PRIMED: A Medicine Search System

Pedro Simões
up202403063@up.pt
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

Miguel Garrido
up202108889@up.pt
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

Emanuel Maia
up202107486@up.pt
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

Guilherme Martins
up202403106@up.pt
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal

Figure 1: Seattle Mariners at Spring Training, 2010.

Abstract

Factors like the simultaneous aging and growth of the population, as well as a larger prevalence of health issues, have lead to a higher demand for medical solutions and associated information systems. In this context, the project tackles the issue of efficiently searching for information about medicines, the diseases they treat and potential side-effects, as well as the companies that produce them and feedback left by other patients. Therefore, this article's goal is to provide clear and extensive insights on the development of a search engine for the medicines themselves and other aforementioned associated information, by integrating previously prepared data collected from multiple sources and properly analysing it.

CCS Concepts

• **Information systems** → Information retrieval query processing.

Keywords

Medicine, Treatments, Sickness, Pipeline, Data, Gathering, Scraping, Preparation, Search Engine

PRI - G51, October 13, 2024, Porto, PT

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Pedro Simões, Miguel Garrido, Emanuel Maia, and Guilherme Martins. 2018. PRIMED: A Medicine Search System. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (PRI - G51).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Nowadays, accurate and comprehensive medicine information plays a vital role in decision-making on the healthcare and investigation sectors. Having access to clean and relevant information is crucial to enhance the efficiency and safety of treatments.

The ***PRIMED*** project aims to compile and organize such data so that it can be accessed in an easy and organized way. It provides insights that can be applied both in practical cases or simply in developing new health safety politics.

In this part of the project, data from various sources was collected, processed and analysed with the goal of creating a solid basis for a pharmaceutical system. The whole process is described in this document, from the choice of the theme itself up to the analysis of gathered data and its subsequent classification.

2 Theme Selection

The choice to focus on medicines as the central theme for this project stems from the critical role they play in modern healthcare. Medication is the primary tool for treating multiple health problems and conditions, making them an essential component in both public healthcare systems and individual patient care. The data surrounding these substances, such as active substances, applicable

cases and clinical trials, offers valuable insights that can improve decision-making in medical practice and pharmaceutics.

3 Data Collection

This chapter outlines the data selection process and the methods utilized to gather it.

3.1 Selection of Data

One of the main challenges of the healthcare sector is ensuring that accurate and up-to-date information on medication is available to healthcare professionals. Given these needs, the selection of data to be collected was made:

- `Medicines`: The main component of the data, consisting of medicines and their respective relevant, intrinsic information.
- `Diseases`: To complement the collected data, information on some diseases was collected so that it would be possible to get more information to allow for an easier medicine selection process.
- `Pharmaceutical Companies`: Some pharmaceutical companies are more trusted than others due to their credibility and higher quality products, which can impact the decision-making process.
- `Reviews`: It is important to know the success rate of the presented medicine and how the people who use it feel about it.

With this data, the aim is to provide the required information to the users.

3.2 Gathering

Finding data suitable for the project proved to be a challenge; not only did it have to be relevant and accurate, but it also had to meet some criteria in terms of quantity and quality. Therefore, the data had to be gathered from multiple sources using different methods - most of the data came from prepared datasets found on Kaggle[1], while the rest came from scraping Wikipedia[2].

3.2.1 Medicines. The *Medicines* dataset[3], retrieved from Kaggle, contains a list of pharmaceutical treatments and some relevant, mostly textual, information about the cases where it is used.

The present information on this dataset is:

- `Medicine Name`: The name of the medicine.
- `Composition`: The active substance present in the medicine.
- `Uses`: A list of cases where the medicine is used (specific diseases, for example).
- `Side Effects`: Lists possible side effects resulting from the medicine's usage.
- `Manufacturer`: The name of the company responsible for producing the medicine.
- `Reviews`: Three additional columns containing the percentage of "Excellent", "Average" and "Poor" reviews for each medicine's treatment results.

This dataset possesses a ***CC0 1.0 Universal***[4] license, which means the data is part of the public domain, allowing for the copying and modification of the data.

3.2.2 Diseases. This dataset was scraped from the tables of a Wikipedia page containing a list of autoimmune diseases[5]; by gathering this data, the goal was to complement the previous dataset's "Uses" and "Side Effects" columns by collecting more information on this specific subset of diseases.

The information extracted from this dataset consists of:

- `Disease`: The name of the disease.
- `Primary Organ/Body Part Affected`: Information on the organs or body parts affected by the disease.
- `Autoantibodies`: The antibodies associated with each specific disease.
- `Acceptance as an Autoimmune Disease`: Classification for each disease related to its acceptance as an autoimmune condition, based on the current scientific consensus and level of evidence supporting its autoimmune nature.
- `Prevalence Rate (US))`: The percentage of people or how many the disease affects in the US.

The gathered dataset contains about 110 lines of diseases, being available under the ***Creative Commons Attribution-ShareAlike 4.0 International*** license[6], which allows for the sharing and adaptation of the contents.

3.2.3 Pharmaceutical Companies. To compliment the data on companies, another scraping of Wikipedia was made so that there could be more data on pharmaceutical companies.

The gathered information for the companies was:

- `Company name`: The name of the company.
- `Year`: The year the company was created and, if available, when the company was shut down.
- `Description`: A short description of each company, it's values and some more information about them.

The origin of the information was a ***pharmaceutical companies list*** where around 700 companies names and years were gathered from and, for each company, their respective Wikipedia page's first description. This dataset also falls under the same licence (***CC0 1.0 UniversalWikipedia:Text of the Creative Commons Attribution-ShareAlike 4.0 International License***) as the previous dataset, as it was gathered in a similar way.

3.2.4 Reviews. Lastly, some reviews for the arranged medicine with more personal descriptions from users was added.

The information is:

- `Unique ID`: The ID of the review.
- `Drug Name`: Name of the drug/treatment.
- `Condition`: The condition of the patient where it was used.
- `Review`: Written review of the experience of taking the drug.
- `Rating`: Number from 0-10 that expresses the quality of the drug.
- `Date`: Date of when the drug was applied.
- `Useful Count`: Similar to likes, this provides the number of other people who found this review useful.

This information was gathered from ***UC Irvine***, containing around 215000 entries. This dataset is covered by the ***Attribution 4.0 International*** licence, which allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

4 Modifications

Modifying the template — including but not limited to: adjusting margins, typeface sizes, line spacing, paragraph and list definitions, and the use of the `\vspace` command to manually adjust the vertical spacing between elements of your work — is not allowed.

Your document will be returned to you for revision if modifications are discovered.

5 Pipeline Description

6 Conceptual Data Model

7 Information Needs

The information needs on this data can vary from the person using it. If the one who is using the tool is, for example, a doctor or pharmaceutical, the needs might be related to the compositions of medicine or cases where it was applied. If the user is an average person, he can be more interested in checking who made the medicine to check if it can be trusted or to check other user's reviews on that treatment/possible side effects to have a better understanding of what can happen to him.

Here is a list of possible information needs, explaining what kind of data on the is needed, why and for whom:

- `Medicine compositions`: The composition of the treatment is important because some might even harm the patient so it is important to keep an eye out for that.
- `Uses`: Not every medicine has clear use cases, some might have a principal use but an also good side use that is not much common.
- `Side effects`: Clearly the users should be able to know the possible side effects of their treatment so that they can be prepared and informed of possible strange events.
- `Manufacturer`: Some people have preferential manufacturers who they trust more than others or can simply want to know more about the one they were prescribed.
- `Reviews`: For all possible types of users, the reviews are always important, both textual or not, they indicate other experiences. Sometimes peculiar cases are described there where the patients got rare effects from said treatment.
- `Diseases' primary organs/autoantibodies`: In some cases, experts might need to know how what part of the body or what are the antibodies that are affected by the diseases in order to prescribe the best solution.
- `Prevalence rate`: It can also be relevant to note how common certain disease when making the diagnosis of the patient or just for statistics.

Based one these needs we can answer some questions like:

- Which medicines can are more commonly used for certain conditions?
- How different manufacturers affect the treatment of specific diseases?
- How can certain health conditions be treated with the most effectiveness?
- With certain symptoms, what treatment should be applied?
- Are these effects normal after taking certain medication?
- What happened to other people who had the same treatment as me?

Lastly, as previously stated, people such as normal users, medicine experts and even investigators can take advantage from this type of information. From gaining awareness, to helping people with special needs on recovering.

8 Dataset Characterization

9 Conclusions

10 Typefaces

The "`acmart`" document class requires the use of the "Libertine" typeface family. Your TEX installation should include this set of packages. Please do not substitute other typefaces. The "`lmodern`" and "`ltimes`" packages should not be used, as they will override the built-in typeface families.

11 Title Information

The title of your work should use capital letters appropriately - https://capitalizemytitle.com/ has useful rules for capitalization. Use the `title` command to define the title of your work. If your work has a subtitle, define it with the `subtitle` command. Do not insert line breaks in your title.

If your title is lengthy, you must define a short version to be used in the page headers, to prevent overlapping text. The `title` command has a "short title" parameter:

```
\title[short title]{full title}
```

12 Authors and Affiliations

Each author must be defined separately for accurate metadata identification. As an exception, multiple authors may share one affiliation. Authors' names should not be abbreviated; use full first names wherever possible. Include authors' e-mail addresses whenever possible.

Grouping authors' names or e-mail addresses, or providing an "e-mail alias," as shown below, is not acceptable:

```
\author{Brooke Aster, David Mehldau}
\email{dave,judy,steve@university.edu}
\email{firstname.lastname@phillips.org}
```

The `authornote` and `authornotemark` commands allow a note to apply to multiple authors — for example, if the first two authors of an article contributed equally to the work.

If your author list is lengthy, you must define a shortened version of the list of authors to be used in the page headers, to prevent overlapping text. The following command should be placed just after the last `\author{}` definition:

```
\renewcommand{\shortauthors}{McCartney, et al.}
```

Omitting this command will force the use of a concatenated list of all of the authors' names, which may result in overlapping text in the page headers.

The article template's documentation, available at https://www.acm.org/publications/proceedings-template, has a complete explanation of these commands and tips for their effective use.

Note that authors' addresses are mandatory for journal articles.

13 Rights Information

Authors of any work published by ACM will need to complete a rights form. Depending on the kind of work, and the rights management choice made by the author, this may be copyright transfer, permission, license, or an OA (open access) agreement.

Regardless of the rights management choice, the author will receive a copy of the completed rights form once it has been submitted. This form contains LaTeX commands that must be copied into the source document. When the document source is compiled, these commands and their parameters add formatted text to several areas of the final document:

- the "ACM Reference Format" text on the first page.
- the "rights management" text on the first page.
- the conference information in the page header(s).

Rights information is unique to the work; if you are preparing several works for an event, make sure to use the correct set of commands with each of the works.

The ACM Reference Format text is required for all articles over one page in length, and is optional for one-page articles (abstracts).

14 CCS Concepts and User-Defined Keywords

Two elements of the "acmart" document class provide powerful taxonomic tools for you to help readers find your work in an online search.

The ACM Computing Classification System — https://www.acm.org/publications/class-2012 — is a set of classifiers and concepts that describe the computing discipline. Authors can select entries from this classification system, via https://dl.acm.org/ccs/ccs.cfm, and generate the commands to be included in the LaTeX source.

User-defined keywords are a comma-separated list of words and phrases of the authors' choosing, providing a more flexible way of describing the research being presented.

CCS concepts and user-defined keywords are required for for all articles over two pages in length, and are optional for one- and two-page articles (or abstracts).

15 Sectioning Commands

Your work should use standard LaTeX sectioning commands: `section`, `subsection`, `subsubsection`, and `paragraph`. They should be numbered; do not remove the numbering from the commands.

Simulating a sectioning command by setting the first word or words of a paragraph in boldface or italicized text is **not allowed.**

16 Tables

The "`acmart`" document class includes the "`booktabs`" package — https://ctan.org/pkg/booktabs — for preparing high-quality tables.

Table captions are placed *above* the table.

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material are found in the *LaTeX User's Guide*.

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
\$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

Always use midrule to separate table header rows from data rows, and use it only for this purpose. This enables assistive technologies to recognise table headers and support their users in navigating tables more easily.

17 Math Equations

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

17.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual `\begin . . . \end` construction or with the short form `$ . . . $`. You can use any of the symbols and structures, from α to ω, available in LaTeX [**?**]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n\to\infty} x = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

17.2 Display Equations

A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

Again, in either environment, you can use any of the symbols and structures available in LaTeX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n\to\infty} x = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{\infty} x + 1$$

Table 2: Some Typical Commands

Command	A Number	Comments
`\author`	100	Author
`\table`	300	For tables
`\table*`	400	For wider tables

and follow it with another numbered equation:

$$\sum_{i=0}^{\infty} x_i = \int_0^{\pi+2} f \tag{2}$$

just to demonstrate LaTeX's able handling of numbering.

18 Figures

The "`figure`" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Figure 2: 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

Your figures should contain a caption which describes the figure to the reader.

Figure captions are placed *below* the figure.

Every figure should also have a figure description unless it is purely decorative. These descriptions convey what's in the image to someone who cannot see it. They are also used by search engine crawlers for indexing images, and when images cannot be loaded.

A figure description must be unformatted plain text less than 2000 characters long (including spaces). **Figure descriptions should not repeat the figure caption – their purpose is to capture important information that is not already provided in the caption or the main text of the paper.** For figures that convey important and complex new information, a short text description may not be adequate. More complex alternative descriptions can be placed in an appendix and referenced in a short figure description. For example, provide a data table capturing the information in a bar chart, or a structured list representing a graph. For additional information regarding how best to write figure descriptions and why doing this is so important, please see https://www.acm.org/publications/taps/describing-figures/.

18.1 The "Teaser Figure"

A "teaser figure" is an image, or set of images in one figure, that are placed after all author and affiliation information, and before the body of the article, spanning the page. If you wish to have such a figure in your article, place the command immediately before the `\maketitle` command:

```
\begin{teaserfigure}
  \includegraphics[width=\textwidth]{sampleteaser}
  \caption{figure caption}
  \Description{figure description}
\end{teaserfigure}
```

19 Citations and Bibliographies

The use of BibTeX for the preparation and formatting of one's references is strongly recommended. Authors' names should be complete — use full first names ("Donald E. Knuth") not initials ("D. E. Knuth") — and the salient identifying features of a reference should be included: title, year, volume, number, pages, article DOI, etc.

The bibliography is included in your source document with these two commands, placed just before the `\end{document}` command:

```
\bibliographystyle{ACM-Reference-Format}
\bibliography{bibfile}
```

where "`bibfile`" is the name, without the "`.bib`" suffix, of the BibTeX file.

Citations and references are numbered by default. A small number of ACM publications have citations and references formatted in the "author year" style; for these exceptions, please include this command in the **preamble** (before the command "`\begin{document}`") of your LaTeX source:

```
\citestyle{acmauthoryear}
```

Some examples. A paginated journal article [**?**], an enumerated journal article [**?**], a reference to an entire issue [**?**], a monograph (whole book) [**?**], a monograph/whole book in a series (see 2a in spec. document) [**?**], a divisible-book such as an anthology or compilation [**?**] followed by the same example, however we only output the series if the volume number is given [**?**] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [**?**], a chapter in a divisible book in a series [**?**], a multi-volume work as book [**?**], a couple of articles in a

proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [**?** **?**], a proceedings article with all possible elements [**?**], an example of an enumerated proceedings article [**?**], an informally published work [**?**], a couple of preprints [**?** **?**], a doctoral dissertation [**?**], a master's thesis: [**?**], an online document / world wide web resource [**?** **?** **?**], a video game (Case 1) [**?**] and (Case 2) [**?**] and [**?**] and (Case 3) a patent [**?**], work accepted for publication [**?**], 'YYYYb'-test for prolific author [**?**] and [**?**]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [**?**]. Boris / Barbara Beeton: multi-volume works as books [**?**] and [**?**]. A couple of citations with DOIs: [**?** **?**]. Online citations: [**?** **?** **?**]. Artifacts: [**?**] and [**?**].

20 Acknowledgments

Identification of funding sources and other support, and thanks to individuals and groups that assisted in the research and the preparation of the work should be included in an acknowledgment section, which is placed just before the reference section in your document.

This section has a special environment:

```
\begin{acks}
...
\end{acks}
```

so that the information contained therein can be more easily collected during the article metadata extraction phase, and to ensure consistency in the spelling of the section heading.

Authors should not prepare this section as a numbered or unnumbered `\section`; please use the "`acks`" environment.

21 Appendices

If your work needs an appendix, add it before the "`\end{document}`" command at the conclusion of your source document.

Start the appendix with the "`appendix`" command:

```
\appendix
```

and note that in the appendix, sections are lettered, not numbered. This document has two appendices, demonstrating the section and subsection identification method.

22 Multi-language papers

Papers may be written in languages other than English or include titles, subtitles, keywords and abstracts in different languages (as a rule, a paper in a language other than English should include an English title and an English abstract). Use `language=...` for every language used in the paper. The last language indicated is the main language of the paper. For example, a French paper with additional titles and abstracts in English and German may start with the following command

```
\documentclass[sigconf, language=english, language=german,
               language=french]{acmart}
```

The title, subtitle, keywords and abstract will be typeset in the main language of the paper. The commands `\translatedXXX`, `XXX` begin title, subtitle and keywords, can be used to set these elements in the other languages. The environment `translatedabstract` is used to set the translation of the abstract. These commands and environment have a mandatory first argument: the language of the second argument. See `sample-sigconf-i13n.tex` file for examples of their usage.

23 SIGCHI Extended Abstracts

The "`sigchi-a`" template style (available only in LaTeX and not in Word) produces a landscape-orientation formatted article, with a wide left margin. Three environments are available for use with the "`sigchi-a`" template style, and produce formatted output in the margin:

`sidebar:` Place formatted text in the margin.
`marginfigure:` Place a figure in the margin.
`margintable:` Place a table in the margin.

Acknowledgments

To Robert, for the bagels and explaining CMYK and color spaces.

References

[1] Kaggle. https://www.kaggle.com, 2024.
[2] Wikipedia. https://www.wikipedia.org/, 2024.
[3] Kaggle. 11000 medicine details. https://www.kaggle.com/datasets/singhnavjot2062001/11000-medicine-details, 2024.
[4] Creative Commons. Cc0 1.0 universal. https://creativecommons.org/publicdomain/zero/1.0/.
[5] Wikipedia. https://en.wikipedia.org/wiki/List_of_autoimmune_diseases, 2024.
[6] Creative Commons. Creative commons attribution-sharealike 4.0 international license. https://en.wikipedia.org/wiki/Wikipedia:Text_of_the_Creative_Commons_Attribution-ShareAlike_4.0_International_License.

A Research Methods

A.1 Part One

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna, id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.

A.2 Part Two

Etiam commodo feugiat nisl pulvinar pellentesque. Etiam auctor sodales ligula, non varius nibh pulvinar semper. Suspendisse nec lectus non ipsum convallis congue hendrerit vitae sapien. Donec at laoreet eros. Vivamus non purus placerat, scelerisque diam eu, cursus ante. Etiam aliquam tortor auctor efficitur mattis.

B Online Resources

Nam id fermentum dui. Suspendisse sagittis tortor a nulla mollis, in pulvinar ex pretium. Sed interdum orci quis metus euismod, et sagittis enim maximus. Vestibulum gravida massa ut felis suscipit congue. Quisque mattis elit a risus ultrices commodo venenatis eget dui. Etiam sagittis eleifend elementum.

Nam interdum magna at lectus dignissim, ac dignissim lorem rhoncus. Maecenas eu arcu ac neque placerat aliquam. Nunc pulvinar massa et mattis lacinia.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009