

Creación de *corpus* médico en español de acceso público

Problema:

La cantidad de información registrada en forma digital en la ficha clínica aumenta cada año¹. Estas fichas contienen una gran cantidad de datos², los cuales pueden estar en formato estructurado o no estructurado³. El texto clínico no estructurado puede ser analizado mediante técnicas de procesamiento de lenguaje natural, sin embargo, existen pocas herramientas computacionales desarrolladas para el lenguaje clínico en idiomas distintos del inglés⁴. En particular, existe insuficiente cantidad de texto clínico en español para entrenar algoritmos⁵.

Solución:

Una forma de obtener textos médicos para entrenar algoritmos es por medio de la aglomeración de artículos de revistas médicas de acceso libre en línea, lo cual ha mostrado un buen desempeño en tareas de procesamiento de texto clínico en inglés cuando son comparados con algoritmos entrenados sobre textos de índole general⁶. Esto se denomina *corpus* biomédico, donde por *corpus* entendemos una muestra representativa de texto.

Métodos:

Se utilizó la biblioteca de Python *scrapy* para extraer el texto de todos los artículos disponibles de las revistas médicas chilenas en SciELO. Posteriormente, por medio de la biblioteca *Beautiful Soup*, también en Python, se procesaron los archivos con el fin de obtener un *corpus* funcional.

Variables y métricas:

Dentro de las métricas a evaluar sobre este *corpus* están la cantidad de artículos por revista, la proporción de palabras por especialidad médica, la diversidad léxica y las palabras más frecuentes, tanto por especialidad como aglomerándolas.

Hipótesis:

Se puede crear un *corpus* biomédico en español en base a todas las revistas médicas de Chile en SciELO mediante *web scraping*.

Objetivos:

- 1) Crear un *corpus* biomédico en español en base a todas las revistas médicas chilenas disponibles en SciELO.
- 2) Analizar las métricas propuestas sobre este *corpus*.
- 3) Hacer accesible de manera online este *corpus* para la comunidad.

Resultados preliminares

Se obtuvo un *corpus* biomédico en español en base a todas las revistas médicas de Chile disponibles en SciELO, que abarcan 13 revistas de especialidad. Además, se hizo disponible su acceso para que pueda ser utilizado por desarrolladores y comunidad científica en general.

Perspectiva:

Se espera que con esta herramienta de acceso libre a la comunidad mejore el procesamiento de texto clínico en español.

Referencias:

- 1) Ghassemi M, Naumann T, Schulam P, L. Beam A, Ranganath R. Opportunities in Machine Learning for Healthcare. arXiv preprint arXiv:180600388, 2018.
- 2) Miotto R, Li L, Kidd B, Dudley J. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports. 2016;6(1).
- 3) Dalianis H. Clinical Text Mining ; 2018.
- 4) Névél A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. Journal of Biomedical Semantics. 2018;9(1).
- 5) Chapman W, Nadkarni P, Hirschman L, D'Avolio L, Savova G, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Journal of the American Medical Informatics Association. 2011;18(5):540-543.
- 6) Chen Z, He Z, Liu X, Bian J. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. BMC Medical Informatics and Decision Making. 2018;18(S2).