supper&supper

Development of the „Product Finder“ for a Generics Manufacturer through the Use of Web Crawling, Text Mining and Power BI

Project Goal

In order to have a sufficient amount of time for the development of generics, it is important for generics manufacturers to identify market niches of active ingredients at an early stage. Furthermore, prior to development, market attractiveness of the potential active ingredients needs to be assessed.

The goal of the project was therefore to develop a tool called “Product Finder“. Taking patent terms into account, this tool easily identifies active ingredients, application areas and pharmaceutical products that can then be added to the development pipeline of a generics manufacturer.

A user-friendly interface in form of a dashboard was developed to visualize the acquired information.

Utilized Datasets

Using the database of the European Medicines Agency (EMA)[1], approximately 1,250 entries were compiled from several sub-sites by the web crawler programmed for this purpose and formed the first consolidated dataset.

The downloadable database of the U.S. Food and Drug Administration(FDA)[2] served as the second record for the “Product Finder“ with approximately 95,000 entries. Through the use of Text Mining, an additional data set of the most important keywords contained in these two records was created.

Applied Methods

As a first step, the data of the EMA database was imported and merged into a structured data set by the Python-based web crawler.

The FDA database was available for download as a CSV file. The use of a web crawler was therefore not necessary for this database. Additionally, in contrast to the EMA database, it already had a high level of detail and structure.

Subsequently, within the framework of text mining, the Natural Language Toolkit (NLTK) - a Python platform for the identification of human language - was used on the extracted EMA data record. Through data cleaning, scoring models and stop word lists, the most important keywords for drug descriptions were obtained.

Conventional Data Wrangling methods were used to determine the content of the EMA and FDA datasets to prepare for better integration with Power BI and to get rid of inconsistencies across the two datasets: columns with multiple values, i.e. different fields of application or ingredients, have been split into multiple columns and different spelling as well as unneeded special characters have been removed. New calculated fields have been inserted, for example the patent expiration date or the deadline for generics development and have been added to the datasets. In addition, the market attractiveness of the individual drugs was determined, following the assumption that the fewer generics there are in a particular application, the more attractive an ingredient becomes - the lower the ratio, the higher the attractiveness of the particular generic.

Challenges

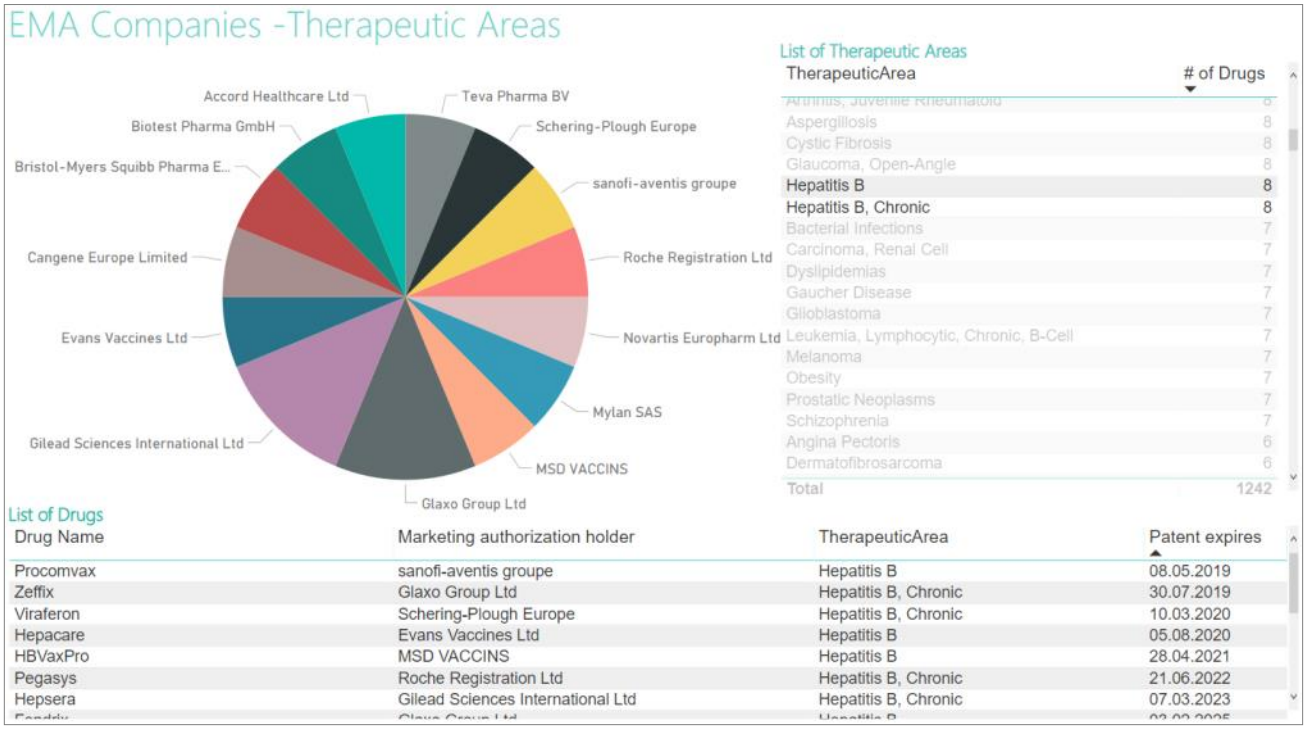
One of the biggest challenges in consolidating the data was the quality of the FDA database. Data cleaning and wrangling ensured that the data was ready for consumption inside the dashboard application.

Another challenging aspect of the project was the difference of granularity between the FDA and EMA datasets as the data from the FDA provided significantly more detail than that of the EMA. To consolidate the two into a single dataset, we used filters, sorting and mapping algorithms.

Project Outcome

The “Product Finder“ supports generics manufacturers in the search for expiring drug patents and identifies active ingredients and generics that could potentially be added to the development pipeline.

The data is presented in a user-friendly, dynamic dashboard. Depending on individual needs and use cases, different dashboards can be selected and displayed. One of the many uses of the dynamic filters is the retrieval of an up-to-date list of patents that are about to expire. Drill downs can be utilized to further analyze details of an active ingredient or to display all application areas and products for a particular manufacturer. Thus you are able to observe the activities of your competitors through the ”Product Finder”.



Sources:

[1] European Medicines Agency: European public assessment reports. Human medicines. http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&mid=WC0b01ac058001d124 July 2018.

[2] U.S. Food and Drug Administration: Medical Device Databases. https://www.fda.gov/drugs/informationondrugs/ucm079750.htm July 2018