HealthHub: A Healthcare Data Management System

Paolo Palumbo, Francesco Panattoni, Nedal Hadam

June 17, 2025

1 Introduction

Healthhub is a web-based platform designed to simplify and centralize the management of medical appointments. It offers dedicated features for both patients and healthcare professionals. Regular users can search for doctors, book visits, and leave reviews, while medical professionals can manage the services they offer and track their appointments.

- Allow users to search for medical professionals and retrieve information on their services
- Allow users to book appointments and review doctors after a visit
- Allow users to manage their appointment history and cancel or modify bookings
- Give users recommendations on doctors based on their location and past activity
- Allow doctors to specify their specializations, availability, and offered services
- Allow doctors to view and manage their upcoming appointments

2 Dataset and Web Scraping

To populate the dataset for our application, we used a combination of web scraping, synthetic data generation, and logical inference based on real-world patterns of user behavior.

2.1 Web Scraping

The initial data was scraped from a platform containing medical reviews, MioDottore[1]. This platform provides a comprehensive database of medical professionals, including their specializations, locations, and user reviews. The scraping process involved extracting the following information:

- Doctor names and specializations
- Locations (cities and regions)
- User reviews and ratings
- Contact information (where available)
- Service offerings (e.g., types of consultations, treatments)

[1] https://www.miodottore.it/

2.2 Synthetic Data Generation

To enhance the dataset, we generated synthetic data to simulate a more realistic user base and appointment history. This involved several steps:

- **Unique Users**: We extracted all unique usernames from the scraped reviews. For each user, we generated a detailed profile including demographic and geographical information.
- **Appointments**: For every review, we created a corresponding medical appointment. These were dated in the weeks preceding the review, to simulate a realistic flow where users leave feedback shortly after being treated.
- **Likes**: To simulate engagement features such as likes on reviews, we analyzed the provinces of the doctors each user had interacted with. We then randomly selected other doctors in the same provinces and associated a number of likes comparable to the number of reviews written by each user.

2.3 Dataset Characteristics: Velocity and Variety

2.3.1 Velocity

The application context is inherently dynamic, and our dataset reflects this through the following characteristics:

- **High review frequency**: The dataset reflects an average of at least 100 new reviews per day, simulating the continuous flow of patient feedback that a live system would experience.
- **Growing doctor base**: Based on trends from medical boards, we estimate around 400 to 500 new doctors would register on the platform annually, contributing to the system's dynamism and evolving content.

2.3.2 Variety

The dataset incorporates diverse data types, which contribute to its heterogeneity:

- **Multi-format records**: Structured user and doctor profiles, unstructured textual reviews, and timestamped interaction logs (likes, appointments) are all represented.
- **Diverse entities**: The inclusion of users, doctors, reviews, likes, and appointments enables multi-relational analysis and feature richness.

2.4 Resulting Dataset

The original scraped dataset, stored in `scraped.json` (265 MB), contains the raw information collected from the web. This dataset served as the foundation for the data generation process, providing:

- 699,987 reviews
- 214,682 unique reviewers
- 87,632 doctors

Building upon this, the final dataset is stored in JSON format, totaling approximately 960 MB, and includes both real and synthetic data entries. Its main components are:

- **Doctors** (`doctors.json`, 289 MB) Contains 87,632 healthcare professionals, each with profile data and linked to 699,987 reviews. Reviews include timestamps and patient feedback.
- **Users** (`users.json`, 66 MB) Comprises 214,682 unique users extracted from the original dataset and extended with synthetic profiles.

- **Appointments** (`appointments.json`, 422 MB) Each review is connected to a synthetic appointment, scheduled in the weeks preceding the review date. Appointments simulate realistic scheduling and clinic visit patterns.
- **Templates** (`templates.json`, 162 MB) Stores template structures for appointment scheduling logic, such as available time slots, weekdays, and timing constraints.
- **User Likes** (`user_likes.json`, 22 MB) Represents user interactions with doctors, generated according to the geographic distribution of the doctors reviewed by each user.

Overall, this dataset offers a realistic simulation of user and doctor activity on a healthcare review platform, reflecting both volume and behavioral complexity.