Report on Locality Sensitive Hashing

Zhao, Zhiyuan

Assumptions:

1. Since our documents are all pretty short, usually no more than 20 words. So, we fixed the length of a shingle to be 1 word.
2. Also, since our documents are all pretty short, we ignored multiple word occurrences in a single document. This is to say, we count words in a document only once.

Approaches:

1. We utilized standard LSH algorithm. When generating signature matrix, we used a parallel implementation which greatly improves the running speed.

Performance analysis:

Algorithm	Time Consumption/s	MSE
pairwise distance	40.1641998	N/A
LSH, k=16	51.209002	0.00475189
LSH, k=128	317.912552	0.00062664

References:

1. https://github.com/rahularora/MinHash/blob/master/minhash.py
2. Other packages include scikit-learn, NumPy, multiprocessing.