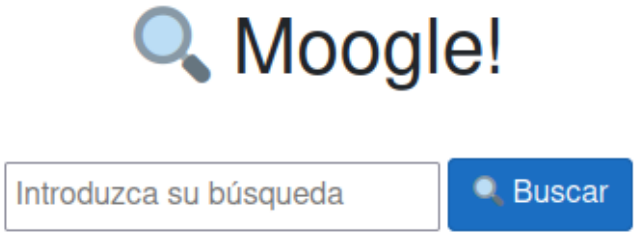
Informe Escrito
Proyecto: Moogle!

Amalia Beatriz Valiente Hinojosa
C-112.

Universidad de la Habana:
Facultad de Matemática y Computación.



1 ¿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 algebráico utilizado, entre otras cosas, para el cálculo de relevancia de información otorgando cierto valor de similitud entre los documentos y la consulta.

2 Funcionamiento:

El modelo de espacio vectorial se divide en una serie de pasos que permite encontrar los documentos con mayor similitud con la consulta que se realice. El primer paso tenido de cada cuya estructura consiste en guardar el condocumento en un diccionario es de la siguiente forma:

```
2 references
public static Dictionary<string, string[]> titleAndContent = Converted(FilesPath, FilesContent);
```

Una vez guardado el contenido de cada documento, se halla la fecuencia de cada término de cada documento (Term Frequency), que es la primera parte necesaria para calcular el peso de cada palabra.

Para calcular la frecuencia de cada término se emplea la siguiente fórmula:

$$tf_{t,d} = \frac{freq_{t,d}}{maxfreq_d} \tag{1}$$

Donde:

- **tf** (t,d) = la frecuencia del término t en el documento d.
- **freq** freq (t,d) = la cantidad de veces que se repite el término t n el documento d.
- **maxfreq** (d) = el término del documento que más se repite en este.

Ahora, los documentos con el tf de cada palabra pasarán a estar almacenados en un nuevo diccionario:

```
2 references
public static Dictionary<string, Dictionary<string, float>> documentsTf = tfFull(Processor.titleAndContent);
1 reference
```

El segundo paso para encontrar el peso de cada palabra es calcular la frequencia inversa de cada palabra (Inverse Document Frequency) que se halla mediante la fórmula:

$$idf_i = \log_{10}(\frac{N}{n_i}) \tag{2}$$

Donde:

- **idf** (i) = frecuencia inversa del término i.
- **N** = cantidad de documentos que se encuentran en la base de datos.
- **n** (i) = cantidad de documentos en los que se encuenta el término i.

Esta nueva información se almacena en un tercer diccionario, para posteriormente hallar el peso de cada palabra de los documentos al calcular:

$$W_{t,d} = tf_{t,d} \times idf_t \tag{3}$$

Donde:

- **W** (t,d) = peso del término t en el documento d.

Para calcular el peso de las palabras de la query se usa la fórmula:

$$QW_i = (b + (1 - b) \cdot \frac{freq_i}{maxfreq}) \times idf_i \tag{4}$$

Donde:

- **QW** (i) = peso del término i de la query.

Una vez que se obtiene el peso de cada palabra, tanto de los documentos como de la query, se procede a evaluar el valor de la similitud entre los documentos y la query usando la fórmula de similitud del coseno:

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} \tag{5}$$

- el denominador es la multiplicación de la norma de los vectores documento y query.

El peso de cada palabra de los documentos y de cada palabra de la query se almacena en los diccionarios:

```
1 reference
public static Dictionary<string, Dictionary<string, float>> documentsTfIdf = TFxIDF(documentsTf,
documentsIdf);
```

y:

```
Dictionary<string, float> queryTfIdf = QueryClass.QueryTFIDF(queryTf, TFIDF.documentsIdf);
```

Una vez se obtiene la simlitud del coseno, cada documento es almacenado en un diccionario ordenado de menor a mayor valor, y, como resultado de la búsqueda, se develven los tres últimos documentos, que son los que tienen mayor similitud con la query.

Además de los documentos, también se devuelve un fragmento del documento en donde aparece al menos una palabra de la query, a este fragmento se le denomina **Snippet**.

3 Estructura

Este proceso fue llevado a cabo en 4 clases:

- **Processor:** En esta clase se normalizan los documentos y se separan y almacenan en diccionarios.
- **Query:**En esta clase se normaliza el contenido de la query y se calcula el peso de las palabras contenidas en ella.
- **TF-IDF:** En esta clase se halla el peso de las palabras de los documtos, la similitud entre los documentos y la query y el snippet.
- **Moogle:** En esta se encuentra el método principal que devuelve el resultado de la búsqueda.

4 Manual

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.