

# Moogle!

## Amalia Beatriz Valiente Hinojosa C-112

Facultad de Matemática y Computación, Universidad de La Habana

- ¿Qué es Moogle!?
- Estructura de Moogle!
- Manual de búsqueda.

# ¿Qué es Moogle!?

Moogle! es una aplicación cuyo propósito es buscar inteligentemente y de forma eficiente un texto en un conjunto de documentos.

Para su funcionamiento fue emplado el modelo de espacio vectorial, un modelo algebraico utilizado, entre otras cosas, para el cálculo de relevancia de información otorgando cierto valor de similitud entre los documentos y la consulta.

Para la construcción del buscador se elaboraron cuatro clases:

- Moogle.
- Processor.
- TF-IDF
- Query.

## Moogle:

En esta clase se encuentra el método principal, Query, este método devuelve los documentos resultantes de la búsqueda.

## Processor:

Esta clase contiene métodos encargados de:

- Normalizar los documentos.
- Separar y almacenar los documentos en diccionarios.

## Query:

En esta clase se encuentran los métodos encargados de:

- Normalizar el contenido de la query eliminando caracteres especiales como tildes, etc.
- Calcular el peso de las palabras de la query mediante la fórmula:

$$QW_i = (b + (1 - b) \cdot \frac{freq_i}{maxfreq}) \times idf_i \quad (1)$$

Donde:

**QW** (i) = peso de la palabra (i) de la query.

**b** = término de suavizado, cuyo valor es 0.5.

**freq** (i) = cantidad de veces que se repite el término (i).

**maxfreq** = el valor del término que más se repite en la query.

**idf** = la frecuencia inversa del término (i) de la query.

## TF-IDF:

Esta clase contiene los métodos encargados de:

- Calcular el peso de las palabras de los documentos mediante la fórmula del tfidf:

$$W_{t,d} = tf_{t,d} \times idf_t \quad (2)$$

Donde:

- $w(t,d)$  = peso del término  $t$  en documento  $d$ .



- Continuando con el peso de las palabras:

- 

$$tf_{t,d} = \frac{freq_{t,d}}{maxfreq_d} \quad (3)$$

Donde:

- **tf** (t,d) = cantidad de veces que se repite el término t en el documento d.
- **maxfreq** (d) = el valor del término que más se repite en el documento d.

- 

$$idf_t = \log_{10}\left(\frac{N}{n_t}\right) \quad (4)$$

Donde:

- **idf** (t) = frecuencia inversa del término t.
- **N** = cantidad de documentos en la base de datos.
- **n** (t) = cantidad de documentos en los que aparece el término t.

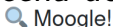
- Hallar el valor de similitud que hay entre los documentos y la query mediante la fórmula de similitud de coseno:

$$sim_{d,q} = \frac{\vec{d} \cdot \vec{q}}{|\vec{d}| \cdot |\vec{q}|} \quad (5)$$

El denominador es el resultado de multiplicar la norma de los vectores documento y query.

- Almacenar en un diccionario los documentos ordenados en función de la similitud.
- Encontrar una porción del texto resultante donde sale al menos una palabra de la query, a esto se le llama Snippet.

Para efectuar una consulta en Moogle! basta con escribir fragmentos o el nombre del documento que se quiere encontrar y dar click en el botón azul que se encuentra a la derecha de la barra de consulta.

¿Quisite decir [azul](#)?

- Neruda, Pablo - Canto General.txt  
... Todas las águilas del cielo nutrían su estirpe sangrienta 20 en el azul inhabitado, y sobre las plumas carnívoras volaba encima del mundo el cóndor, rey asesino, fraile solitario del cielo, 25 talismán negro de la nieve, huracán de la cetrería La ingeniería del hornero hacía del barro fragante pequeños teatros sonoros 30 donde aparecía cantando ...
- El corazon delator.txt  
... ¡Era uno de sus ojos, sí, esto es! Se asemejaba al de un buitre y tenía el color azul pálido Cada vez que este ojo fijaba en mi su mirada, se me helaba la sangre en las venas; y lentamente, por grados, comenzó a germinar en mi cerebro la idea de arrancar la vida al viejo, a fin de librame para siempre de aquel ojo que me molestaba ...
- NERUDA, Pablo - Residencia en la tierra.txt  
... Sus copas duras cubren tu alma derramada en la tierra fría con sus pobres chispas azules volando en la voz de la lluvia 20 [54] Colección nocturna He vencido al ángel del sueño, el funesto alegórico: su gestión insistía, su denso paso llega envuelto en caracoles y cigarras, marino, perfumado de frutos agudos ...