El dilema de las palabras vacías en la revisión humana del procesamiento automatizado

The stopword dilemma in human review of automated processing

Fernanda Peset

Cómo citar este artículo:

Peset, Fernanda (2023). "El dilema de las palabras vacías en la revisión humana del procesamiento automatizado [The stopword dilemma in human review of automated processing]". *Infonomy*, v. 1, e23011.
https://doi.org/10.3145/infonomy.23.011

Fernanda Peset
https://orcid.org/0000-0003-3706-6532
https://www.directorioexit.info/ficha234
Universitat Politècnica de València
Instituto de Matemática Pura y Aplicada (*IUMPA*)
Camí de Vera, s/n. Edificio 8E
46022 Valencia, España
mpesetm@upv.es

Resumen
Se discute la necesidad de pre-procesar los corpus de términos para eliminar las palabras vacías o stopwords, y se presenta el dilema de hacerlo manualmente o utilizando un sistema automatizado. Se sugiere que los documentalistas trabajen en la construcción de diccionarios y en la creación semiautomática de vocabularios específicos por dominios.

Palabras clave
Palabras vacías; Diccionarios; Pre-procesamiento; Procesos manuales versus automatizados; Rol de los documentalistas.

Abstract
The need to pre-process term corpora to eliminate stopwords is discussed, and the dilemma of doing it manually or using an automated system is presented. It is suggested that librarians-information scientists should work on the construction of dictionaries and the semi-automatic creation of domain-specific vocabularies.

Keywords
Empty words; Stopwords; Dictionaries; Pre-processing; Manual versus automated processes; Weak review; Role of librarians and information scientists.

1. Introducción

En la actualidad, el procesamiento del lenguaje natural se ha convertido en una herramienta multifuncional con aplicaciones exitosas, especialmente en la inteligencia generativa. Estos algoritmos, fundamentales para la creación de sintaxis correctas, requieren previamente "entender" el vasto contenido disponible en Internet.

2. Palabras vacías

De particular interés para documentalistas, o para analistas de comunicación, es la utilización de diccionarios de palabras vacías, también conocidas como *stopwords*. Se asume que la identificación precisa de estos términos es crucial para lograr resultados efectivos en el procesamiento de textos. Es importante destacar que un diccionario no solo comprende palabras vacías "puras", como conectores gramaticales, adverbios e incluso números, sino también sustantivos no significativos. Estos últimos, al ser demasiado generales o frecuentes, carecen de valor para interpretar el dominio que se esté analizando. Por ejemplo, en este mismo texto, el término "*stopword*" no agrega información relevante cuando ya se está hablando de *stopwords*, al igual que la palabra "fines", que carece de aportación específica en este mismo texto.

3. Diccionarios

Comencemos desde el principio. Existen múltiples diccionarios de palabras vacías, gratuitos o no, en diversos idiomas. Podemos asumir que, para la lengua inglesa serán más numerosos que para el castellano. En cualquier caso, todos los diccionarios que hemos explorado en nuestro idioma han requerido enriquecimiento desde el inicio. Por ejemplo, muchos no incluían la enumeración de meses, días o símbolos como el "%". La omisión de estos elementos puede afectar a la calidad de los resultados al procesar el corpus. Por tanto, se vuelve esencial realizar un preprocesamiento de los textos, adaptado a cada dominio de estudio.

4. Pre-procesamiento

El procedimiento inicia con la elección de un diccionario estándar de palabras vacías. Posteriormente, se procesa el corpus documental para identificar los términos más frecuentes. Establecido un umbral mínimo de frecuencia, el investigador debe identificar aquellos términos que podrían distorsionar los resultados. Estas palabras, símbolos o números, se añaden al diccionario de palabras vacías inicial. A partir de este punto, comienza el procesamiento final, donde se logra una adaptación más precisa a las particularidades del dominio en estudio.

Visto este trabajo semi-manual, nos preguntamos, ¿cuál es la cantidad ideal, equilibrada, de supervisión humana en estos procesos?

En estos momentos, la automatización de procesos que tradicionalmente dependían de la intervención humana se presenta como una panacea. De hecho, la intención es tender a disminuir al máximo la supervisión humana (**Burns** *et al.*, 2023). Suena bien. El objetivo es interesante y alcanzable. De hecho, en un trabajo anterior (**Blasco-Gil** *et al.*, 2020) probamos que el procesamiento con *KH Coder*[1] ofrecía unos resultados similares al análisis de los expertos para entender las actas de claustros universitarios del siglo XVI. Trabajar con la transcripción del castellano antiguo fue un gran desafío, por la variabilidad de las palabras, las alteraciones aleatorias en la grafía… El trabajo previo con las *stopwords* fue exhaustivo, pero los resultados probaron su validez cuando se utilizó una parte del corpus específica. En cualquier caso, los mejores análisis se lograron con el uso de Mapas Auto-Organizados (SOM), modelo de neurona artificial con aprendizaje no supervisado descrito por primera vez por el finlandés Kohonen[2].

5. Entre calidad y rapidez

Este tipo de experiencias nos alertan sobre un cambio potencial en el equilibrio entre calidad y rapidez. La supervisión de los algoritmos y de la inteligencia artificial no solo debe centrarse en

su transparencia o su ética (**Calabuig** *et al.*, 2023) sino también en encontrar la justa medida entre eficiencia y calidad de los resultados.
Este dilema destaca la importancia de que los documentalistas desempeñen un papel crucial en la construcción, no solo de estos diccionarios, sino también en la creación semiautomática de vocabularios específicos por dominios. Estos vocabularios son fundamentales para ayudar a las inteligencias artificiales a "comprender" de manera más precisa y contextual lo que se comunica. En este sentido, la búsqueda de un proporción cuidadosa entre eficiencia y calidad se convierte en una tarea esencial para garantizar que la automatización no sacrifique la fiabilidad en la interpretación del lenguaje.

Notas

1. *https://khcoder.net/en*
2. *https://es.wikipedia.org/wiki/Mapa_autoorganizado*

6. Referencias

Blasco-Gil, Yolanda; González, Luis M.; Pavón-Romero, Armando; Mercado-Estrada, Mariano; Pavón-Romero, Carlos; Cabrera, Ana M.; Garzón-Farinós, Fernanda; Peset, Fernanda (2020). "Enriqueciendo la investigación en humanidades digitales. Análisis de textos de claustros académicos de la Universidad de Valencia (1775-1779) con KH Coder". *Revista española de documentación científica*, v. 43, n. 1, e257.
https://doi.org/10.3989/redc.2020.S1

Burns, Collin; Izmailov, Pavel; Kichner, Jan H.; Baker, Bowen; Gao, Leo; Aschenbrenner, Leopold; Chen, Yining; Ecoffet, Adrien; Joglekar, Manas; Leike, Jan; Sutskever, Ilya; Wu, Jeff (2023). *Weak to strong generalization: Eliciting strong capabilities with weak supervision*.
https://cdn.openai.com/papers/weak-to-strong-generalization.pdf

Calabuig, José-Manuel; Ferrer-Sapena, Antonia; Garcia-Raffi, Lluís-Miquel; Peset, Fernanda; Sánchez-Pérez, Enrique A.; Sánchez-Del-Toro, M. Isabel (2023). "Algoritmos matemáticos para una inteligencia artificial responsable, ética y transparente". *Revista Valenciana d'Estudis Autonòmics*, n. 68, pp. 283-305.
https://presidencia.gva.es/es/web/begv-gavina/politica/-/asset_publisher/MBYQ47LTEnde/content/revista-valenciana-d-estudis-autonomics