Análisis de textos y contenidos semánticos con asistencia computacional e inteligencia artificial

José Carlos Machicao
GestioDinámica©

Todos usamos el lenguaje para comunicarnos. Sin embargo, la complejidad de la comunicación se ha elevado tanto en estos días que se hace difícil llevar la cuenta de las ideas que debemos procesar o las decisiones que debemos tomar ante un cúmulo de información que recibimos. Muchos interlocutores esperan que sus ideas sean recibidas por nosotros y a veces nosotros llegamos a niveles de saturación que nos hacen pensar en respuestas que si prestamos atención son comunes en cualquier puesto de trabajo: "es que yo no soy un robot para responder tan rápido".

Bueno, qué pasaría si tuviéramos un robot que pudiera, no vamos a decir, responder todo, pero al menos ayudarnos a organizar las ideas y la información que recibimos o tenemos que emitir en nuestro lenguaje. Esa capacidad de ayuda ya no es ciencia ficción, es una realidad.

Este curso está orientado a experimentar con la capacidad computacional para el procesamiento de lenguaje natural, aprendiendo de la experiencia ya vivida por expertos en el tema, pero también abriendo la posibilidad de reflexionar respecto al lenguaje a medida que se aplican las herramientas existentes.

1 Definiciones básicas del procesamiento de lenguaje natural

El lenguaje natural es aquel que usa un ser humano a través del uso libre de la palabra o la escritura. La forma de emisión son la voz o los textos. Las tablas de números no son lenguaje natural. Un gráfico o una imagen no son lenguaje natural. Un código computacional no es lenguaje natural. Este párrafo escrito o su lectura en voz alta sí es lenguaje natural. A la disciplina que estudia cómo procesar computacionalmente el lenguaje natural se le llama procesamiento de lenguaje natural (PLN, o por sus siglas en inglés NLP[1]).

- **Documento**: Es una unidad de análisis. Puede ser un párrafo o un documento entero como un libro. Generalmente es la unidad que se procesa.
- **Oración**: Es la oración gramatical en un idioma.
- **Palabra**: Es la palabra gramatical en un idioma.
- **Token**: Es una unidad de análisis dentro de un documento, en la mayoría de aplicativos puede ser una oración o una palabra, pero también pueden ser, por ejemplo, pares o tríos de palabras.
- **Concepto**: Expresión de una idea comprensible dentro de un documento o corpus.
- **Corpus**: Es el conjunto total de documentos a analizarse. Generalmente un análisis determinado tiene un solo corpus. Si es complejo puede tener varios corpora.

[1] Las siglas NLP se usan para Natural Language Processing o para Neuro Linguistic Programming. En este texto será siempre el primer significado

En el desarrollo de la experimentación de aplicación de instrumentos, no debe descartarse el encontrar otro tipo de unidades de análisis. El concepto, por ejemplo, o la idea compleja, podría servir para poder concluir algo respecto a un documento o a un corpus. En general los conceptos iniciales son solo una forma inicial de ordenar cómo entender los sistemas de comunicación, y la inteligencia artificial aplicada a ellos debería abrir puertas a ensayos de nuevas unidades, teniendo como norte la mayor comprensión y transmisión de las ideas.

2 Demostración de instrumentos prácticos directos de análisis de textos

Tres aplicativos muy demostrativos de cómo la inteligencia artificial y los algoritmos están ayudando a procesar el lenguaje natural son: (i) el dictado por voz de Google Docs, (ii) el narrador de voz NaturalReader, y (iii) la versión libre de MeaningCloud. Se puede experimentar directamente el narrado o registro de voz, poniendo énfasis en las imperfecciones que puede tener los algoritmos tanto para saber sus limitaciones como para pensar en qué soluciones se puede dar a estas limitaciones para que podamos contar con textos posteriormente analizables. Por ejemplo, qué hacer para ganar detección de la puntuación o cómo aprovecharlo cuando se está usando en una reunión. En cuanto al análisis de textos, se puede experimentar con textos conocidos haciendo comparaciones entre el análisis con algoritmos y el análisis manual que se pueda hacer. Esto puede ayudar a tener una idea clara de la potencia de cada herramienta de análisis. En esta introducción se cubre sólo dos herramientas.

3 Introducción a Python para el análisis computacional de textos

Python© es un lenguaje de programación que permite hacer prácticamente lo mismo que hacen las hojas de cálculo como Google Spreadsheets o Microsoft Excel, pero sin necesariamente tener que ponerlo en un formato de casilleros en filas y columnas, y con un potencial computacional mucho mayor. Con Python se puede hacer cosas como buscar palabras con patrones definidos por el usuario, de modo que se pueda saber, por ejemplo, qué textos contienen cierta palabra y en qué lugar está esta palabra, pero también puede encontrar todas las palabras que obedezcan a cierto patrón (por ejemplo, todas las palabras que empiecen en mayúscula y estén juntas). Esto da una gran flexibilidad para búsquedas en un gran volumen de documentos o párrafos. Otras librerías más bien convierten un texto en una estructura completamente rotulada por sus valores gramaticales y obviamente otras librerías permiten vectorizar textos para analizarlos con inteligencia artificial.

4 Estructuración de información natural con inteligencia artificial

Los algoritmos de inteligencia artificial permiten hacer aplicaciones prácticas para el análisis de textos. Algunas de las más conocidas son:

- **Web Scraping**: Navegación automática por páginas web extrayendo documentos o contenidos de los documentos o páginas web para estructurarlos.
- Extracción de tópicos: Permite identificar las ciudades o entidades, o personas con nombre propio, pero también valores monetarios, o cantidades o expresiones de tiempo. Todas estas son identificables por patrones o por inteligencia artificial.
- **Análisis de sentimiento**: Permite identificar si un texto es positivo o negativo en su sentido, o si es objetivo o subjetivo, o si es concreto o irónico. Muchos de estos algoritmos todavía están en experimentación, pero es factible explotarlos si se conoce su potencial.

- **Clasificación de textos**: Permite clasificar párrafos o frases en función a sus contenidos semánticos, combinados a veces por la presencia de vocabularios específicos.
- **Identificación de afinidades**: También conocido como "text clustering" permite identificar las frases o conceptos coincidentes

5 Visualización de estructuras de textos

¿Un texto se puede ver de una manera gráfica? La capacidad computacional de hacer estadísticas más o menos de gran volumen en un corto tiempo (o con un esfuerzo computacional más o menos reducido) brinda la posibilidad de graficar esta estadística con las mismas herramientas con las que se grafica los cambios en el clima o la evolución de la economía. Cuando los textos no son demasiado largos es posible utilizar visualización estadística, o numérica de características numéricas, ordinales o de vinculaciones entre los textos. Esto involucra herramientas de conteo, caracterización manual de algunos elementos como vocabularios previamente definidos o diagramas de red.

6 Transcripción de audios

Un contenido de voz es una sucesión consistente de fonemas. Durante muchos años la inteligencia artificial tenía muchos problemas interpretando sonidos hasta que finalmente logró vincularlos al contexto. Hoy en día los algoritmos de transcripción alcanzan precisiones superiores al 70% en casi cualquier acento en cualquier idioma.

El algoritmo con el que vamos a experimentar se llama Google Speech-to-Text y se puede descargar de la plataforma de Google Cloud. Lo más importante es saber en qué formato de archivo de audio se necesita preparar y acceder a una cuenta de GCP[2]. Además, es posible hacerlo desde muchas otras plataformas como AWS, IBM o Azure.

7 Preparación de documentos para procesamiento

No siempre es fácil acceder a la información contenida en un documento de texto o de voz. Los principales problemas surgen de la imperfección del registro de muchos documentos. Por ejemplo, una de las imperfecciones más típicas es el defecto en el escaneo o los errores de escritura (tipeo) de alguna palabra. Existe la posibilidad del uso de algoritmos tanto usando simples reglas o algoritmos con capacidad de inteligencia artificial para identificar patrones de defectos lo cual puede ayudar a transformar un texto desestructurado en un texto más estructurado que sea más analizable. La experimentación de estos métodos se hace más importante en contextos en los que los contenidos registrados obedecen a patrones muy diversos de registro como estilos diversos de redacción.

8 Vectorización de documentos, frases y palabras

Un texto escrito en lenguaje natural puede expresarse en números. Por ejemplo, si tenemos los siguientes textos:

- "Las calles están vacías, aparentemente la gente ha salido de vacaciones."
- "Se nota que hay vientos intensos hacia el norte, no sé si debamos seguir navegando."
- "No hay signos de personas en este lado de la ciudad, será que están descansando."

Una forma de "vectorizarlos" de una manera muy sencilla es por el número de palabras que tienen. Por tanto los vectores serían 11, 15 y 15. Si además agregamos sus verbos como componente del vector entonces los vectores serían: [11, 3], [15, 6], [15, 4]. De este modo se puede

[2] Google Cloud Platform

ir agregando componentes a un vector hasta caracterizar una frase de manera única. Pero qué ocurre cuando hay un número muy grande de frases que pueden parecerse mucho. Entonces es necesario recurrir a modelos más sofisticados de vectorización masiva y que al mismo tiempo permita denotar la similaridad o diferencia entre dos frases o más. A estas técnicas se les llama vectorización de textos. La técnica más popular el Word2Vec que hoy es aplicable en cualquier contexto y tiene alta capacidad de explicabilidad, a diferencia de otras más modernas como BERT que requieren mayor capacidad computacional y no tienen tanta explicabilidad.

9 Extracción de contenidos semánticos complejos

Tres aplicaciones regularmente expandidas en el campo técnico pueden servir de ejemplo para acercarse más a aplicaciones más integrales o cercanas a problemas complejos:

- **Estructura Semántica y Gramática**: Uno de los aspectos que se puede analizar en textos de investigación o textos universitarios o escolares es la idoneidad de redacción. El análisis hoy en día está bastante disponible con herramientas completas y encuentra sus aplicaciones prácticas en análisis masivo de documentos.
- **Sumarización**: A menudo se requiere tener idea de los temas principales de un documento. Existen muchos instrumentos libres para resumir textos, sin embargo, se puede diversificar los enfoques a través de los cuales se quiere resumir. Su aplicación práctica alcanza los puestos de trabajo con alta diversidad de documentos.
- **Comparación de textos**: La clasificación de frases o textos brinda la oportunidad de sofisticar la aplicación y ser capaz de comparar las distancias semánticas entre dos documentos o cualquiera de los elementos semánticos, en especial a través de la vectorización.

10 Cierre

Dados los temas anteriores explicados es casi ineludible deducir que lo que se puede hacer con textos hoy aprovechando la capacidad computacional y más aún la inteligencia artificial y las diferentes arquitecturas de redes neuronales es prácticamente infinito. Una de las restricciones de una introducción a un conjunto basto de herramientas es que durante el aprendizaje sólo se podrá cubrir determinados espacios de un problema, sin embargo, será posible discutir y poner en práctica al menos ciertos tramos de metodologías más integrales. El objetivo de un curso como este estaría cumplido si se logra abrir las posibilidades de aplicación y poner en manos de más gente la aplicación de técnicas que ayuden a elevar la calidad de los productos con ayuda de la inteligencia artificial y de la capacidad computacional en general.