Generador automático de contenido.

Authomatic generator content.

Autor 1: Esteban Sánchez López

*Departamento de sistemas, Universidad Tecnológica de Pereira, Pereira, Colombia*

Correo-e: esteban.sanchez@utp.edu.co

***Resumen*— En el presente documento se pretende realizar un trabajo de investigación acerca del área de generación automática de contenido, la cual pasa bastante desapercibida pero que posee elementos bastantes interesantes y útiles en el conocimiento y acercamiento a la inteligencia artificial.**

**Estudiaremos métodos y herramientas usadas en el área, tales como scraping, y generaremos un modelo-diseño del como creemos que estaría compuesto dicho generador.**

***Palabras clave—* Automático, Generador, Inteligencia artificial, Scraping.**

***Abstract*— This paper intends to carry out a research project about the area of ​​automatic content generation, which goes quite unnoticed but has quite interesting and useful elements in knowledge and approach to artificial intelligence.**

**We will study methods and tools used in the area, such as scraping, and generate a model-design of how we believe that generator would be composed.**

***Key Word* — Automatic, Generator, Artificial Intelligence, Scraping.**

1. INTRODUCCIÓN

A medida que pasa el tiempo, en nuestra época tecnológica, surgen grandes cantidades de información , la cual se vuelve cada vez más difícil de procesar dicha información, labor que por parte humana es demasiado tediosa y requiere enormes cantidades de tiempo, específicamente para llegar a escribir algo como un artículo, es necesario tener en cuenta mucha información acerca del tema de este y realmente para una persona llegar a leer tantos escritos e información sobre el tema buscado para poder desarrollar el artículo en condiciones requiere demasiado tiempo ( incluyendo el tiempo de búsqueda).

* Para resolver este problema se creó una rama de la inteligencia artificial que se enfocaría en desarrollar una manera de generar artículos de manera automática, por medio de algoritmos y técnicas que ahorraran tiempo a las personas.
* En este documento describiremos el proceso que tiene un generador automático de textos, con cada uno de los posibles módulos que puede tener.
* Describiremos cada uno de los “Módulos”, funcionamiento y el objetivo de este, con el fin de unirlos y tener la idea de funcionamiento de un generador de contenido por medio de inteligencia artificial.

1. CONTENIDO

Para entender cómo funciona un generador automático de contenido, vamos a descomponerlo en distintos módulos, los suficientes para describir cada una de las áreas o procesos que realiza para obtener un resultado.

Los módulos están nombrados de manera ordenada de acuerdo la ocurrencia de estos, desde el que ocurre inicialmente hasta el último. Los módulos en este caso son:

1. Scraping o extracción de información.

En primer lugar, se realiza el proceso de extracción de conocimiento, para lo cual, teniendo una lista de páginas web, referencias que vamos a usar para sacar la información. Luego, se usa el código HTML de cada una de las paginas para obtener la información que vamos a almacenar (Se pueden presentar casos en los que aparte de HTML las paginas contengan javascript, para lo cual se usan API´s tales como Selenium Web Driver, que simulan el comportamiento de un usuario para llegar a la información que está más oculta).

1. Procesamiento de la información.

En este módulo se realiza un proceso de selección e identificación de toda la información obtenida en el módulo de scraping. En este módulo se realizan algunos subprocesos con la información obtenida.

Estos subprocesos son:

1. Inicialmente la separación de la información obtenida en párrafos o renglones.
2. Luego se realiza un proceso de separación en “Tokens”, que consiste en separar una por una cada una de las palabras que conforman los renglones y párrafos.

Aquí se pueden presentar varios problemas pues hay que tomar en cuenta los errores que puede tener el documento o elementos que no aportan nada al proceso, estos errores pueden ser:

* Errores de ortografía.
* Los signos de puntuación.
* Teléfono, números y correos electrónicos.
* Y una de las más importantes son las expresiones multi-palabras. (más adelante se explica mejor que se hace con estas expresiones).

1. Identificaciones y almacenamiento de las palabras con más peso o las que por sí solas poseen un significado

sobre algo relevante para tener en cuenta (Generalmente son las palabras del tipo verbos o bien los sustantivos).

1. Eliminar las palabras “sin sentido”, es decir, las que no aportan información o contenido (Generalmente estos son pronombres, conjunciones o preposiciones).
2. Durante la eliminación de palabras se toman en cuenta conjunto de palabras que pueden cambiar significado o la intención al combinarse con las palabras de su alrededor, como mencionamos anteriormente en los errores comunes, las multi-palabras son bastante importantes pues le dan el significado correcto de acuerdo con la situación necesitada, por ejemplo, la palabra Madrid, Cambia de significado al agregar la palabra “Real” tras ella, pasa de referirse a una ciudad a un equipo de futbol.

Esto es muy importante para evitar confusiones y evitar el procesamiento de información irrelevante que solo generarían un costo adicional.

1. Definición de Temas y subtemas.

Una vez obtenemos todas las palabras relevantes de la información, procedemos a definir los temas mencionados en las páginas web, tomando en cuenta la definición de las palabras y de acuerdo con cuáles son los temas más mencionados.

Los títulos son las partes que generalmente más ayudan pues en estos se mencionan las palabras que más aportan a la hora de definir un tema.

Para la definición de subtemas, tomamos todas las palabras o frases seleccionadas como temas y luego para cada una de las palabras o frases restantes, tomamos su significado y lo relacionamos con el tema que tenga más similitudes.

1. Definición de diccionarios.

Los diccionarios son el conjunto de palabras o frases que pueden ser mencionados por un tema o subtema y aportan algo a este. Son el conjunto de palabras que entraron como temas o subtemas pues no son tan complejas pero que pueden tener relación con uno-muchos temas o subtemas. Por ejemplo, la palabra Barcelona, puede entrar en diferentes temas de articulo como pueden ser deportes y política.

1. Estructuración del artículo.

Una vez terminado el proceso de extracción y clasificación de la información, continuamos a la estructuración del nuevo documento usando dicha información, para esto, empezamos con un título, en este caso, se puede requerir la ayuda humana, debido a que este puede acarrear problemas legales con otras páginas web, por lo que requerirá de una revisión más cuidadosa.

Para el resto, la maquina tomara por si misma los fragmentos previamente almacenados de acuerdo con el tema seleccionado, usando los diccionarios para relacionar las palabras que tienen relación y así colocar el párrafo o renglón al cual corresponde dicha palabra.

Inicialmente, tomara en cuenta las palabras que tiene relación única y exclusivamente con el tema elegido, pero luego, realizara adiciones de párrafos o renglones que tienen relación con el tema, pero no es una relación exclusiva, así se añadirán matices de otros temas sin salirse del tema primordial.

1. Finalización del artículo.

Una vez terminado el grueso del documento, se realiza la adición de las partes finales, estas son las referencias, que previamente son guardadas en el módulo de extracción y almacenamiento, además de también añadir la bibliografía (todas las páginas de las que se extrajo la información usada, no se incluirá aquella que se haya guardado, pero no usado en el artículo.

1. Etapa de revisión y publicación.

Como último paso, se hará una revisión por parte humana, en la que se revisara el documento entregado por la máquina, donde se realizara correcciones, pero principalmente se enfocaran en revisar que el documento tenga coherencia con el tema y entre los distintos párrafos escritos, tomando en cuenta que no haya información irrelevante.

1. CONCLUSIONES

* La inteligencia usada puede llegar a aprender sobre las páginas que utiliza, llegando a identificar en que página puede hallar un tema.
* Este tipo de inteligencia facilitan en gran medida el

BIBLIOGRAFIA

* <http://sedici.unlp.edu.ar/bitstream/handle/10915/74098/Tesis.pdf-PDFA.pdf?sequence=1&isAllowed=y> Tesis sobre generación automática de resúmenes.