685 Mental Disorders: Data Collection and Preparation

Francisco Ribeiro
up202104797@fc.up.pt

Marisa Azevedo
up202108624@fc.up.pt

Ricardo Ferreira
up201907835@fe.up.pt

Toni Grgurevic
up202408625@fe.up.pt

Abstract

This report focuses on the preparation and characterization of a document collection containing 685 mental disorders from Wikipedia. We aim to gather, process, and analyze mental disorders data to create a comprehensive resource. In order to do this, we used techniques like web scraping and API data extraction. We developed a reproducible pipeline for processing the data. This phase involved collecting, organizing, and evaluating the datasets, resulting in a valuable document collection that will serve as a foundation for the remaining milestones.

CCS Concepts: • **Information Systems** → Information Retrieval.

Keywords: Web Scraping, API Data Extraction, Information Retrieval, Pipeline Documentation

1 Introduction

Understanding mental health conditions is essential for advancing research, increasing awareness, and offering better support to those affected. With the information available online, particularly on platforms like Wikipedia, there is a chance to organize and structure this data in a more accessible format. In this project, we aim to collect 685 mental health disorders from Wikipedia.

We utilized web scraping techniques and API data extraction to create a reproducible system that efficiently collects and organizes the information. This resource will serve as a foundation for future research and development in mental health. This initial phase is an important first step toward deepening our understanding of mental health conditions and the data that supports them.

2 Data Sources - Wikipedia

In this project, the primary source of information for mental and neurological disorders was Wikipedia. As one of the largest, most up-to-date, and freely accessible repositories of knowledge on the web, Wikipedia offers extensive information on a wide variety of medical and psychological topics, including mental health disorders. By utilizing both API data extraction and web scraping techniques, we were able to gather comprehensive data from Wikipedia.

2.1 API Data Extraction

Wikipedia provides a robust API (Application Programming Interface), which facilitates automated access to its content.

For this project, we used the API to extract basic information for each disorder, such as, a summary (description), the Wikidata ID and the number of revisions, which can provide insights into the history and credibility of the article.

2.2 Web Scraping with BeautifulSoup

While the API was helpful for extracting structured data, not all the necessary information was readily available via this method. To overcome these limitations, we employed web scraping techniques using the Python library BeautifulSoup. Web scraping involves programmatically accessing a webpage's HTML structure and extracting the desired content.

Through BeautifulSoup, we gathered detailed information from each Wikipedia mental disorder page that was not accessible via the API, such as sections: the causes, symptoms, treatments, and epidemiology of each disorder.

3 Data Collection and Preparation

The process of collecting, preparing, and structuring the data for this project was organized into several key phases, which are illustrated in 9 in the Appendix.

As shown in 9, the pipeline begins with the data collection phase, where information is gathered from Wikipedia using both API data extraction and web scraping techniques. This is followed by data cleaning, where duplicate and incomplete entries are removed, and data enrichment, which includes adding links to Wikidata and further information from additional sources.

3.1 Data Collection

The first phase of the pipeline involves gathering raw data on 685 mental and neurological disorders from Wikipedia pages [1] and [2]. Both types of disorders were stored in separate JSON files for better organization.

Information Processing and Retrieval, FEUP, 2024

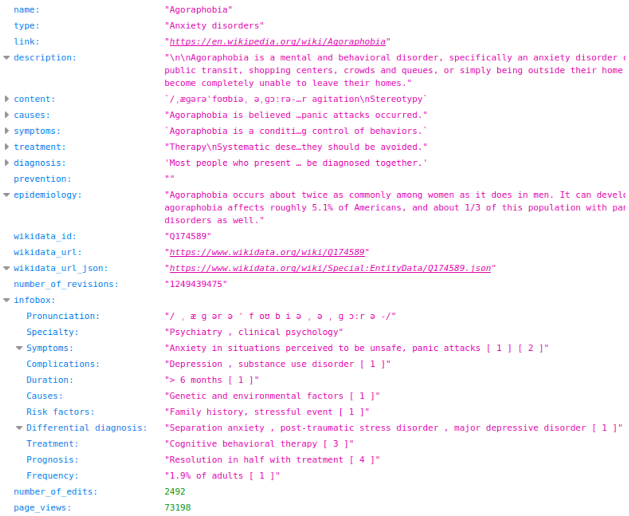
3.2 Data Cleaning

After the initial data collection, the next phase focuses on cleaning the collected data to ensure quality and consistency. During this step, we:

- Remove duplicates: Disorders that appeared on both Wikipedia pages were consolidated into a single entry to ensure only one complete version was retained for each disorder.
- Correct missing data: We noticed that some information of certain disorders were incomplete, so we find out the bug that e was causing the problem and handle that.
- Content Formatting: The raw text contained citation markers (e.g., "[1]", "[2]"), which we removed to improve readability and ensure cleaner, more structured data.

3.3 Data Structuring

We decided to structure the data in a way that captures all relevant information about each disorder while ensuring consistency across entries. The structured data schema for each disorder is illustrated in *Figure 2* in the appendix. Below are the key components of this structure:



```
  name:                       "Agoraphobia"
  type:                       "Anxiety disorders"
  link:                       "https://en.wikipedia.org/wiki/Agoraphobia"
▼ description:                "\n\nAgoraphobia is a mental and behavioral disorder, specifically an anxiety disorder
                              public transit, shopping centers, crowds and queues, or simply being outside their home
                              become completely unable to leave their homes."
▶ content:                    `/ˌæɡərəˈfoʊbiə, ə,ɡɔːrə-…r agitation\nStereotypy`
▶ causes:                     "Agoraphobia is believed …panic attacks occurred."
▶ symptoms:                   `Agoraphobia is a conditi…g control of behaviors.`
▶ treatment:                  "Therapy\nSystematic dese…they should be avoided."
▶ diagnosis:                  'Most people who present … be diagnosed together.'
  prevention:                 ""
▼ epidemiology:               "Agoraphobia occurs about twice as commonly among women as it does in men. It can devel
                              agoraphobia affects roughly 5.1% of Americans, and about 1/3 of this population with par
                              disorders as well."
  wikidata_id:                "Q174589"
  wikidata_url:               "https://www.wikidata.org/wiki/Q174589"
▼ wikidata_url_json:          "https://www.wikidata.org/wiki/Special:EntityData/Q174589.json"
  number_of_revisions:        "1249439475"
▼ infobox:
    Pronunciation:            "/ ˌ æ ɡ ər ə ˈ f oʊ b i ə ˌ ə ˌ ɡ ɔːr ə -/"
    Specialty:                "Psychiatry , clinical psychology"
  ▼ Symptoms:                 "Anxiety in situations perceived to be unsafe, panic attacks [ 1 ] [ 2 ]"
    Complications:            "Depression , substance use disorder [ 1 ]"
    Duration:                 "> 6 months [ 1 ]"
    Causes:                   "Genetic and environmental factors [ 1 ]"
    Risk factors:             "Family history, stressful event [ 1 ]"
  ▼ Differential diagnosis:   "Separation anxiety , post-traumatic stress disorder , major depressive disorder [ 1 ]"
    Treatment:                "Cognitive behavioral therapy [ 3 ]"
    Prognosis:                "Resolution in half with treatment [ 4 ]"
    Frequency:                "1.9% of adults [ 1 ]"
  number_of_edits:            2492
  page_views:                 73198
```

Figure 1. Data Structure

- **Name**: The official name of the disorder as it appears on Wikipedia.
- **Type**: The classification or category of the disorder (e.g., "Anxiety disorders").
- **Link**: The URL link to the full Wikipedia article for further details.
- **Description**: A brief, summary of the disorder, providing essential information.
- **Content**: The full textual content extracted from the article, offering a more in-depth explanation and context.
- **Causes**: A description of the factors or events believed to contribute to the disorder.
- **Symptoms**: The main symptoms or characteristics that are typically observed in individuals with the disorder.
- **Treatment**: Information about available treatments, therapies, or interventions for managing the disorder.
- **Diagnosis**: Criteria and methods used to diagnose the disorder.
- **Prevention**: Any noted strategies or practices aimed at preventing the the disorder.
- **Epidemiology**: Statistical data and prevalence information, highlighting how common the disorder is and any relevant demographic factors.
- **Wikidata ID, URL, JSON**: These fields link the disorder to its corresponding Wikidata entry.
- **Number of Revisions**: This field tracks the number of revisions made to the Wikipedia article, giving an indication of its revision history.
- **Infobox**: The infobox is a structured set of additional attributes that offer a quick reference for key aspects of the disorder.
- **Number of Edits**: This field captures how many times the Wikipedia article has been edited, which can indicate the level of activity.
- **Page Views**: The total number of views the Wikipedia article has received in the last 30 days, providing insight into its popularity or relevance over time.

3.4 Corrections

During the data processing phase, we notice several disorders with missing information, some were because they did not exist, others because the script was not detecting the sections, after some hours trying to understand the situation, we find out that some class names were very common and it pointed to other content that we did not wanted so we had to specify a little more, while doing the scraping.

3.5 Final Data

After passing through the phases of collection, cleaning, structuring, and enrichment, the final dataset was stored in JSON format. This structured format allows for clear visualization and easy navigation of the data, making it ideal for further exploration and analysis. The dataset includes not only the essential attributes such as disorder name, type, and description, but also enriched metadata like the number of revisions, detailed sections on symptoms and treatments, and links to external resources.

4 Data Characterization

In this section, we describe the dataset used to build the search engine, consisting of 685 documents related to different disorders. Each document contains multiple attributes such as name, type, link, description, wikidata id, wikidata url, wikidata url json, number of revisions, content and various optional fields like causes, symptoms, treatment, diagnosis, prevention and epidemiology.Most of data in document is in context and other alredy mentioned optional fields that can be empty if there is no info on that subject on wikipedia. The primary goal of this analysis is to explore the structure and distribution of the dataset in terms of document length, attribute completion (distribution of empty fields), and other characteristics.

4.1 Document length analysis

One key aspect of the dataset is length of documents for each disorder. Figure 2 shows the histogram of document lengths. The distribution is highly skewed, with many documents being relatively short, but there are also a few outliers with very high word counts. Figure also shows mean and median value

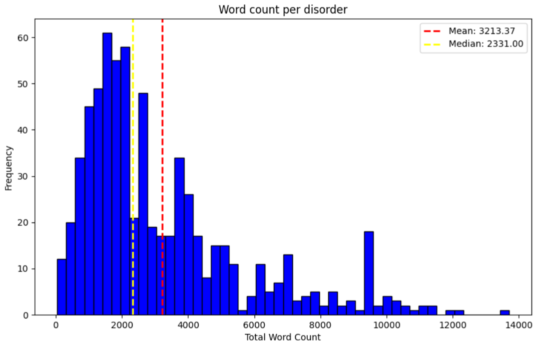


Figure 2. Distribution of document length (number of words) across the dataset

To further explore the variability in document length, Figure 3 presents a box plot. The plot highlights a large variance between documents, with a significant number of outliers having much longer text than the majority. This variance could be due to the nature of the disorders or the completeness of the information provided.

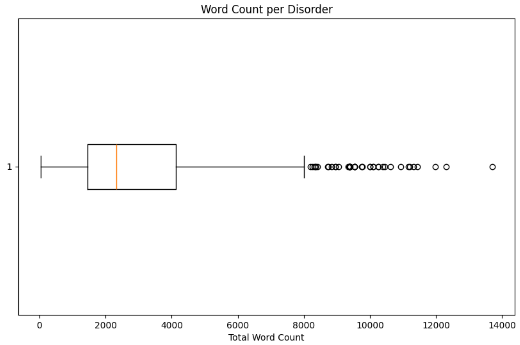


Figure 3. Box plot showing the distribution of document lengths with several outliers

We also examined the differences in document length between different disorder types. Figure 4 shows the mean word count for each type. The analysis reveals that some disorder types tend to have longer or more detailed documents, which may reflect the complexity of the condition or the availability of information.

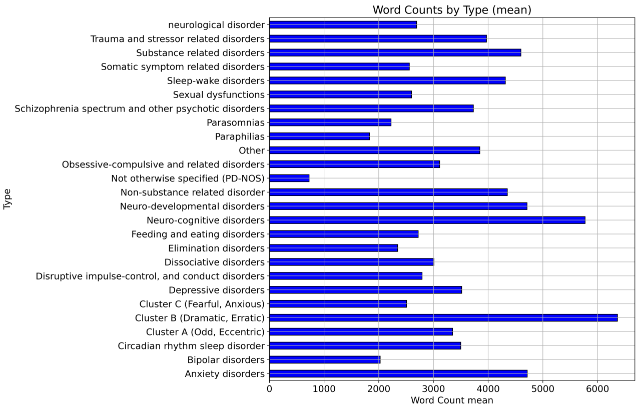


Figure 4. Mean document length by disorder type

Most of documents (disorders) are of type "neurological disorders", 385 different disorders of this type.This distribution may influence the overall performance of the search engine if certain types of disorders are overepresented.

Figure 5 shows a pie chart of how the document length is distributed across attributes. This helps to understand where most of the content lies.Attributes that are short (one sentence) ore not of text type are left out for clearer chart. Content is clearly most populated attribute since information that can't be placed in other optional fields ends up in there.

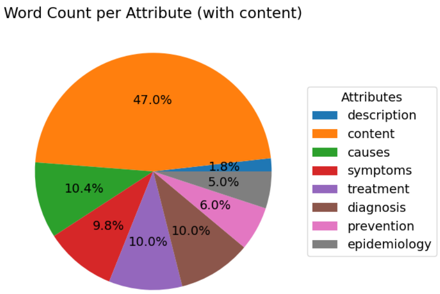


Figure 5. Distribution of document content across different attributes

4.2 Null Values

An important factor in understanding the completeness of the dataset is the presence of null (empty) fields. Figure 6 presents the distribution of null values across the different attributes. There are 2016 empty fields in total, with certain attributes being more commonly left out than others, since not all disorders had sections for that attribute.Most common empty attribute is prevention witch could be even left out from document.

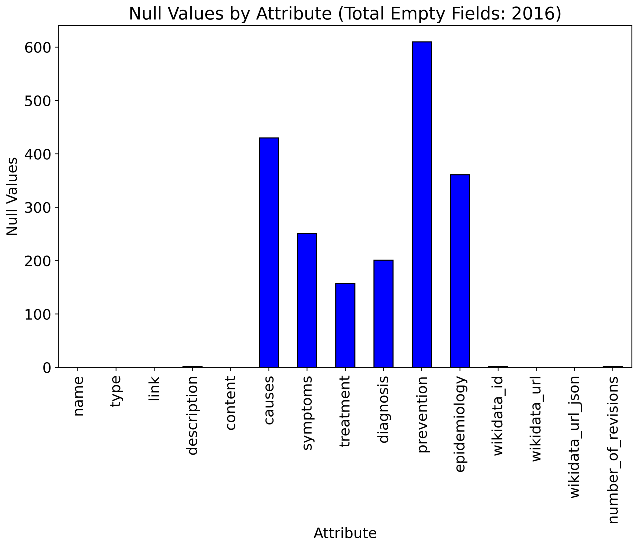


Figure 6. Distribution of null values across different attributes

Since it could also be interesting to see how many empty fields are in documents based on types, we present that in Figure 7.Most of empty fields are in neurological type disorders witch is understandable since most of disorders in data are of that type.Unexpected result is relatively high count of null values in documents with type substance related disorders

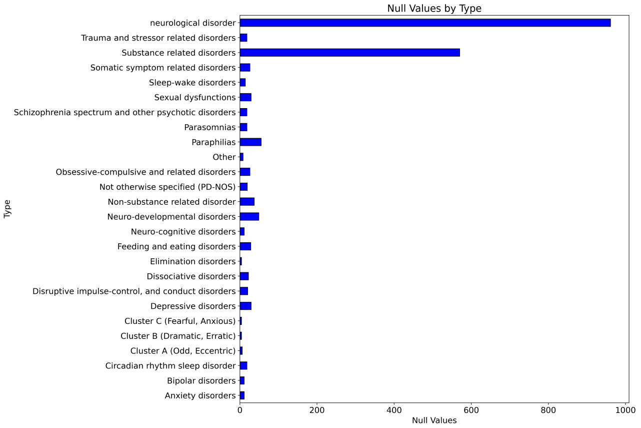


Figure 7. Distribution of null values across different types.

4.3 Word Cloud

To capture the common themes and terms present across the dataset, we generated a word cloud. Figure 8 visualizes the most frequent words across all documents, emphasizing the key concepts and terminology used in describing various disorders.

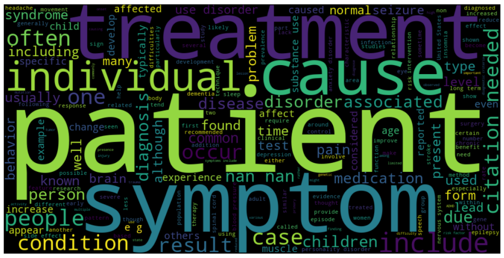


Figure 8. Word cloud representing the most common terms in the dataset.

4.4 Summary

The data characterization revealed that the dataset is highly diverse in terms of document length and attribute completion. Certain types of disorders, such as neurological disorders, are overepresented, and optional attributes like prevention, epidemiology and causes are often missing. This variability may impact the search engine's performance, and future iterations could focus on improving data completeness and balance across disorder types.

References

[1] Wikipedia. 2024. Mental Disorders. https://en.wikipedia.org/wiki/List_of_mental_disorders

[2] Wikipedia. 2024. Neurological Disorders. https://en.wikipedia.org/wiki/List_of_neurological_conditions_and_disorders

A Appendix

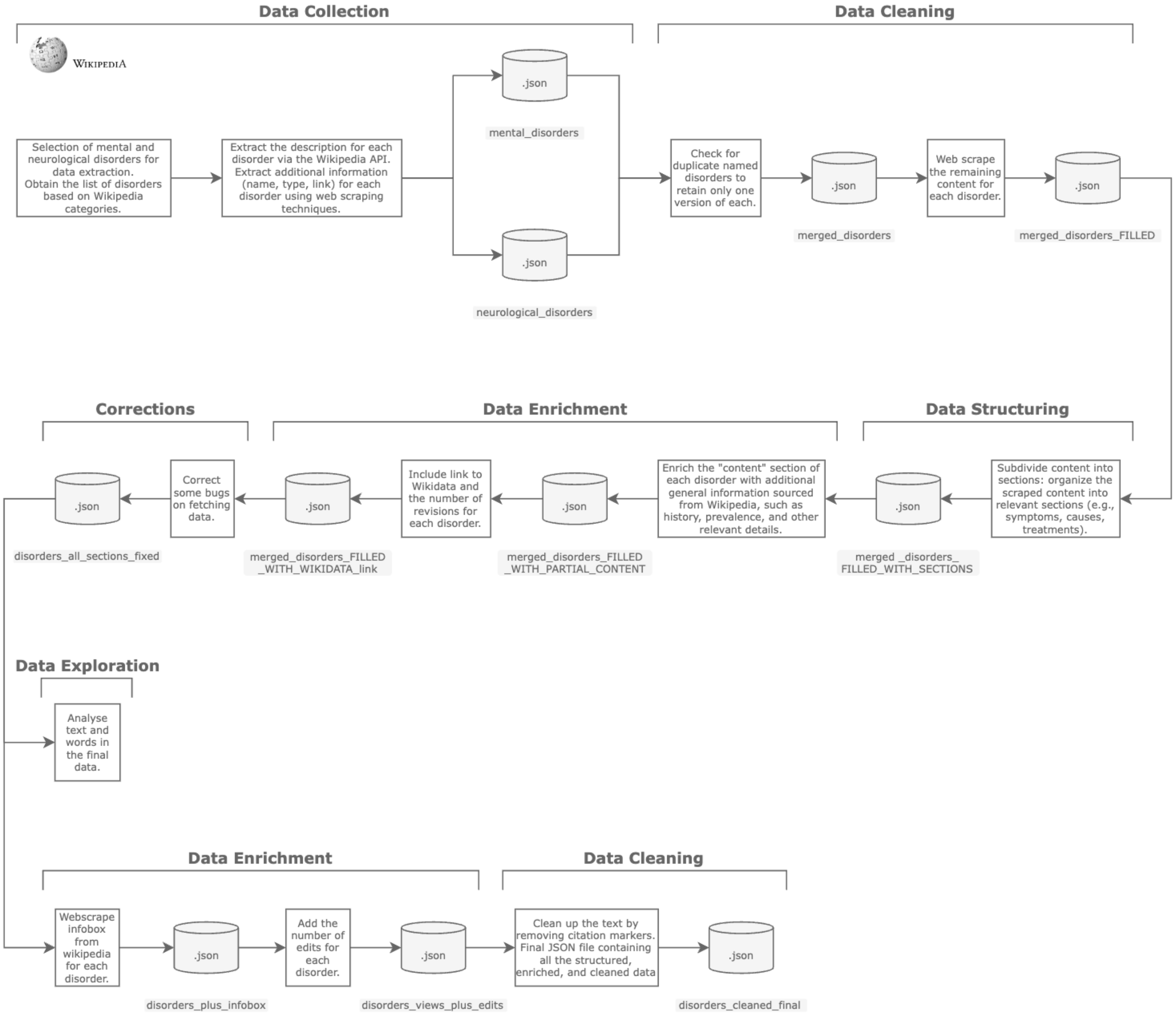

Data Collection
WIKIPEDIA
Selection of mental and neurological disorders for data extraction. Obtain the list of disorders based on Wikipedia categories.
Extract the description for each disorder via the Wikipedia API. Extract additional information (name, type, link) for each disorder using web scraping techniques.
.json
mental_disorders
.json
neurological_disorders
Data Cleaning
Check for duplicate named disorders to retain only one version of each.
.json
merged_disorders
Web scrape the remaining content for each disorder.
.json
merged_disorders_FILLED
Data Structuring
Subdivide content into sections: organize the scraped content into relevant sections (e.g., symptoms, causes, treatments).
.json
merged _disorders_ FILLED_WITH_SECTIONS
Data Enrichment
Enrich the "content" section of each disorder with additional general information sourced from Wikipedia, such as history, prevalence, and other relevant details.
.json
merged_disorders_FILLED _WITH_PARTIAL_CONTENT
Include link to Wikidata and the number of revisions for each disorder.
.json
merged_disorders_FILLED _WITH_WIKIDATA_link
Corrections
Correct some bugs on fetching data.
.json
disorders_all_sections_fixed
Data Exploration
Analyse text and words in the final data.
Data Enrichment
Webscrape infobox from wikipedia for each disorder.
.json
disorders_plus_infobox
Add the number of edits for each disorder.
.json
disorders_views_plus_edits
Data Cleaning
Clean up the text by removing citation markers. Final JSON file containing all the structured, enriched, and cleaned data
.json
disorders_cleaned_final

Figure 9. Pipeline Diagram