medications, etc.
- Querying the Knowledge graph
- This allows chatbots to comprehend context and provide more accurate and context-aware responses.
- Traversal: Breadth First Search

Aho-Corasick Algorithm

- It is a powerful string searching algorithm that efficiently finds all occurrences of a set of patterns (keywords) within a text.
- Key points:
 - a. Building the Automation
 - 1. Trie Construction
 - 2. Failure Construction

BERT MODEL

15

- Bidirectional Encoder Representations from Transformers, is a machine learning (ML) framework for natural language processing.
- It converts words into numbers and allows you to train machine learning models on your textual data.
- Key points:
 - a. BERT Tokenizer
 - 1. Word piece algorithm
 - b. BERT for Classification
 - 1. BERT Encoder

FLASK Frame work

16

- Flask is flexible Python web framework, plays a crucial role in creating chatbots.
- **Core Functionalities:**
 - a. Server-side logic
 - b. Communicates with the webpage to take input from user
 - c. Communicates with the webpage to display the final results
- **Decorators:**
 - 1. /
 - 2. /grammar
 - 3. /api

Work Flow

17

Front End

Designed using html, css, jquery

User types query in input field(body) ----->
Restful

Receives and extracts

↓
appended to the chat log,
effectively updating the conversation history
displayed to the user.

Backend

Designed using python, flask framework

process_query() has users query data.

↓
Extracts query data and processed.

1. Grammar Enhancement
2. Named Entity Recognition
3. Answer retrieval (KG)

↓
Converts into Json format

←-----<
Sends

This loop continues for every query.

Results

18

Scraped data in MongoDB

My Queries

Databases

Search

admin

config

local

medical

medlineplus

disease_names_url

entire_data

medlineplus.entire_data - Documents

medlineplus.entire_data

6361
DOCUMENTSINDEXES

DocumentsAggregationsSchemaIndexesValidation

FilterType a query: { field: 'value' }

ExplainResetFindOptions

ADD DATAEXPORT DATA

1 - 20 of 3555

```
_id: ObjectId('650bec8db687440565a13c0b')
name: "A1C test"
description: "A1C is a lab test that shows the average level of blood sugar (glucose..."
alternate_name: "HbA1C test; Glycated hemoglobin test; Glycohemoglobin test; Hemoglobin..."
How the Test is Performed: Array (4)
How to Prepare for the Test: Array (1)
How the Test will Feel: Array (2)
Why the Test is Performed: Array (3)
Normal Results: Array (8)
What Abnormal Results Mean: Array (12)
Risks: Array (8)
Alternative Names: Array (1)
Patient Instructions: Array (2)
Images: Array (2)
References: Array (2)
Review Date 4/29/2022: Array (1)
No Title Found: Array (empty)
Related MedlinePlus Health Topics: Array (7)
```

Cleaned data

19

```
data > disease_cleaned.json > {} 1 > [ ] symptoms
```

```
1 [
2   {
3     "id": "650bec8eb687440565a13c0d",
4     "name": "Aarskog syndrome",
5     "description": "Aarskog syndrome is a very rare disease that affects",
6     "alternate_names": [
7       "Aarskog disease",
8       "Aarskog-Scott syndrome",
9       "AAS",
10      "Faciodigitogenital syndrome",
11      "Gaciogenital dysplasia",
12      "Aarskog disease; Aarskog-Scott syndrome; AAS; Faciodigitogenital",
13    ],
14    "symptoms": [
15      "Symptoms of this condition include:",
16      "Belly button that sticks out Bulge in the groin or scrotum Delayed sexual maturity",
17      "Belly button that sticks out",
18      "Bulge in the groin or scrotum",
19      "Delayed sexual maturity",
20      "Delayed teeth",
21      "Downwardpalpebral slantto eyes ( palpebral slant is the direction of the hairline)",
22      "Hairline with a \"widow's peak\"",
23      "Mildly sunken chest ( pectus excavatum )",
24      "Mild to moderate cognitive problems",
25      "Mild to moderate short height which may not be obvious until the child is older",
26      "Poorly developed middle section of the face",
27      "Rounded face",
28      "Scrotum surrounds the penis ( shawl scrotum )",
29      "Short fingers and toes with mild webbing",
30      "Single crease in the palm of the hand",
31      "Small, broad hands and feet with short fingers and curved-in fifth finger",
32      "Small nose with nostrils tipped forward",
33      "Testicles that have not come down ( undescended )",
34      "Top portion of the ear folded over slightly",
35      "Wide groove above the upper lip, crease below the lower lip",
36      "Wide-set eyes with droopy eyelids"
37    ],
38  }
39 ]
```

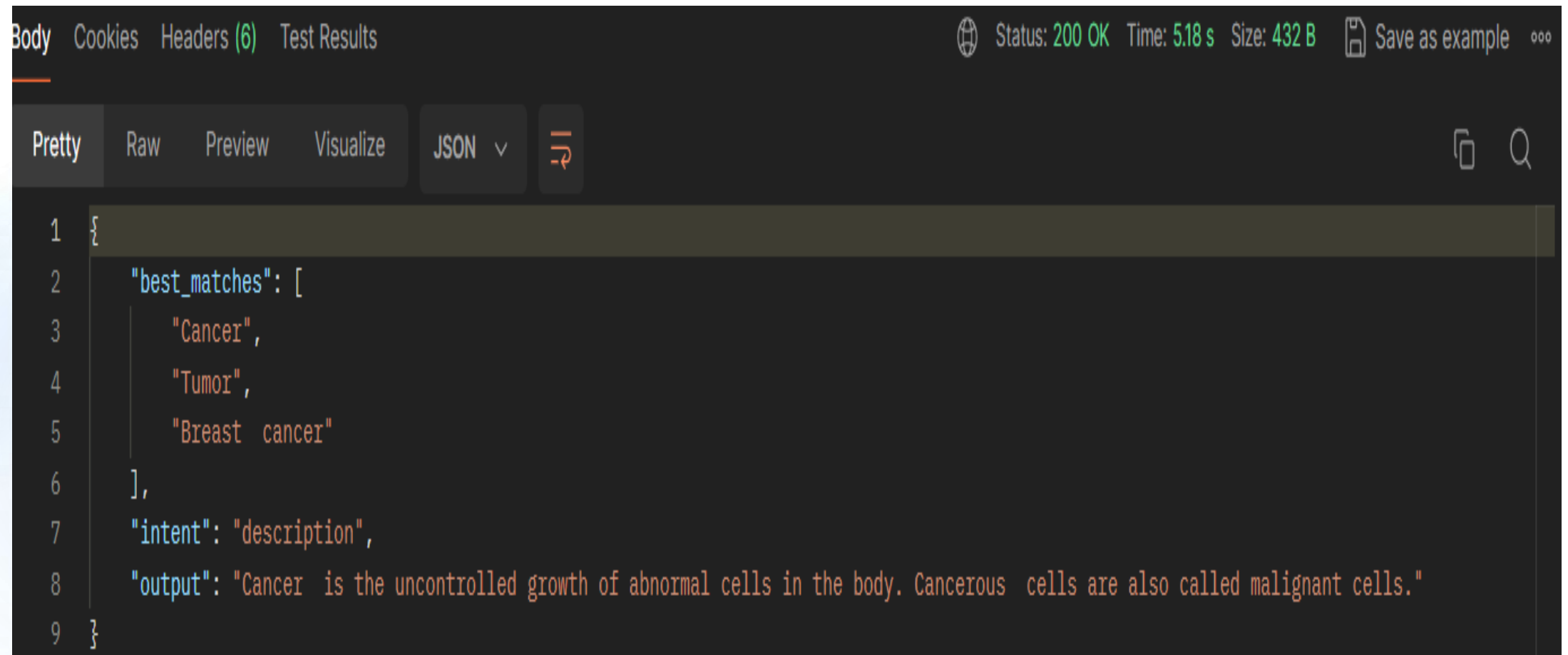
```
data > medlineplus.json > {} 1 > [ ] Support Groups
```

```
89 {
90   "_id": {
91     "$oid": "650bec8eb687440565a13c0d"
92   },
93   "name": "Aarskog syndrome",
94   "description": "Aarskog syndrome is a very rare disease that affects a person's physical appearance and health.",
95   "alternate_name": "Aarskog disease; Aarskog-Scott syndrome; AAS; Faciodigitogenital syndrome",
96   "Causes": [
97     "Aarskog syndrome is a genetic disorder that is linked to the X chromosome.",
98   ],
99   "Symptoms": [
100    "Symptoms of this condition include:",
101    "Belly button that sticks outBulge in the groin or scrotumDelayed sexual maturity",
102    "Belly button that sticks out",
103    "Bulge in the groin or scrotum",
104    "Delayed sexual maturity",
105    "Delayed teeth",
106    "Downwardpalpebral slantto eyes (palpebral slant is the direction of the hairline)",
107    "Hairline with a \"widow's peak\"",
108    "Mildly sunken chest (pectus excavatum)",
109    "Mild to moderate cognitive problems",
110    "Mild to moderate short height which may not be obvious until the child is older",
111    "Poorly developed middle section of the face",
112    "Rounded face",
113    "Scrotum surrounds the penis (shawl scrotum)",
114    "Short fingers and toes with mild webbing",
115    "Single crease in the palm of the hand",
116    "Small, broad hands and feet with short fingers and curved-in fifth finger",
117    "Small nose with nostrils tipped forward",
118    "Testicles that have not come down (undescended)",
119    "Top portion of the ear folded over slightly",
120    "Wide groove above the upper lip, crease below the lower lip",
121    "Wide-set eyes with droopy eyelids"
122  ],
123   "Exams and Tests": [
124     "These tests may be done:",
125     "Genetic testing for mutations in theFGDN1geneX-rays".
```

```
PS C:\Users\Sai Tarun\OneDrive - Amrita Vishwa Vidyapeetham\Desktop\knowledge_graph> python -u "c:\Users\Sai Tarun\OneDrive - Amrita Vishwa Vidyapeetham\Desktop\knowledge_graph\build_graph.py"
What would you like to know about?
1. id
2. name
3. description
4. alternate_names
5. symptoms
6. causes
7. exams_and_tests
8. treatment
9. possible_complications
10. images
11. related_medlineplus_health_topics
Enter the number of your choice: 3
Which disease are you referring to? ABO incompatibility
Description: A, B, AB, and O are the 4 major blood types. The types are based on small substances ( molecules ) on the surface of the blood cells.
PS C:\Users\Sai Tarun\OneDrive - Amrita Vishwa Vidyapeetham\Desktop\knowledge_graph>
```

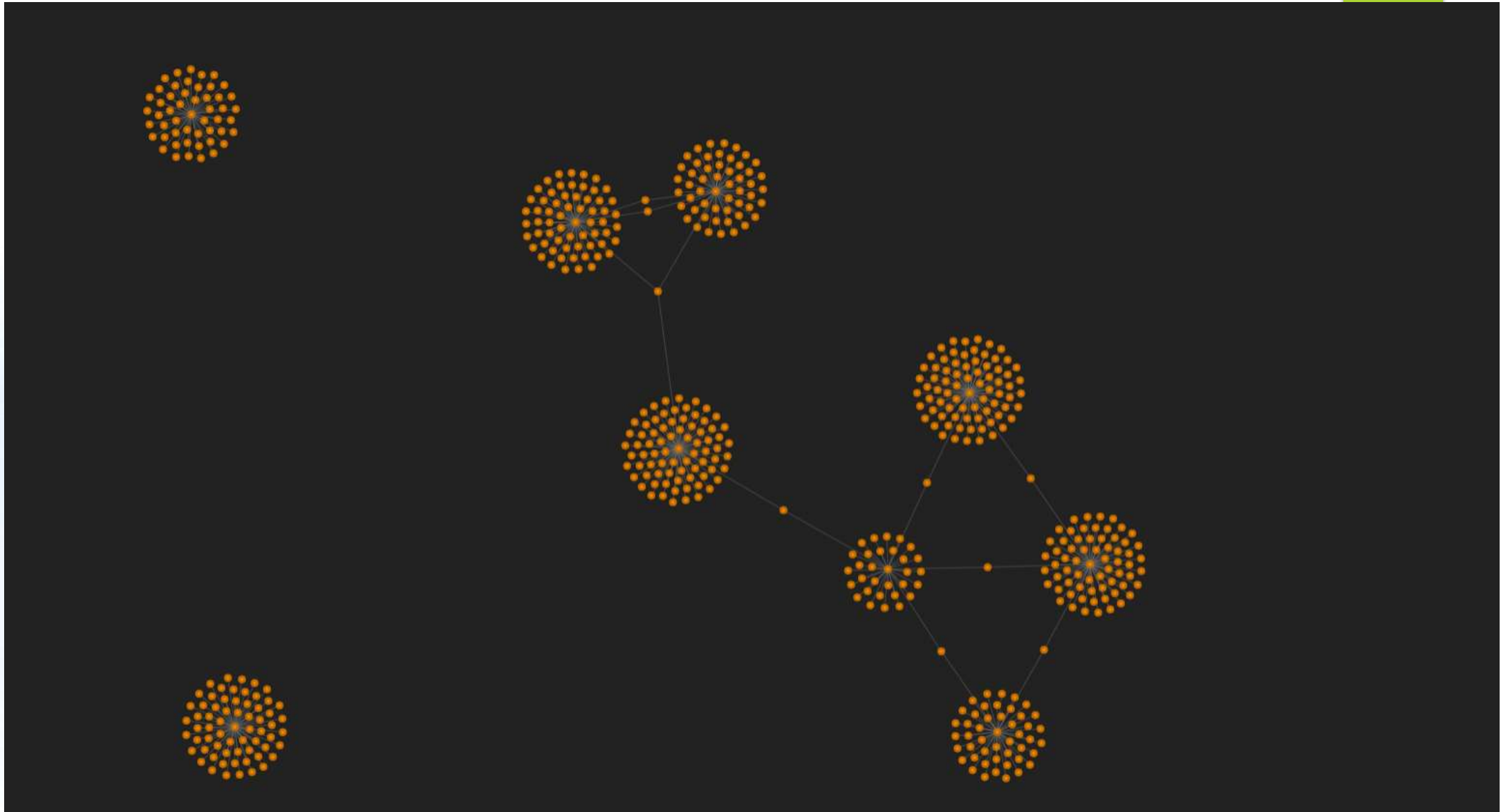
Testing

21



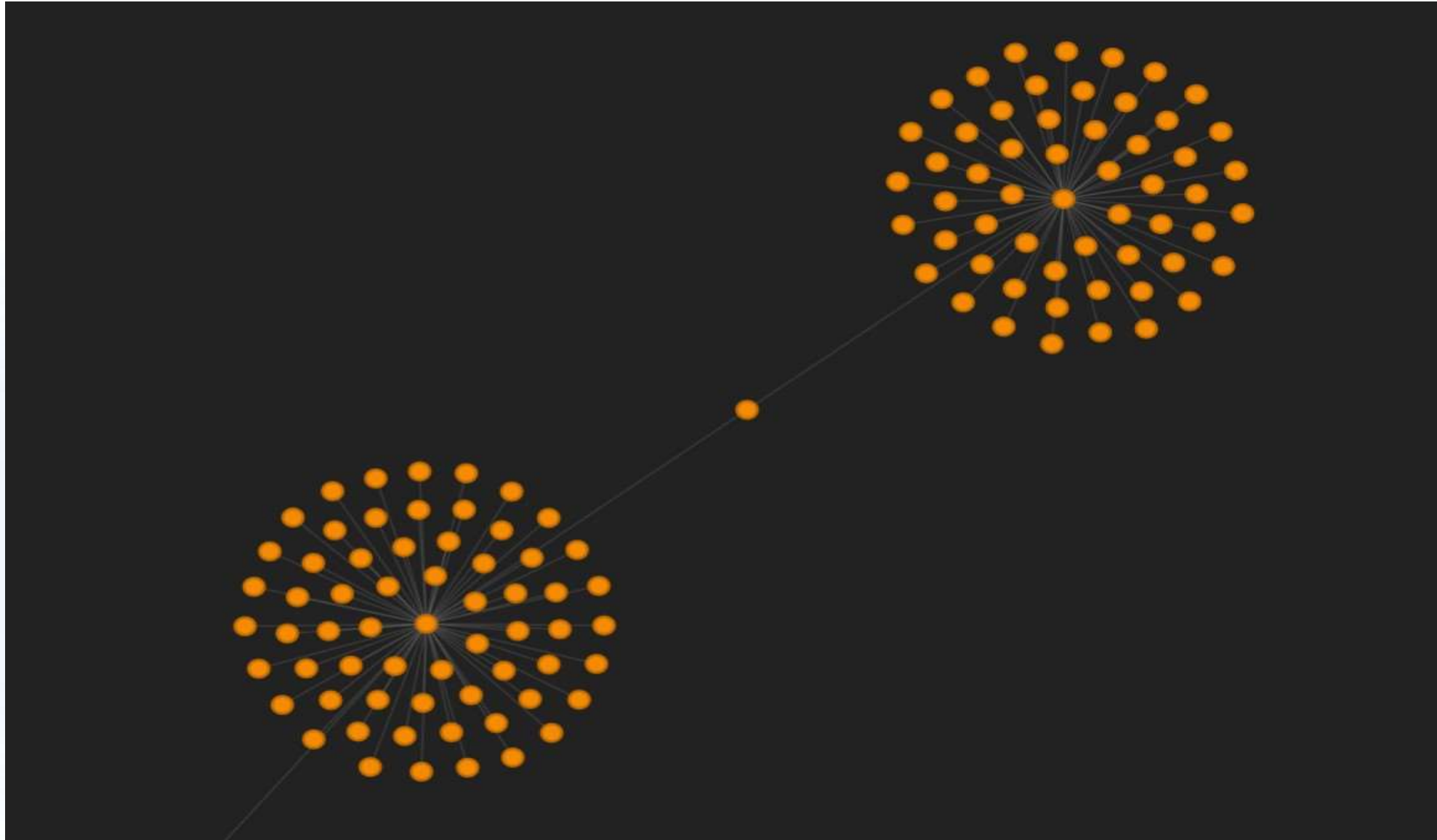
The screenshot shows a web browser's developer console with the 'Body' tab selected. The response is a JSON object with a status of 200 OK, a time of 5.18 s, and a size of 432 B. The JSON data is as follows:

```
1 {  
2   "best_matches": [  
3     "Cancer",  
4     "Tumor",  
5     "Breast cancer"  
6   ],  
7   "intent": "description",  
8   "output": "Cancer is the uncontrolled growth of abnormal cells in the body. Cancerous cells are also called malignant cells."  
9 }
```

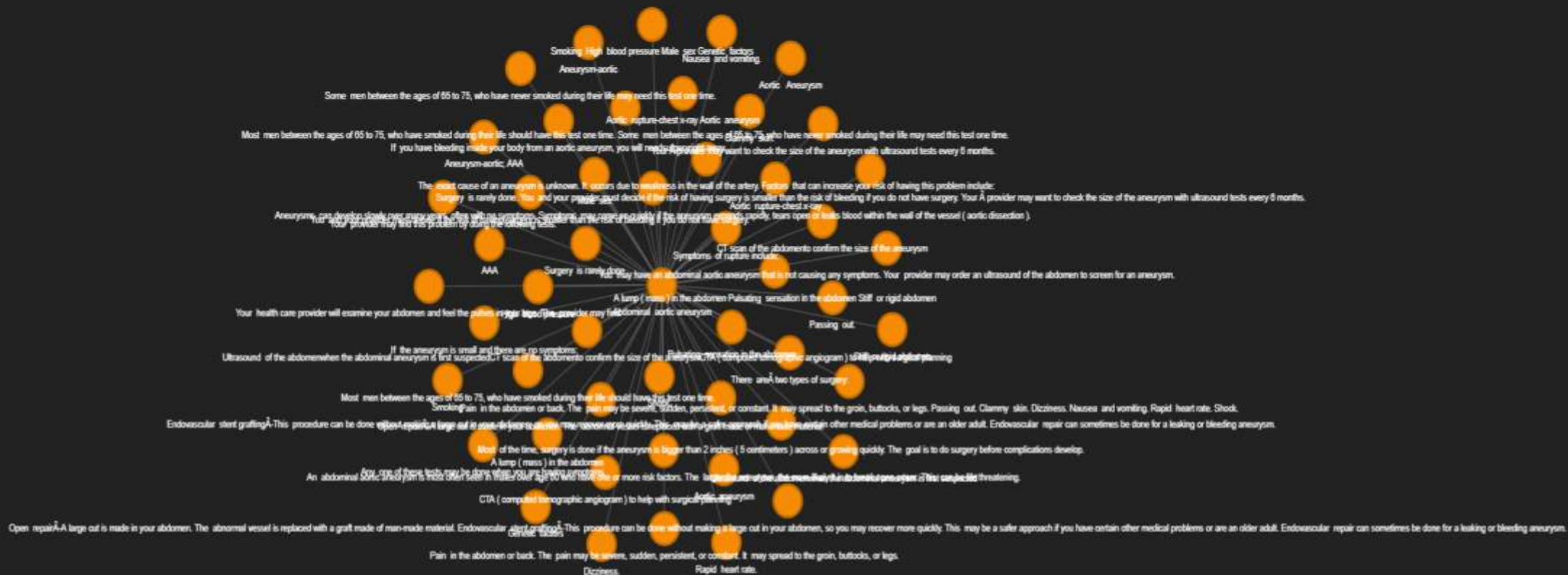



Two disease with common attribute

23

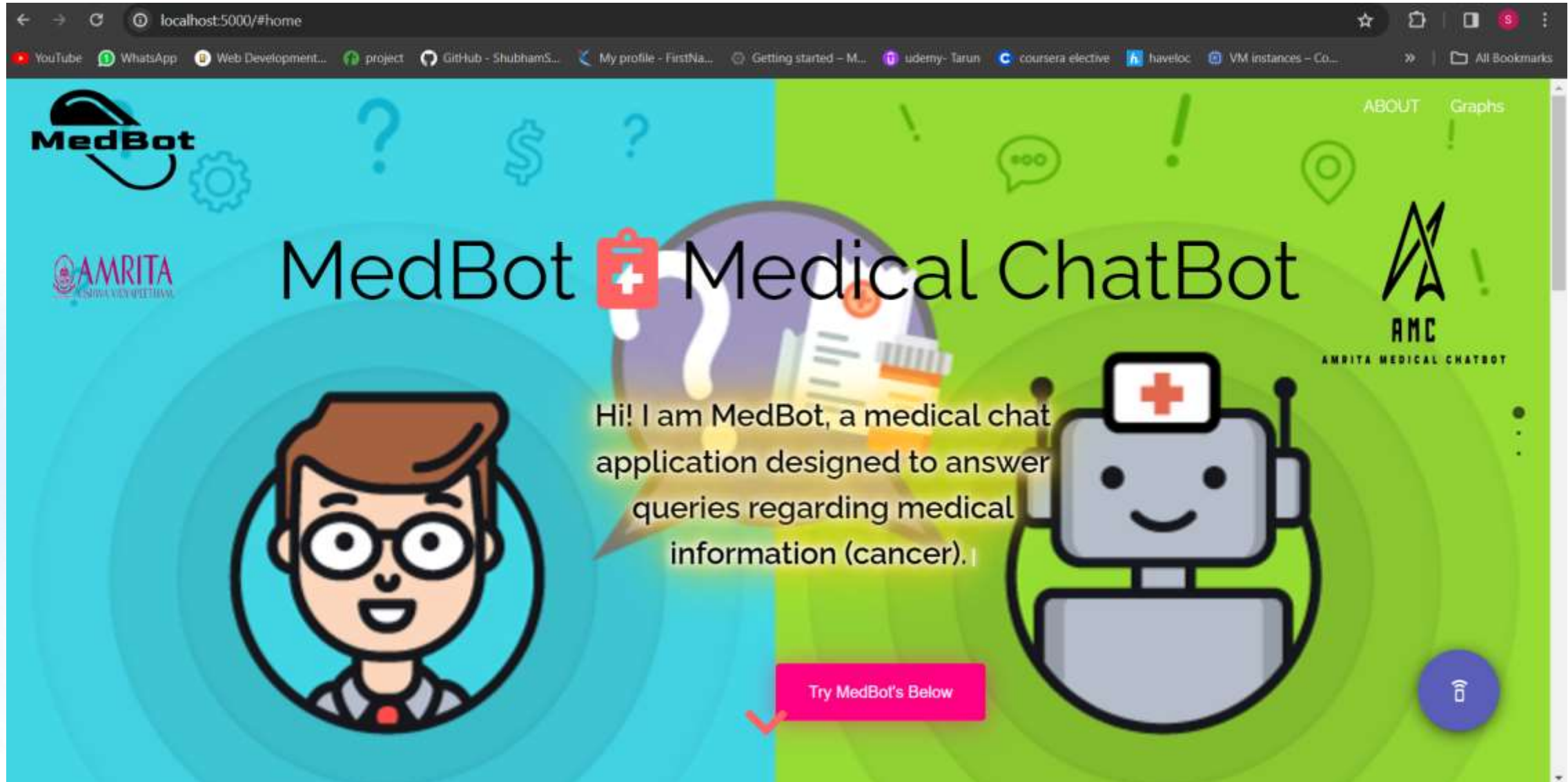


Node and Edges(single node)



Home page

25



About page

26

localhost:5000/#about


YouTube WhatsApp Web Development... project GitHub - ShubhamS... My profile - FirstNa... Getting started - M... udemy- Tarun coursera elective haveloc VM instances - Co... All Bookmarks

MedBot

ABOUT Graphs


WHAT CAN MEDBOT DO ?

Tasks Performed By MedBot




Analyse User's Questions

MedBot analyses the user's text using NLP. Grammar is enhanced for user's questions if it lacks.




Parse Questions

MedBot parses the questions and identifies the type of question being asked.



Fetching Answers

MedBot fetches answers from the medical knowledge graph based on the question category.





Knowledge Graph

MedBot is fast because of the graph structure and traversal algorithms are used to fetch through the graph.

Wi-Fi

Portfolio page

27



Diseases Graph

A single disease node connected to all of its relational details




ABOUT

Graphs

◀

▶



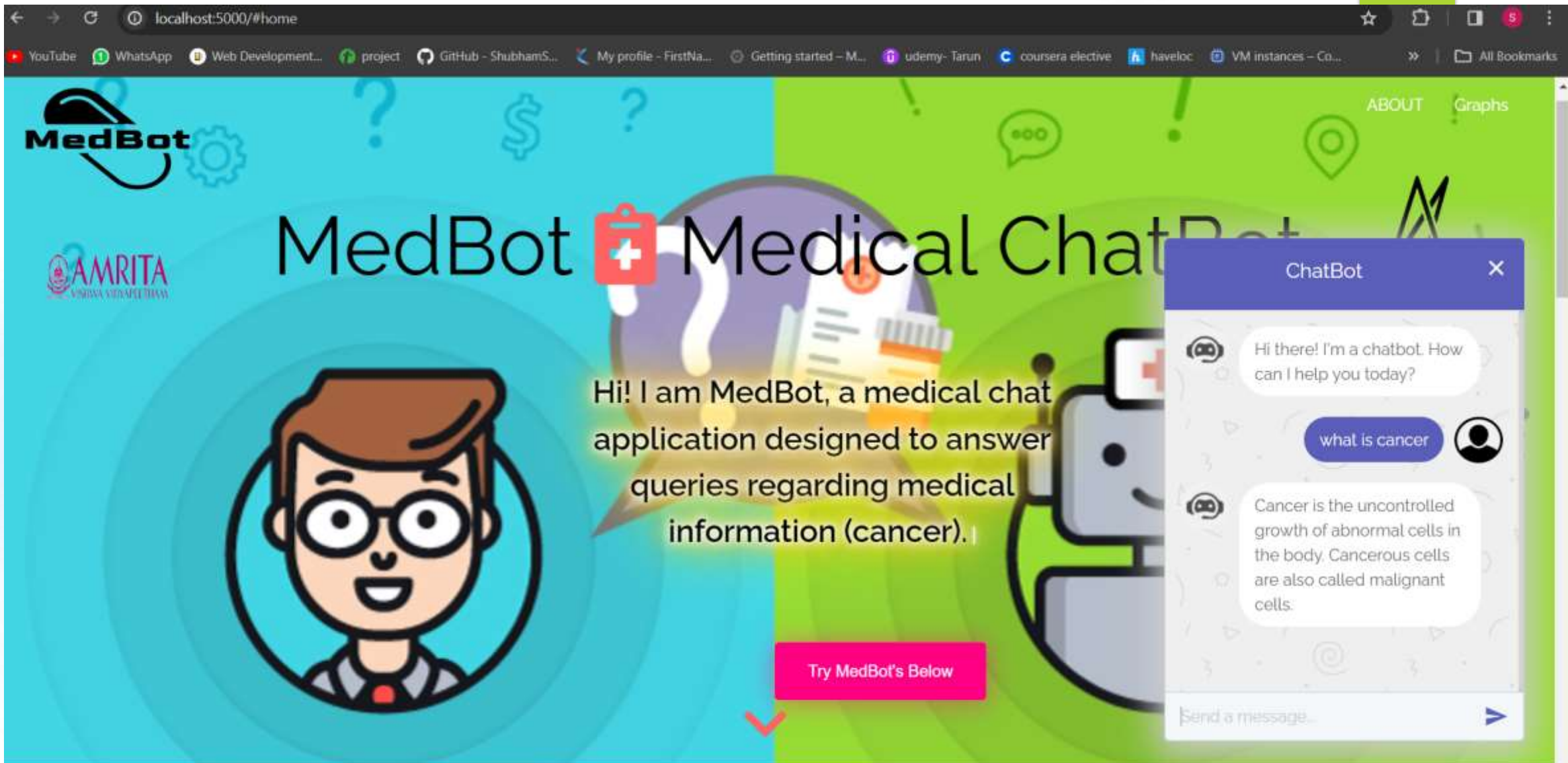
[HOME](#) [FEEDBACK](#) [MEDICAL DISCLAIMER](#) [FAQS](#) [ABOUT](#)

© 2020-2024 Computer and Communication Engineering ["CCE"]



Integration of UI and Chatbot

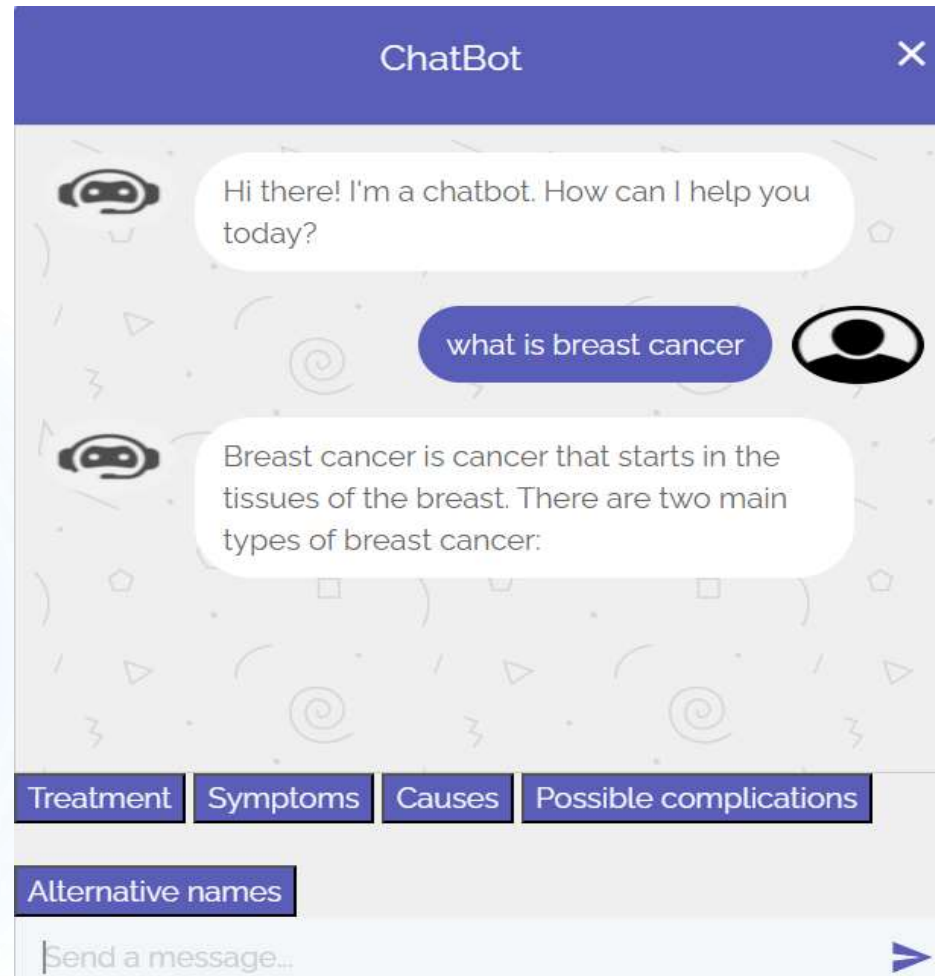
29



The screenshot displays a web browser window at the URL `localhost:5000/#home`. The browser's address bar and tabs are visible at the top. The main content area has a light blue and green background with medical-themed icons like question marks, a dollar sign, and a location pin. On the left, there is a 'MedBot' logo and a smaller logo for 'AMRITA'. The central part of the page features a large speech bubble from a cartoon doctor character that says: "Hi! I am MedBot, a medical chat application designed to answer queries regarding medical information (cancer).". Below this, there is a pink button labeled "Try MedBot's Below". On the right side, a chatbot window titled "ChatBot" is open, showing a conversation. The chatbot's first message is "Hi there! I'm a chatbot. How can I help you today?". The user has typed "what is cancer", and the chatbot has responded with "Cancer is the uncontrolled growth of abnormal cells in the body. Cancerous cells are also called malignant cells." The chat window includes a text input field at the bottom labeled "Send a message..." and a blue send button.

Chat bot

30



What are its symptoms?



Early breast cancer often does not cause symptoms. This is why regular breast exams and mammograms are important, so cancers that don't have symptoms may be found earlier. As the cancer grows, symptoms may include:

- Breast lump or lump in the armpit that is hard, has uneven edges, and usually does not hurt.
- Change in the size, shape, or feel of the breast or nipple.
- For example, you may have redness, dimpling, or puckering that looks like the skin of an orange.
- Fluid from the

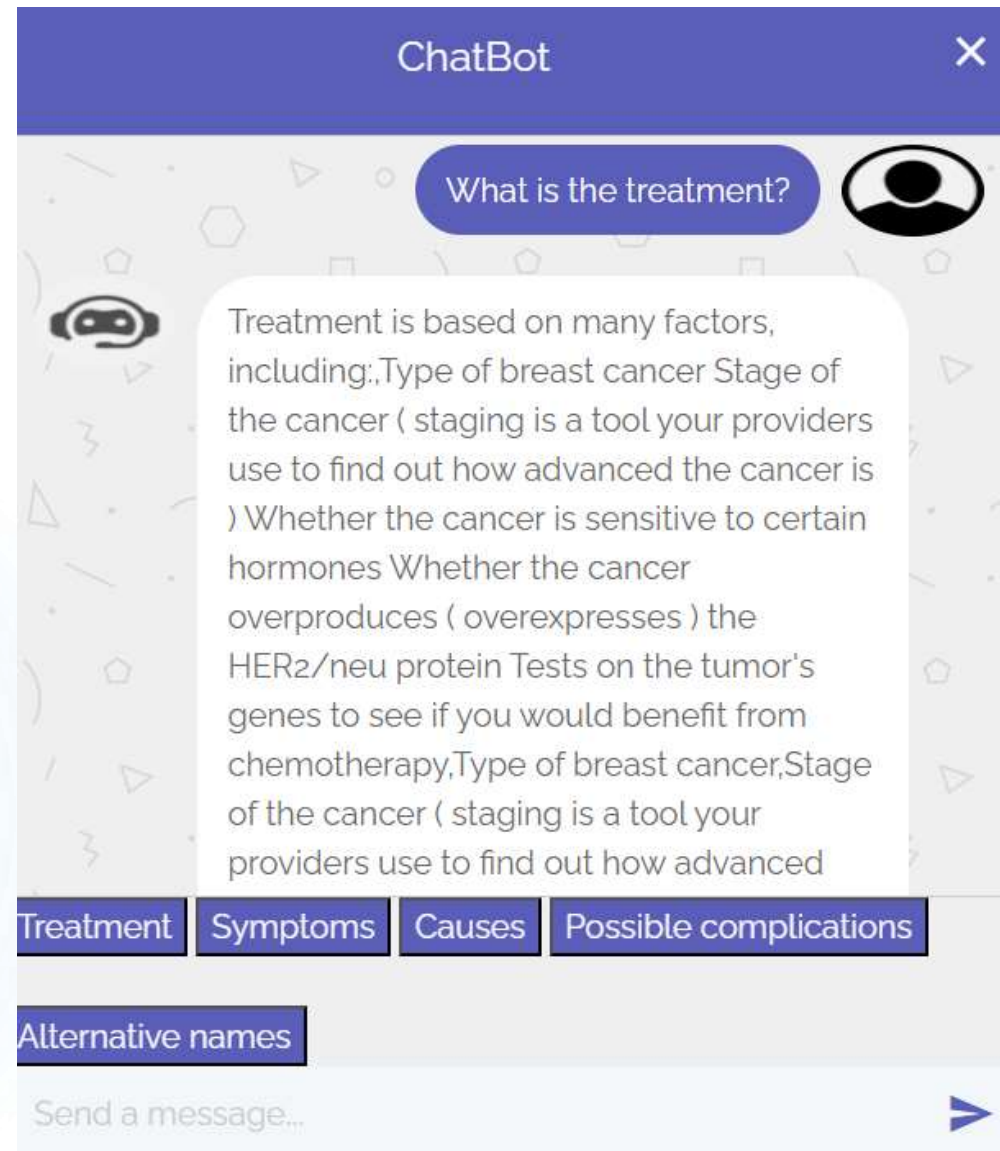
Treatment

Symptoms

Causes

Possible complications

Alternative names



Performance Metrics of intention finding

Accuracy	Precision	Recall
0.91	0.90	0.90

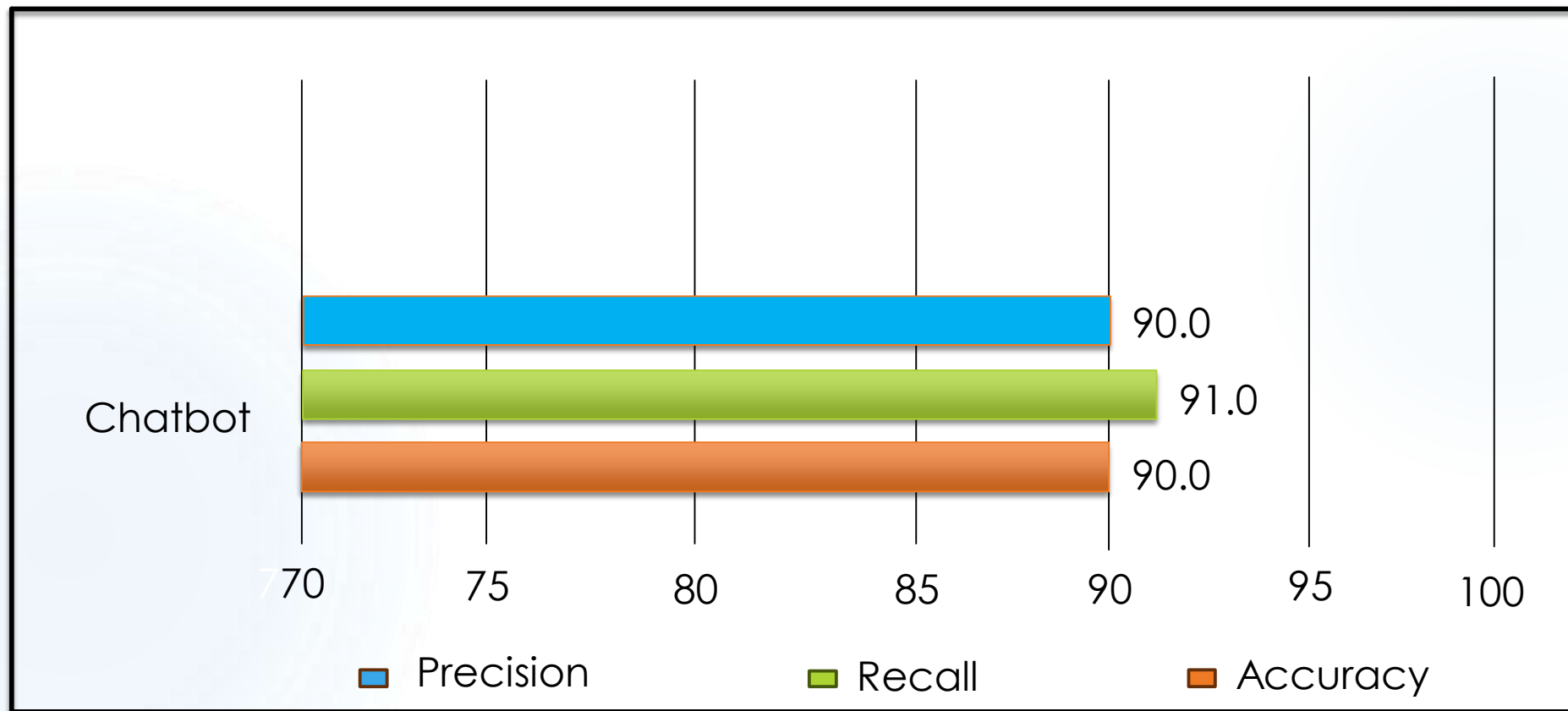
Confusion matrix:

45	5
4	46

Performance of Knowledge graph:

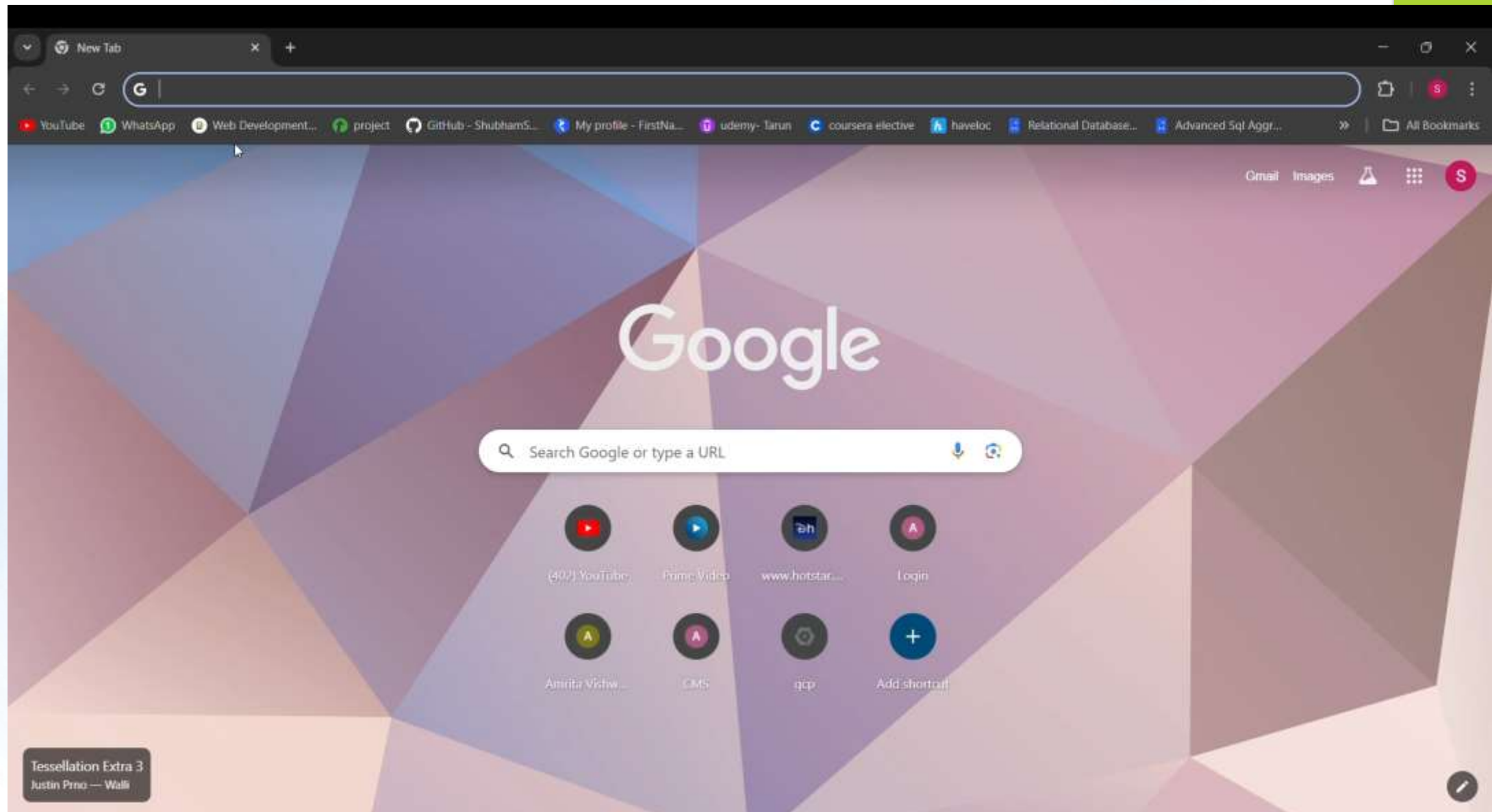
1. Average Time taken for 100 queries : 0.0053sec

Performance Metrics of Amrita Medical Chatbot (AMC)



Live Demo

35



Bibliography

36

- [1]. E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," in IEEE Access, vol. 8, pp. 61726-61740, 2020, doi: 10.1109/ACCESS.2020.2984503.
- [2]. H. Lan et al., "COVID-Scraper: An Open-Source Toolset for Automatically Scraping and Processing Global Multi-Scale Spatiotemporal COVID-19 Records," in IEEE Access, vol. 9, pp. 84783-84798, 2021, doi: 10.1109/ACCESS.2021.3085682.
- [3]. S. Ji, S. Pan, E. Cambria, P. Marttinen and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 2, pp. 494-514, Feb. 2022, doi: 10.1109/TNNLS.2021.3070843.
- [4]. Jiang, Zhixue, Chengying Chi, and Yunyun Zhan. "Research on medical question answering system based on knowledge graph." IEEE Access 9 (2021): 21094-21101
- [5]. Lino Murali, G. Gopakumar, Daleesha M. Viswanathan, Prema Nedungadi, Towards electronic health record-based medical knowledge graph construction, completion, and applications: A literature study, Journal of Biomedical Informatics, Volume 143, 2023, 104403, ISSN 1532-0464
- [6]. Y. Lan, G. He, J. Jiang, J. Jiang, W. X. Zhao and J. -R. Wen, "Complex Knowledge Base Question Answering: A Survey," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11196-11215, 1 Nov. 2023, doi: 10.1109/TKDE.2022.3223858.

Bibliography

37

- [7]. D. N. and N. P. K. S., "Design and Development of We-CDSS Using Django Framework: Conducting Predictive and Prescriptive Analytics for Coronary Artery Disease," in *IEEE Access*, vol. 10, pp. 119575-119592, 2022, doi: 10.1109/ACCESS.2022.3220899.
- [8]. S. Aghaei, E. Raad and A. Fensel, "Question Answering Over Knowledge Graphs: A Case Study in Tourism," in *IEEE Access*, vol. 10, pp. 69788-69801, 2022, doi: 10.1109/ACCESS.2022.3187178.
- [9]. X. Wu, T. Jiang, Y. Zhu and C. Bu, "Knowledge Graph for China's Genealogy11.A shorter version of this paper won the Best Paper Award at IEEE ICKG 2020 (the 11th IEEE International Conference on Knowledge Graph, ickg 2020.bigke.org).," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 634-646, 1 Jan. 2023, doi: 10.1109/TKDE.2021.3073745.
- [10]. Fakhare Alam, Hamed Babaei Giglou, Khalid Mahmood Malik, Automated clinical knowledge graph generation framework for evidence based medicine, *Expert Systems with Applications*, Volume 233, 2023, 120964, ISSN 0957-4174.

THANK YOU