supper&supper

Entwicklung des „Product Finders“ für einen Generika-Hersteller durch die Verwendung von Web Crawling, Text Mining und Power BI

Projektziel

Um genügend Zeit für die Entwicklung eines Generikums zu haben ist es für Generika-Hersteller wichtig, frühzeitig Marktnischen pharmazeutischer Wirkstoffe zu identifizieren. Im Vorfeld des Entwicklungsprozesses muss zudem die Marktattraktivität der potentiellen Wirkstoffe bestimmt werden.

Ziel des Projektes war deshalb die Entwicklung eines Tools namens „Product Finder“. Mit diesem sollen unter Berücksichtigung der Patentlaufzeiten Wirkstoffe, Anwendungsgebiete und pharmazeutische Produkte leicht identifizierbar gemacht und zur Pipeline der Entwicklung eines Generika-Herstellers hinzugefügt werden. Eine benutzerfreundliche Bedienoberfläche in Form eines Dashboards sollte die gewonnenen Informationen visualisieren.

Verwendete Datensätze

Die rund 1.250 Einträge der Datenbank der European Medicines Agency (EMA)[1] wurden durch den dafür programmierten Webcrawler aus mehreren Unterseiten zusammengetragen und bildeten den ersten verfügbaren Datensatz.

Die herunterladbare Datenbank der U.S. Food and Drug Administration (FDA)[2] mit ca. 95.000 Einträgen diente als zweiter Datensatz für den „Product Finder“. Ein weiterer Datensatz war das Ergebnis des Text Minings auf die beiden Datensätze und umfasste die wichtigsten enthaltenen Schlagworte.

Angewandte Methoden

Zuerst wurden die Daten der EMA-Datenbank durch Webcrawling, basierend auf einem Python-Skript, ausgelesen und zu einem strukturierten Datensatz zusammengefügt.

Die Datenbank der FDA stand als CSV-Datei zum Download zur Verfügung. Der Einsatz eines Webcrawlers war daher für diese Datenbank nicht notwendig. Zudem wies sie im Gegensatz zur EMA-Datenbank bereits einen sehr hohen Detaillierungsgrad auf.

Anschließend wurden im Rahmen eines Text Minings auf dem gewonnenen EMA-Datensatz das Natural Language Toolkit (NLTK) - eine Python-Plattform zur Identifikation der menschlichen Sprachdaten - genutzt. Durch Cleaning, Scoring Modelle und Stopword Listen wurden die wichtigsten Schlagworte zu den Medikamentenbeschreibungen aus den Daten gewonnen.

Klassische Methoden des Data Wranglings wurden angewandt, um den Inhalt der EMA- und FDA-Datensätze zur besseren Integration in Power BI zu vereinheitlichen und konsistent zu halten. Spalten mit mehreren Werten, wie zum Beispiel verschiedene Anwendungsgebiete und Inhaltsstoffe wurden aufgesplittet, abweichende Schreibweisen angepasst und nicht benötigte Sonderzeichen entfernt. Es wurden neu berechnete Felder eingefügt, zum Beispiel das Patentablaufdatum oder die Deadline zur Generika-Entwicklung. Außerdem wurde die Marktattraktivität der einzelnen Medikamente ermittelt: es wurde die Annahme getroffen, dass ein Wirkstoff attraktiver wird, je weniger Generika in einem bestimmten Anwendungsgebiet existieren – je geringer dieses Verhältnis, desto höher die Attraktivität.

Herausforderungen

Eine der größten Herausforderungen war die Qualität der FDA-Datenbank. Data Cleaning und Wrangling sorgten dafür, dass die Daten in der Dashboard-Anwendung nutzbar wurden.

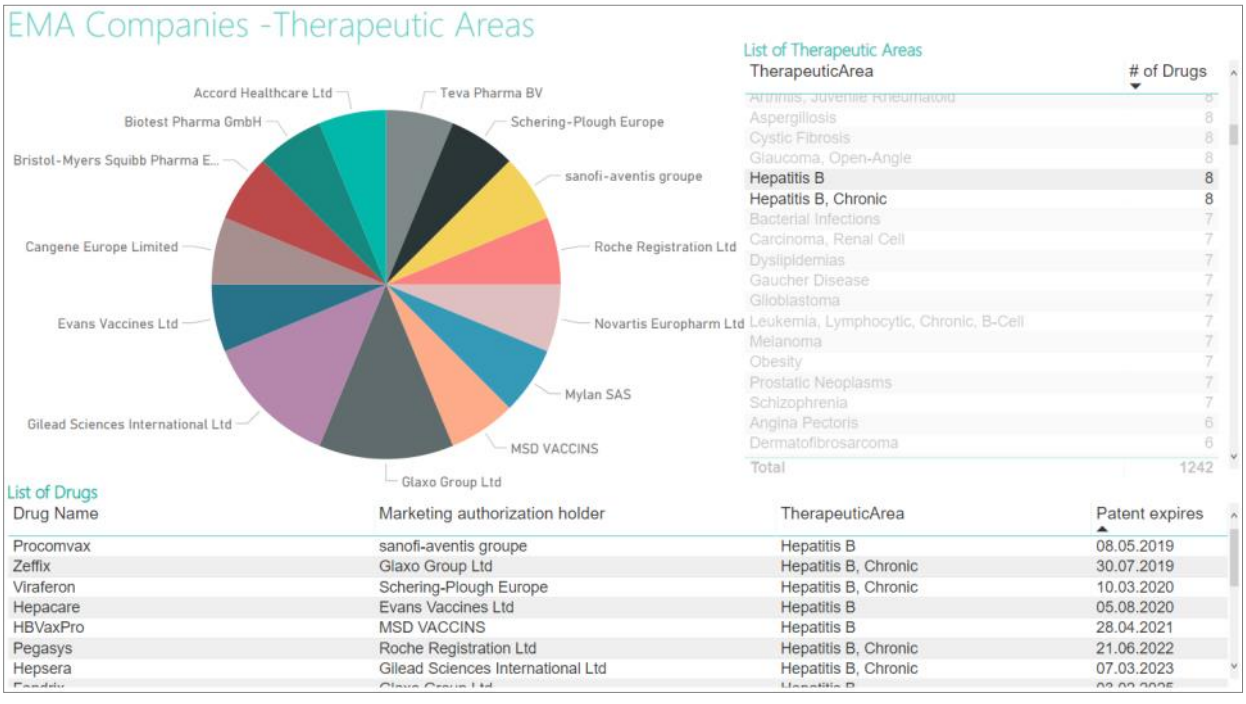
Gleichzeitig wiesen die FDA- und EMA-Datenbanken eine unterschiedliche Granularität auf – die der FDA enthielt deutlich mehr Details als die der EMA. Die beiden Datensätze konsolidierten wir in einen Datensatz, indem wir Filter, Sortierungen und Mapping-Algorithmen verwendeten.

Projektergebnis

Der „Product Finder“ unterstützt Generika-Hersteller bei der Suche nach auslaufenden Medikamentenpatenten und identifiziert potentiell zur Entwicklungs-Pipeline hinzufügbare pharmazeutische Wirkstoffe und Generika.

Die Daten werden in einem benutzerfreundlichen, dynamischen Dashboard präsentiert. Je nach individuellem Anwendungsfall können unterschiedliche Dashboards ausgewählt und angezeigt werden. Dynamische Filter ermöglichen unter anderem eine aktuelle Liste mit bald auslaufenden Patenten zu extrahieren. Mittels Drilldowns können weitere Details eines Wirkstoffes analysiert werden oder für einen Hersteller alle Anwendungsgebiete und Produkte angezeigt werden. Im „Product Finder“ können Sie so die Aktivitäten Ihrer Wettbewerber beobachten.

Aufgrund der technischen Flexibilität des „Product Finders“ können je nach Kundenwunsch zusätzliche oder andere Datenbanken eingebunden sowie weitere Dashboards ergänzt werden.



Quellen:

[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 Juli 2018.

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