Tokenized product information for centrally approved medicines within EU (extracted May 3, 2022)

SND-ID: 2022-157-1. **Version**: 1. **DOI**: https://doi.org/10.57804/ggrw-hr06

Download data

PL_export_all_220822.csv *(43.34 MB)*

SmPC_export_all_220822.csv *(150.28 MB)*

Associated documentation

PI data desc.txt *(986 bytes)*

Download all files

2022-157-1-1.zip *(~193.62 MB)*

Citation

Westman, G. (2022) Tokenized product information for centrally approved medicines within EU (extracted May 3, 2022) (Version 1) [Data set]. Uppsala University. Available at: https://doi.org/10.57804/ggrw-hr06

Creator/Principal investigator(s)

Gabriel Westman - Uppsala University

Research principal

Uppsala University - Department of Medical Sciences

Description

The text corpus was compiled on May 3, 2022, by scripted downloading of all available English language product information files for all centrally approved medicinal products within the EU, from the European Medicines Agency website. Package Leaflet (PL) and Summary of product characteristics (SmPC) documents for each medicinal product, excluding multiplicate documents for medicinal products with more than one strength or pharmaceutical preparation, were used. The PDF files were scraped using the pdfplumber version 0.6.1 package in Python 3.8.10 to extract all text except page numbering, headers, and footers.

Line breaks and special characters (excluding punctuation characters) were removed, and punctuation was added to sentences where this was missing (such as headings) to avoid false aggregation. All paragraphs were tokenized on a sentence level using the Natural Language Toolkit (NLTK) version 3.7 tokenizer

This database contains sentence-level tokenized product infomation from all centrally approved medicinal products within the EU (May 3, 2022) including Summary of product characteristics (SmPC) and Package leaflet (PL) documents.

A total of 1258 medicinal products were initially included, of which 5 were subsequently excluded due to document compatibility issues. From these, a total of 783 K sentences were extracted from PL and SmPC documents.

Data contains personal data

No

Language

English

Population

All centrally approved medicinal products within EU

Study design

Observational study

Description of study design

Health informatics study on information about approved medicinal products.

Data format / data structure

Text

Responsible department/unit

Department of Medical Sciences

Commissioning organisation

Swedish Medical Products Agency

Research area

Computer and information science (Standard för svensk indelning av forskningsämnen 2011)

Basic medicine (Standard för svensk indelning av forskningsämnen 2011)

Keywords

Information science, Linguistics, Artificial intelligence, Pharmacy

Publications

Bergman E, Sherwood K, Forslund M, Arlett P, Westman G (2022) A natural language processing approach towards harmonisation of European medicinal product information. PLoS ONE 17(10): e0275386. https://doi.org/10.1371/journal.pone.0275386
DOI: https://doi.org/10.1371/journal.pone.0275386

Accessibility level

Access to data through SND
Data are freely accessible

Use of data

Things to consider when using data shared through SND

Versions

Version 1. 2022-09-29

Homepage

http://www.lakemedelsverket.se/ai

Contact for questions about the data

Gabriel Westman

gabriel.westman@medsci.uu.se

Download metadata

DataCite

DDI 2.5

DDI 3.3

DCAT-AP-SE 2.0

JSON-LD

PDF

Citation (CSL)

File overview (CSV)

Published: 2022-09-29
Last updated: 2022-10-25