SIW - P4 - SimHash

Primera versión (simhash_v1)

Para realizar esta práctica he reutilizado el código (un poco mejorado) de la clase BagOfWords. De este modo en el método simhash, el vector coge los valores del bag-of-words de la linea que se pasa por parámetro.

En el método calculate_hashes, se calculan y se ordena.

Sobre los resultados obtenidos podemos decir que con valor de restrictiveness 3, obtenemos buenos resultados pero no todos los que podríamos obtener.

Segunda versión (simhash_v2)

La vectorización de la linea se hace mediante la función ngrams del módulo NLTK, “sustituyendo” al bag-of-words.
Para ello he implementado la función “ngram_to_vector”, que se encarga de transformar el n-grama que obtenemos de NLTK a forma vectorial.

Con restrictiveness por defecto (es decir 10) y n=5 obtenemos todos los resultados que articles_1000.truth nos indica, para articles_1000.train.