ILLINOIS INSTITUTE
OF TECHNOLOGY

PROJECT ON INFORMATION RETRIVAL (CSP – 429)

HealthSphere Chatbot

By
Uday Kumar swamy(A20526852)

Under The Guidance of

Mr. Jawahar Panchal
Professor (CSP-429)
Department of Computer Science

ABSTRACT

HealthSphere Chatbot is a specialized tool tailored to assist users seeking information about skin and bone-related diseases. Leveraging a combination of technologies and techniques, the chatbot offers an intuitive and informative experience.

The backbone of HealthSphere Chatbot lies in its ability to understand and process natural language. NLTK (Natural Language Toolkit) is used for text processing, enabling the chatbot to interpret user queries and extract relevant information. This allows users to interact with the chatbot in a conversational manner, making the experience more engaging and user-friendly.

To further enhance its capabilities, the chatbot employs machine learning algorithms from Scikit-learn. These algorithms help the chatbot in tasks such as text classification, enabling it to categorize user queries and provide more accurate responses. For example, the chatbot can classify a user query about symptoms into relevant categories such as skin conditions or bone diseases, allowing it to provide targeted information.

The backend of HealthSphere Chatbot is built using Flask, a lightweight web framework for Python. Flask allows the chatbot to handle user requests and generate responses dynamically, ensuring a seamless user experience. Additionally, Flask provides a scalable and maintainable architecture, making it easier to add new features and improve existing ones.

To retrieve information about skin and bone-related diseases, the chatbot utilizes web scraping techniques with Beautifulsoup and Requests. These tools allow the chatbot to extract data from the NIAMS website, ensuring that the information provided is up-to-date and accurate.

To improve the relevance of its responses, the chatbot uses cosine similarity and TF-IDF. Cosine similarity is used to compare the similarity between user queries and existing data, allowing the chatbot to identify relevant information. TF-IDF is used to determine the importance of words in a document relative to a collection of documents, helping the chatbot to rank the relevance of its responses.

OVERVIEW

HealthSphere Chatbot is a project aimed at providing a conversational interface for users seeking information about skin and bone-related diseases. The chatbot utilizes a combination of natural language processing (NLP), machine learning, and web scraping techniques to provide accurate and relevant information to users. Queries are made specific to the PDFs crawled from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) website to ensure the information provided is tailored to the content of these documents.

Solution Outline:

1. **Natural Language Processing (NLP):**
    - NLTK (Natural Language Toolkit) is used for text processing, enabling the chatbot to understand and interpret user queries.
    - Techniques such as tokenization, stemming, and lemmatization are employed to process the text and extract relevant information.
2. **Machine Learning (ML):**
    - Scikit-learn is used for machine learning tasks, such as text classification.
3. **Web Scraping:**
    - BeautifulSoup and Requests are used for web scraping to extract information about skin and bone-related diseases from the NIAMS website.
    - Web scraping ensures that the information provided by the chatbot is up-to-date and accurate.
4. **Cosine Similarity and TF-IDF:**
    - Cosine similarity is used to compare the similarity between user queries and existing data.
    - TF-IDF is used to determine the importance of words in a document relative to a collection of documents, helping the chatbot to rank the relevance of its responses.
5. **Specific Query to PDFs:**
    - Queries are made specific to the content of the PDFs crawled from the NIAMS website.
    - PDF parsing libraries are used to extract text from the PDFs, which is then indexed for efficient searching.
    - Search functionality is implemented using a search engine library to retrieve relevant PDFs based on user queries.

Relevant Literature:

- The project is based on existing research and literature in the fields of NLP, machine learning, and web scraping.
- Relevant studies and papers on chatbot development, NLP techniques, and healthcare information retrieval are consulted to inform the design and implementation of the chatbot.

Proposed System:

- The proposed system is a chatbot interface that allows users to interact with the chatbot using natural language queries.
- The system utilizes NLP techniques to process user queries and extract relevant information.
- Machine learning algorithms are used to categorize user queries and provide accurate responses.
- Web scraping is used to retrieve information about skin and bone-related diseases from the NIAMS website.
- Cosine similarity and TF-IDF are used to enhance the relevance of the chatbot's responses.
- Queries are specific to the content of the PDFs crawled from the NIAMS website, ensuring that users receive information tailored to the documents available.
- Flask, a lightweight web framework, is used to handle user requests and responses, making the chatbot's interactions smooth and user-friendly.

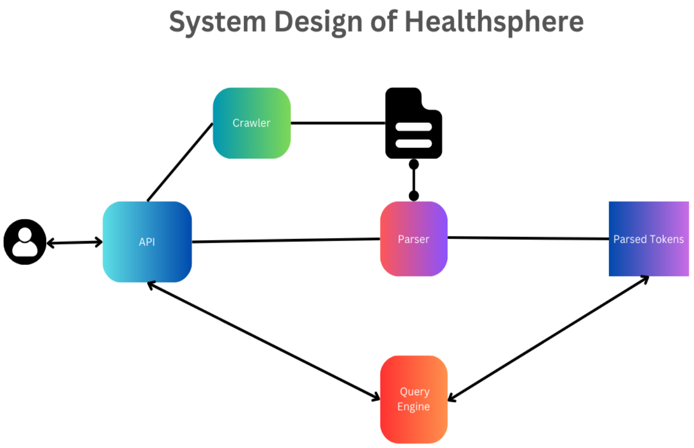
User Feature Enhancements

The HealthSphere Chatbot can be further enhanced in the future by integrating with large language models like Open Ai's GPT-3 or LLAMA-2. These advanced language models have the capability to understand and generate human-like text, which can significantly improve the chatbot's query response quality and versatility.

Potential Enhancements:

1. **Natural Language Understanding**: Integration with large language models can enhance the chatbot's ability to understand natural language queries with greater accuracy and context awareness.
2. **Improved Responses**: By leveraging advanced language models, the chatbot can generate more informative and contextually relevant responses to user queries, leading to a more engaging user experience.
3. **Expanded Knowledge Base**: Large language models have access to vast amounts of information from various sources. Integrating with these models can enable the chatbot to provide answers to a wider range of queries and offer more comprehensive information on skin and bone-related diseases.
4. **Personalized Interaction**: Advanced language models can analyse user preferences and behaviour to personalize interactions and tailor responses to individual users' needs and interests.

DESIGN & ARCHITECTURE



Crawler Components Description

This crawler is designed to retrieve disease information data from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) website. It utilizes the Python libraries requests and Beautiful Soup (bs4) for this purpose.

Components

1. **Requests Module**: The requests module is used to send HTTP requests to the NIAMS website and retrieve the webpage content.
2. **Beautiful Soup (bs4)**: The Beautiful Soup library is employed for parsing the HTML content of the webpage and extracting relevant information.

Functionality

1. **Fetching Disease Information**: The crawler navigates to the NIAMS website's page containing information on various diseases related to arthritis, musculoskeletal, and skin conditions.
2. **Scraping Data**: Using Beautiful Soup, the crawler parses the HTML content of the webpage and extracts relevant information about different diseases, including their names and associated links to PDF documents.
3. **Saving PDF Documents**: The crawler identifies the links to PDF documents for each disease and downloads them into a local directory for further processing.
4. **Error Handling**: The crawler includes error handling mechanisms to handle exceptions gracefully, ensuring smooth execution even in the presence of unexpected issues such as network errors or malformed HTML content.

Parser Component Description

The parser component utilizes the Natural Language Toolkit (NLTK) to parse downloaded documents. It also utilizes the Fitz module for PDF parsing capabilities and includes TfidfVectorizer and cosine similarity for text analysis.

Functionality

1. **Parsing Downloaded Documents**: The parser component processes the downloaded PDF documents using the Fitz module, which provides functionalities for parsing PDF content.
2. **Tokenization**: After parsing the documents, NLTK's tokenization techniques are employed to break down the text into individual tokens. Tokenization involves splitting the text into words, sentences, or other meaningful units.
3. **Vectorization with TfidfVectorizer**: The parsed text data is then vectorized using TfidfVectorizer, which converts text data into numerical vectors based on the term frequency-inverse document frequency (TF-IDF) representation. This step is crucial for preparing the text data for further analysis.
4. **Similarity Calculation with Cosine Similarity**: Once the text data is vectorized, cosine_similarity is used to calculate the similarity between text documents. Cosine similarity measures the cosine of the angle between two vectors and is commonly used to determine the similarity between text documents.
5. **Natural Language Processing**: NLTK provides various tools and methods for natural language processing tasks, such as lemmatization, stemming, and part-of-speech tagging. These capabilities can be further utilized to enhance the parsing and analysis of the text data.

Query Engine Component Description

The Query Engine component is responsible for processing user queries and generating responses based on the tokens provided by the user. It employs techniques for text pre-processing, tokenization, and similarity matching to identify the most relevant responses.

Functionality

1. **Text Pre-processing**: The Query Engine pre-processes user queries by removing punctuation marks and applying other text normalization techniques to ensure consistency in token representation.
2. **Tokenization**: After pre-processing, the user query is tokenized into individual tokens, typically words or phrases, using tokenization techniques provided by libraries like NLTK.
3. **Similarity Matching**: The Query Engine matches the user tokens with precomputed tokens from the dataset or corpus. It calculates the similarity between the user tokens and the dataset tokens using techniques like cosine similarity.
4. **Response Generation**: Based on the similarity scores, the Query Engine selects the tokens from the dataset that are most similar to the user tokens. It then

generates a response by selecting or aggregating information associated with the most similar tokens.

5. **Ranking**: The Query Engine ranks the responses based on the similarity scores, ensuring that the most relevant and highly ranked responses are presented to the user.

Data Flow:

1. **Scraping Data:**
   - User triggers scraping by clicking the "Scrape" button in the user interface.
   - The scraper component visits the website specified in the source URL configured in the **config.properties** file.
   - Using web scraping techniques, the scraper parses the HTML structure of the website.
   - It identifies and retrieves the desired PDF files available on the website.
2. **Parsing PDFs:**
   - After scraping, the user initiates parsing by clicking the "Parse" button.
   - The parsing component processes the downloaded PDFs using libraries such as **fitz**, **nltk**, and **sklearn**.
   - **fitz** is utilized for extracting text and metadata from PDF documents.
   - **nltk** is employed for natural language processing tasks such as tokenization and lemmatization.
   - **sklearn** may be used for tasks like text vectorization and similarity calculation.
   - The parsed data is stored or made available for querying.
3. **Querying Parsed Data:**
   - Once the processing of PDFs is complete, users can query the PDFs to retrieve relevant information.
   - Users interact with the query engine through the user interface.
   - The query engine analyses user queries and retrieves relevant information from the parsed PDFs.
   - Responses containing relevant information are displayed to the user in the user interface.

OPERATION

Steps to Run the Code
Follow these steps to run the HealthSphere Chatbot code:

Step 1: Navigate to the HealthSphere Folder
Go to the "healthsphere" folder located in the Step 1 location.

Step 2: Run the Flask Application
Open a terminal or command prompt, navigate to the "healthsphere" folder, and run the following command:
python app.py

Step 3: Access the Web Application
Open your web browser and enter the following URL in the address bar:
http://127.0.0.1:8080/

Step 4: Crawl Data
Once the web application is open in your browser, click on the "Crawl Data" button.

Step 5: Parse Data
After crawling the data, click on the "Parse Data" button to parse the crawled data.

Step 6: Query the Engine
You can now type your message and press Enter in the chat box to query the HealthSphere chatbot engine.

Examples of Commands and Responses

- **User Query:** Who are you?
    - **Response:** "I am a Healthsphere chatbot developed by Uday Kumar Swamy under the guidance of Professor Jawahar Panchal. I am built on a limited training dataset of skin and bone related diseases from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) (https://www.niams.nih.gov/health-topics/all-diseases). "
- **User Query:** list of diseases?
    - **Response:** "Acne, Sjogren's Syndrome, Scleroderma, Hip Replacement Surgery, Osteogenesis Imperfecta, Vitiligo, (and so on...)"
- **User Query:** Acne?
    - **Response:** "Living With Acne If you have acne, the following recommendations may help you in taking care of your skin. Researchers believe that you may be more likely to get acne if your parents had acne. Who Treats Acne? Who Gets Acne? People of all races and ages get acne, but it is most common in teens and young adults. When acne appears during the teenage years, it is more common in males. Acne can continue into adulthood, and when it does, it is more common in women. Types of Acne Acne causes several types of lesions, or pimples. Doctors usually

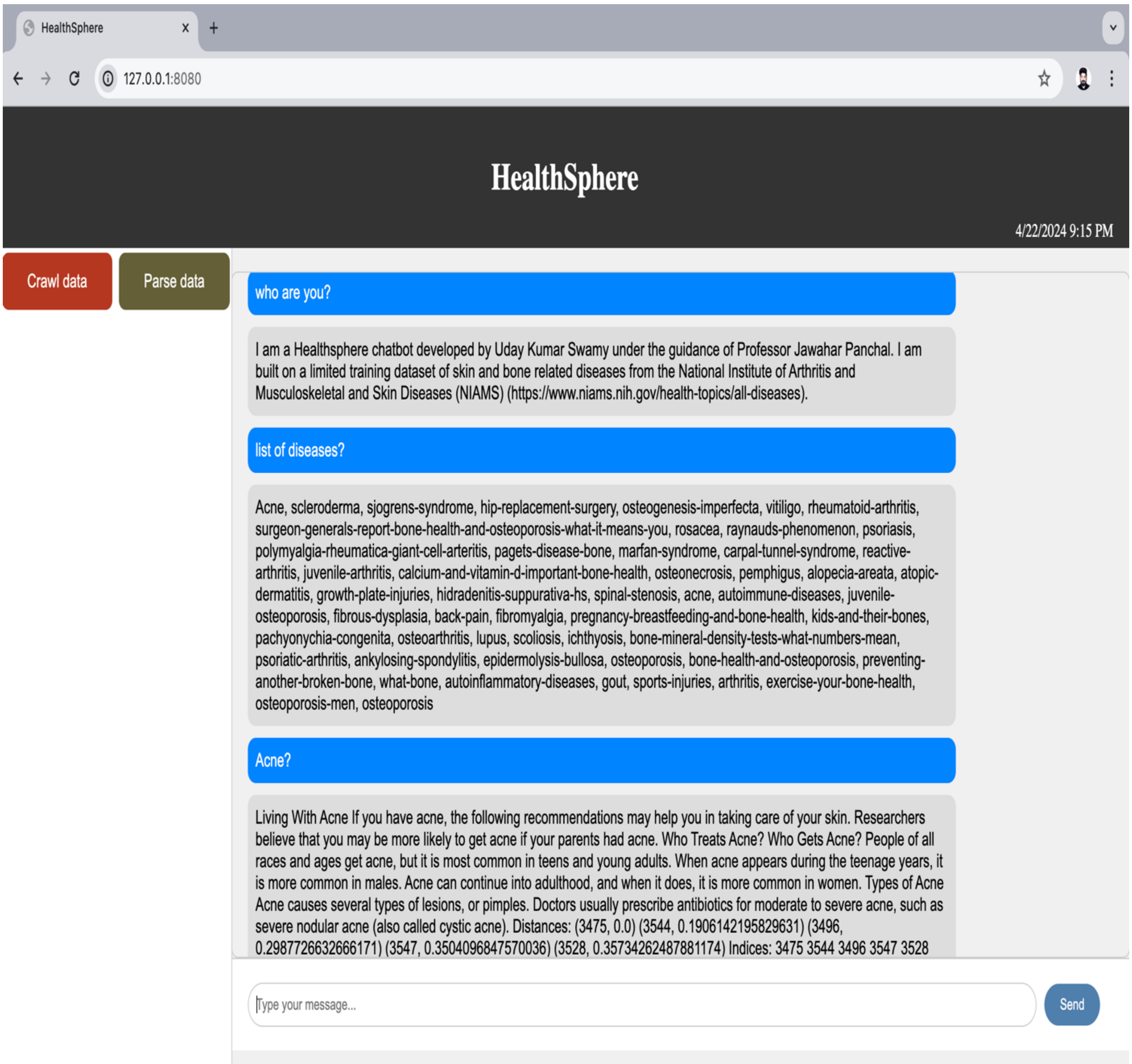
prescribe antibiotics for moderate to severe acne, such as severe nodular acne (also called cystic acne). Distances: (3475, 0.0) (3544, 0.1906142195829631) (3496, 0.2987726632666171) (3547, 0.3504096847570036) (3528, 0.35734262487881174) Indices: 3475 3544 3496 3547 3528”

- **User Query:** Who treats acne?
  - **Response:** "Living With Acne If you have acne, the following recommendations may help you in taking care of your skin. Researchers believe that you may be more likely to get acne if your parents had acne. Who Gets Acne? Who Treats Acne? The following health care providers may diagnose and treat acne: Dermatologists, who specialize in conditions of the skin, hair, and nails. Primary health care providers, including family doctors, internists, or pediatricians. Clean your skin gently. Doctors usually prescribe antibiotics for moderate to severe acne, such as severe nodular acne (also called cystic acne). Distances: (3544, 0.0) (3475, 0.19035778066569053) (3496, 0.4322567428292967) (3547, 0.47406425550858666) (3528, 0.47967745653531935) Indices: 3544 3475 3496 3547 3528”
- **User Query:** calcium?
  - **Response:** "A good way to get enough calcium is to eat foods that have calcium. Calcium is a mineral. How much calcium do we need? If we do not get enough calcium in our diets, our body takes the calcium we need from our bones. Over time, this process makes bones weaker and raises the risk of osteoporosis. Osteoporosis is a disease that makes bones weak and brittle, and people with osteoporosis have a higher risk of fractures (broken bones). The amount of calcium each person needs depends on their age and sex. Also, our bones are partly made of calcium salts (mixtures of calcium with other minerals), especially calcium phosphate. Distances: (2544, 0.26251747584106333) (2547, 0.2770751368052815) (2534, 0.347562678176245) (5566, 0.3602630321447623) (2536, 0.375089929366364) Indices: 2544 2547 2534 5566 2536"

CONCLUSION

In conclusion, HealthSphere Chatbot is a valuable tool for accessing information about skin and bone-related diseases. Through its use of NLP, machine learning, and web scraping, the chatbot provides accurate and tailored responses to user queries. Moving forward, improvements can be made by expanding the dataset of crawled PDFs and incorporating more advanced machine learning algorithms. Overall, HealthSphere Chatbot demonstrates the potential of technology in providing accessible healthcare information.

OUTPUTS:



who are you?

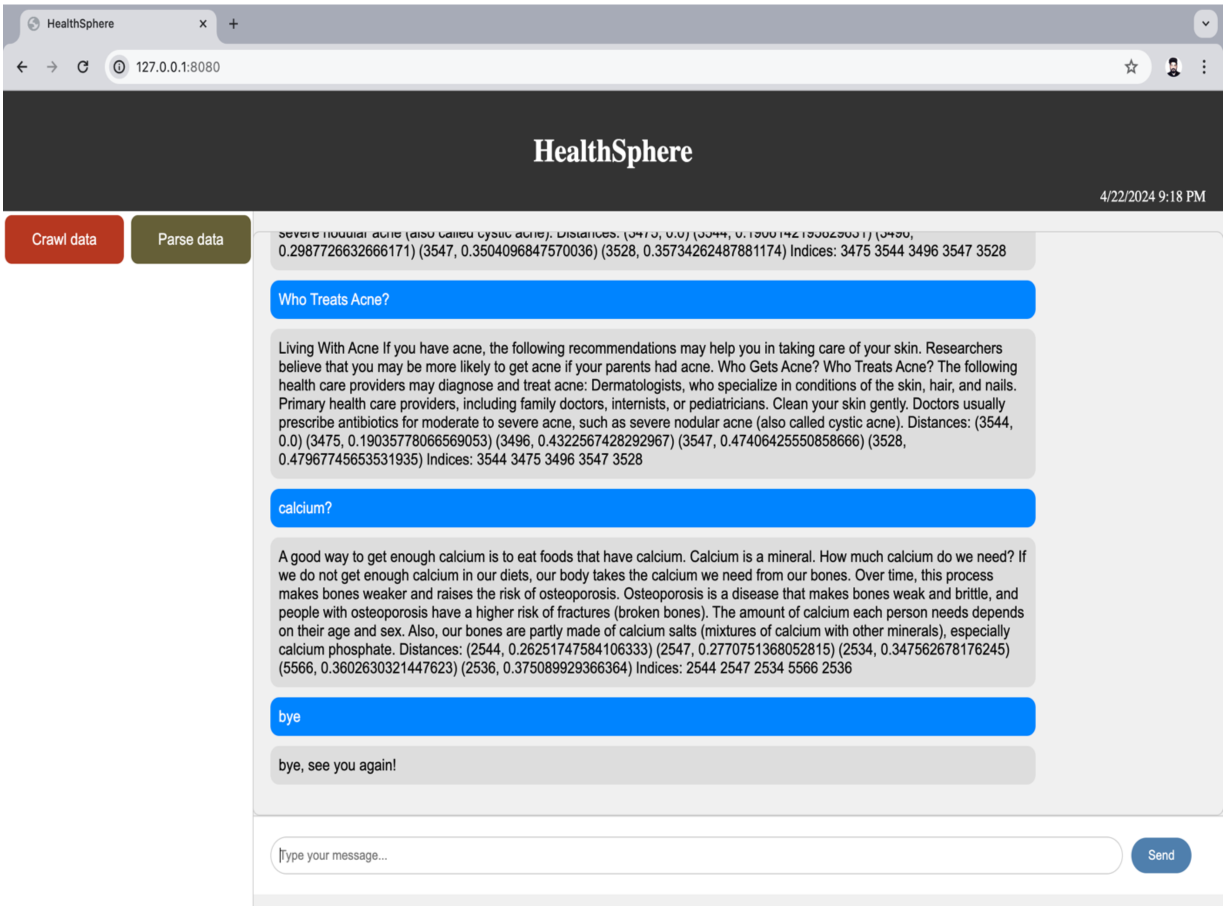
I am a Healthsphere chatbot developed by Uday Kumar Swamy under the guidance of Professor Jawahar Panchal. I am built on a limited training dataset of skin and bone related diseases from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) (https://www.niams.nih.gov/health-topics/all-diseases).

list of diseases?

Acne, scleroderma, sjogrens-syndrome, hip-replacement-surgery, osteogenesis-imperfecta, vitiligo, rheumatoid-arthritis, surgeon-generals-report-bone-health-and-osteoporosis-what-it-means-you, rosacea, raynauds-phenomenon, psoriasis, polymyalgia-rheumatica-giant-cell-arteritis, pagets-disease-bone, marfan-syndrome, carpal-tunnel-syndrome, reactive-arthritis, juvenile-arthritis, calcium-and-vitamin-d-important-bone-health, osteonecrosis, pemphigus, alopecia-areata, atopic-dermatitis, growth-plate-injuries, hidradenitis-suppurativa-hs, spinal-stenosis, acne, autoimmune-diseases, juvenile-osteoporosis, fibrous-dysplasia, back-pain, fibromyalgia, pregnancy-breastfeeding-and-bone-health, kids-and-their-bones, pachyonychia-congenita, osteoarthritis, lupus, scoliosis, ichthyosis, bone-mineral-density-tests-what-numbers-mean, psoriatic-arthritis, ankylosing-spondylitis, epidermolysis-bullosa, osteoporosis, bone-health-and-osteoporosis, preventing-another-broken-bone, what-bone, autoinflammatory-diseases, gout, sports-injuries, arthritis, exercise-your-bone-health, osteoporosis-men, osteoporosis

Acne?

Living With Acne If you have acne, the following recommendations may help you in taking care of your skin. Researchers believe that you may be more likely to get acne if your parents had acne. Who Treats Acne? Who Gets Acne? People of all races and ages get acne, but it is most common in teens and young adults. When acne appears during the teenage years, it is more common in males. Acne can continue into adulthood, and when it does, it is more common in women. Types of Acne Acne causes several types of lesions, or pimples. Doctors usually prescribe antibiotics for moderate to severe acne, such as severe nodular acne (also called cystic acne). Distances: (3475, 0.0) (3544, 0.1906142195829631) (3496, 0.2987726632666171) (3547, 0.3504096847570036) (3528, 0.3573426248788174) Indices: 3475 3544 3496 3547 3528



severe nodular acne (also called cystic acne). Distances: (3475, 0.0) (3544, 0.1906142195829631) (3496, 0.2987726632666171) (3547, 0.3504096847570036) (3528, 0.35734262487881174) Indices: 3475 3544 3496 3547 3528

Who Treats Acne?

Living With Acne If you have acne, the following recommendations may help you in taking care of your skin. Researchers believe that you may be more likely to get acne if your parents had acne. Who Gets Acne? Who Treats Acne? The following health care providers may diagnose and treat acne: Dermatologists, who specialize in conditions of the skin, hair, and nails. Primary health care providers, including family doctors, internists, or pediatricians. Clean your skin gently. Doctors usually prescribe antibiotics for moderate to severe acne, such as severe nodular acne (also called cystic acne). Distances: (3544, 0.0) (3475, 0.19035778066569053) (3496, 0.4322567428292967) (3547, 0.47406425550858666) (3528, 0.47967745653531935) Indices: 3544 3475 3496 3547 3528

calcium?

A good way to get enough calcium is to eat foods that have calcium. Calcium is a mineral. How much calcium do we need? If we do not get enough calcium in our diets, our body takes the calcium we need from our bones. Over time, this process makes bones weaker and raises the risk of osteoporosis. Osteoporosis is a disease that makes bones weak and brittle, and people with osteoporosis have a higher risk of fractures (broken bones). The amount of calcium each person needs depends on their age and sex. Also, our bones are partly made of calcium salts (mixtures of calcium with other minerals), especially calcium phosphate. Distances: (2544, 0.26251747584106333) (2547, 0.2770751368052815) (2534, 0.347562678176245) (5566, 0.3602630321447623) (2536, 0.375089929366364) Indices: 2544 2547 2534 5566 2536

bye

bye, see you again!

DATA SOURCE

CODEBASE : https://github.com/udaykumarswamy/healthsphere
REF_WEB : https://www.niams.nih.gov/health-topics/all-diseases

BIBLIOGRAPHY

Dave, Tirth, Sai Anirudh Athaluri, and Satyam Singh. 2023. "ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations." *Frontiers in Artificial Intelligence* 6 (May). https://doi.org/10.3389/frai.2023.1169595.

Lahitani, Alfirna Rizqi, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. "Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment." IEEE Xplore. April 1, 2016. https://doi.org/10.1109/CITSM.2016.7577578.