PRÁCTICA 4 - SIMHASH

Sistemas de información para la Web

Antonio Payá González

UO251065

Para la realización de la función de simhash se ha tenido en cuenta la posibilidad de utilizar ngramas, por medio de la librería NLTK.

Además, se ha añadido un archivo llamado *score_to_csv.py* en el que se ejecuta el archivo *score_simhash_implementation.py* con diferentes valores de ngramas y de restrictiveness, por medio del cual he llegado a la conclusión que con los valores con los que mejor trabaja simhash es con 5 gramas y un valor de restrictiveness que puede ser tanto 4 como 5.

Resultados para el fichero articles_100.train:

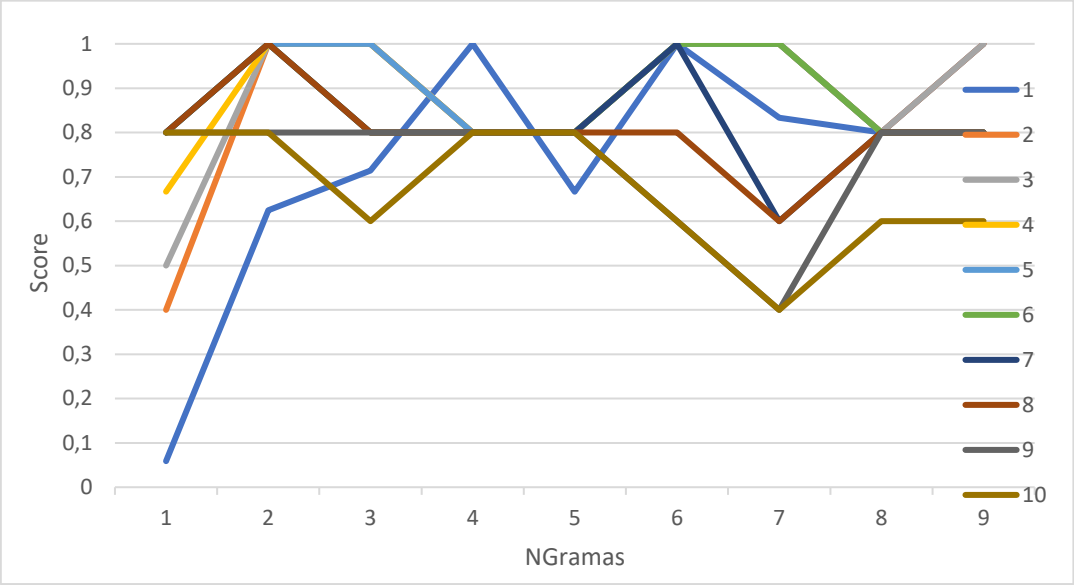

Figura 1. Score del simhash para los diferentes valores de Ngramas (eje x) y restrictiveness (leyenda)

Resultados para el fichero articles_10000.train:

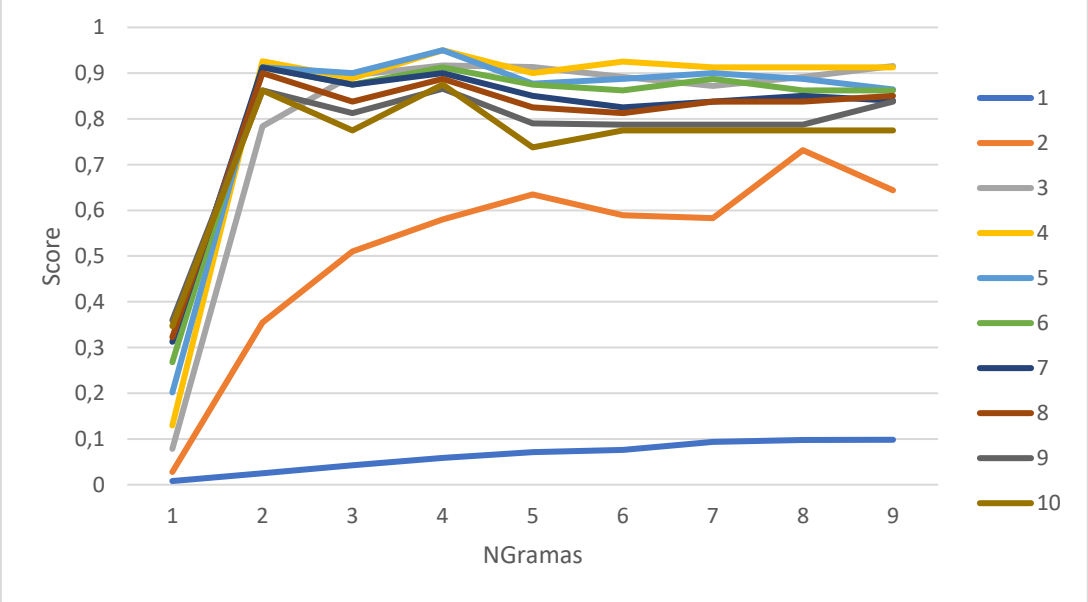

Figura 2. Score del simhash para los diferentes valores de Ngramas (eje x) y restrictiveness (leyenda).