SmartMedicalCodex - ICD-10 Augmented Dataset

Juan Carlos Rivera Domínguez, M.Sc. (@jrivd), Adrian Ramos Cápitas, M.Sc. (@aramcap)

Español

Resumen

Este repositorio contiene un conjunto de datos aumentado, que tiene como objetivo compartir con la comunidad científica el resultado de un gran trabajo realizado en nuestro laboratorio de innovación, y así facilitar a los investigadores conjuntos de datos de valor.

Para abordar las limitaciones de los conjuntos de datos públicos existentes centrados en CIE-10, que son de tamaño reducido, poco representativos y desequilibrados en cuanto a la distribución de clases, hemos desarrollado un sistema de generación de datos sintéticos sobre la información pública existente, con diversas técnicas para mejorar los resultados, para obtener conjuntos de datos de prueba con información de gran fidelidad, proporcionando un gran volumen de datos con una distribución equilibrada de clases y un léxico variado para entrenar los modelos de aprendizaje automático de SmartMedicalCodex.

Contenido

El conjunto de datos aumentado que se comparte aquí contiene la clave de mayor nivel de CIE-10 y un texto relacionado descriptivo de lo que puede contener la clave.

Desarrollo

Para crear este conjunto de datos se utilizaron las descripciones de tres niveles de CIE-10, además de un procedimiento de enriquecimiento con información pública y datos de elaboración propia (apoyado en diversas técnicas de inteligencia artificial), para hacerlo más natural y completo.

Propiedad

RED.es ha subvencionado parcialmente el proyecto SmartMedicalCodex, el cual está catalogado como "Investigación industrial con amplia difusión de resultados".

English

Summary

This repository contains an augmented dataset so the result of a large body of work performed in our innovation lab can be shared with the scientific community, providing researchers with a valuable dataset.

The provided dataset was created to address the limitations of existing public datasets focused on ICD-10, most of which are: small in size, unrepresentative and disequilibrated in terms of class distribution. To create it, we developed a synthetic data generator based on datasets publicaly available and in-house

techniques to improve the results, then we used it to generate test datasets of almost-real quality information, thus providing a large volume of data with balanced distribution of classes and a varied lexicon that we used to train SmartMedicalCodex's machine learning models.

Content

The augmented dataset published here contains the highest ICD-10 level key and a related descriptive text of what the key may contain.

Development

To create this dataset, the three-level descriptions of ICD-10 were used, in addition to an enrichment procedure with public information and self-developed data (supported by various artificial intelligence techniques), to make it more natural and complete.

Property

This development is property of Keedio, although it is released under the Apache 2.0 license.

RED.es has partially subsidized the SmartMedicalCodex project, which is classified as "Industrial research with wide dissemination of results".

KEEDIO