Extraer palabras clave de un texto corto

El concepto de extraer palabras claves de un texto corto se puede lograr mediante el uso de machine learning. De acuerdo al autor del texto “Extracting Keywords From Short Text”, tomo tres opciones distintas como soluciones:

1. Estilo Scrabble: Que es tomar la oración y calcular cada valor de cada palabra como en el juego de Scrabble y se toman las mayor puntuadas como las mas importantes. Luego se toman las palabras de la siguiente oración y se compara con las anteriores para buscar aquellas palabras relacionadas o repetidas. En este caso, las palabras con mayores puntajes tienden a ser las más importantes y las que más se repiten excepto por ciertos casos; para evitar esto se implementa una lista que contenga este tipo de palabras consideradas como excepciones y no contarlas en el ranking. Al finalizar, el ranking debe de tener una lista de puntajes en el cual las palabras más importantes estarían de primeras.
2. Método RAKE (Rapid Automated Keyword Extraction): Este método consiste en eliminar las palabras más comunes en el idioma como conectores entre otros para al final dejar las palabras que se podrían considerar no comunes o las importantes del texto. En este caso no se garantiza perfectamente que sean las palabras más importantes del texto corto. Por último, RAKE se encuentra implementado en varios lenguajes de programación lo cual lo hace bastante simple y sencillo de usar.
3. Inventar un nuevo Lenguaje: En este caso sería crear un lenguaje mediante el uso de palabras importantes como secuencia y hacer que la maquina aprenda a reconocerlas mediante una base de datos, tomándolas y creando un nuevo lenguaje basándose en ellas. Primero es necesario generar un set de datos, en el cual se pueda relacionar las palabras claves a algún tema y así empezar a categorizarlas. Luego, se cargan los datos del programa en la red neuronal con una estructura base. Para continuar, se ejecuta la red neuronal y se deja entrenando la secuencia de secuencias hasta que termine, al terminar la rede neuronal ya debería de tener la capacidad de identificar algunas o varias palabras claves dentro del texto que se ingrese. Para mejores resultados es preferible que se entrene la red neuronal con una cantidad de datos mayor.

<https://towardsdatascience.com/extracting-keywords-from-short-text-fce39157166b>

1. Consulta Mas Como Esto o More Like this Query, consiste en encontrar documentos parecidos en base a un conjunto de documentos mediante la búsqueda de términos similares dentro de los documentos. Para lograr este proceso, se seleccionan una cantidad representativa de términos en el documento ingresado, y luego se genera una petición de búsqueda en la base de datos para encontrar textos que contengan términos similares. Para que esta consulta funcione, primero se debe de clasificar el texto ingresado dentro de un tipo de consulta, en este caso se utiliza como método la formula de puntaje de Lucene (buscar similitudes dentro del texto, asignarles un puntaje y comparar donde pertenece). Luego, se toma el texto del documento ingresado y se analiza para poder extraer n cantidad de términos clave y utilizarlos en la consulta para la búsqueda de textos similares. Al finalizar el proceso, se retorna un conjunto de documentos similares al que se ingresó.

<https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html>