Universidad de Oviedo

Prácticas de laboratorio

Sesión 3

Sistemas de Información para la Web

Similitud léxica entre textos y detección de documentos cuasi duplicados

Omar Teixeira González, *UO281847*

26/10/2023

Tabla de contenidos

UO281847

Documentación sobre *SimHash*

Descripción

Práctica 3, correspondiente a *SimHash*, que se basa en la búsqueda de *cuasi-duplicados* de un archivo de documentos, para su posterior filtrado y realizar consultas en esos documentos, utilizando la medida de similitud asignada (en mi caso Coeficiente de Solapamiento).

Forma de uso

Para utilizarlo es necesario el paso de 5 parámetros, los cuales son:

- *documents_file*: nombre del archivo donde se encuentran los documentos con los que se va a realizar la detección de *cuasi-duplicados* y de donde buscarán las consultas
- *queries_file*: nombre del archivo donde se encuentran las *queries* con las que se van a realizar las consultas.
- *k*: longitud de los *shingles* a generar.
    - valor por defecto: 3.
- *f*: dimensión de las firmas, a mayor número de documentos se recomienda que el valor sea 128.
    - valor por defecto: 64.
- *tokenizer_name*: Tokenizador que se va a utilizar tanto para los documentos como para las *queries*.
    - valor por defecto: *WhitespaceTokenizer*.

Medida de similitud a utilizar

La medida de similitud a implementar dependerá de tu número de DNI. Coge el número y calcula su módulo 4. En función del resultado implementarás la siguiente medida: 0 = Dice, 1 = Jaccard, 2 = Coseno, 3 = Solapamiento.

$$32892095 \ mod \ 4 = 3 \rightarrow Coeficiente\ de\ solapamiento: \frac{|\boldsymbol{X} \cap \boldsymbol{Y}|}{\mathbf{min}\ (|\boldsymbol{X}|, |\boldsymbol{Y}|)}$$

Decisiones tomadas

Para la realización de la práctica se ha empleado un *tokenizador* que podrá ser seleccionado por el usuario (por defecto está el *WhitespaceTokenizer*, como ya se ha mencionado). Posteriormente, tanto los documentos como las consultas serán pasados a minúsculas, además de la *tokenización* ya realizada.

Un ejemplo de cómo serían los documentos o las queries tras la *tokenización* y vectorización es el siguiente:

Documentos:

```
{
 (t980, [a, man, was, shot, dead, and, fifteen, others, injured, when, zambian, policemen, clashed, with, ...]),
 (t1088, [russian, prime, minister, viktor, chernomyrdin, on, thursday, proposed, a, three-phase, solution, ...]),
 ...
}
```

Queries:

```
[
 [bangladesh, anti-government, strike],
 [baseball, players, legal, victory],
 ...
]
```

U028184?