**BDA Assignment 2: LSH**

**Group 3**

[**Code**](https://colab.research.google.com/drive/1ueZLoG7XwgOa5CELB92DrOVyQTLAf5e1#scrollTo=FYTr4cG3a7Qi)

**Shingling Methodology**

As suggested in the recommended [textbook](http://infolab.stanford.edu/~ullman/mmds/ch3n.pdf) for this course, we created shingles by considering substrings of length k, with k varying from 1 to 13. Also we replaced chunks of white spaces with a single white space, and included them as a part of our shingles, as it is useful in distinguishing shingles that cover more than 1 word from shingles that cover only 1 word. The text recommends a shingle size of 5 when the corpus consists of shorter entities like emails, and k=9 when it consists of larger entities like research papers. As our corpus consists of references found in the research papers, it lies somewhere between emails and research papers. Thus we decided to experiment with ks ranging from 1 to 13. It was also recommended that rather than storing the shingles created in the string form, they should be stored as hashes, thus we hashed each of the shingles to a 32 bit integer. The following were the number of shingles created as a result

| K | No. of shingles |
| --- | --- |
| 1 | 280075 |
| 2 | 2355568 |
| 3 | 6239799 |
| 4 | 8824813 |
| 5 | 10386603 |
| 6 | 11548081 |
| 7 | 12450123 |
| 8 | 13155252 |
| 9 | 13740916 |
| 10 | 14235311 |
| 11 | 14656938 |
| 12 | 15007758 |
| 13 | 15306210 |

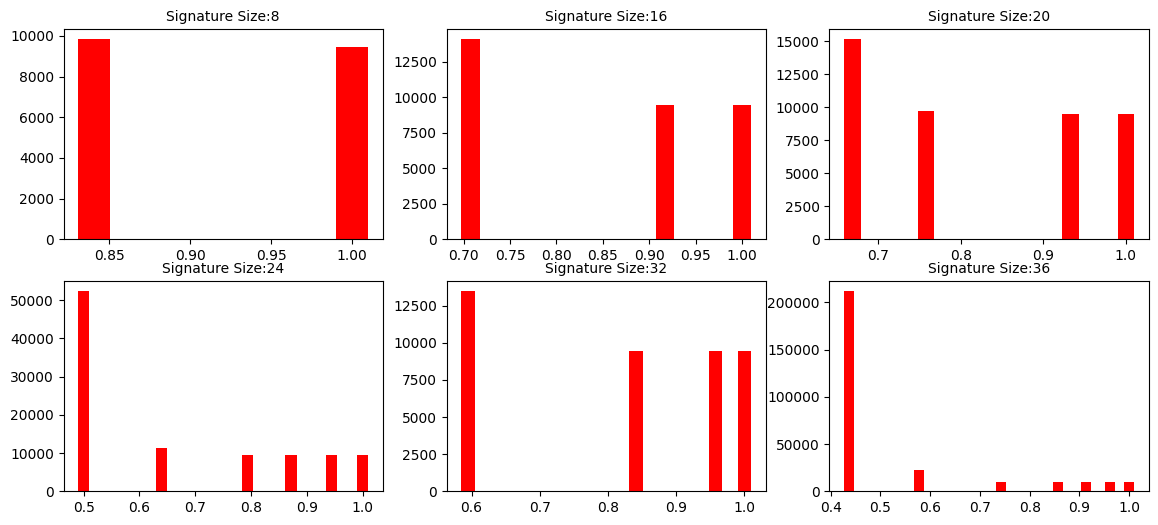
**Varying the similarity threshold**

We tried out all possible combinations of b and r for signature sizes 36, 32, 24, 20, 16, 8.

| b | r | similarity ((1/b)^(1/r)) | Number of pairs |
| --- | --- | --- | --- |
| Signature size = 36 | | | |
| 1 | 36 | 1 | 9453 |
| 2 | 18 | 0.9622238369 | 9453 |
| 3 | 12 | 0.9125147548 | 9453 |
| 4 | 9 | 0.8572439829 | 9453 |
| 6 | 6 | 0.7418363756 | 9532 |
| 9 | 4 | 0.5773502692 | 23038 |
| 12 | 3 | 0.4367902324 | 211835 |
| 18 | 2 | 0.2357022604 | 3722277 |
| Signature size = 32 | | | |
| 1 | 32 | 1 | 9453 |
| 2 | 16 | 0.9576032807 | 9453 |
| 4 | 8 | 0.8408964153 | 9455 |
| 8 | 4 | 0.5946035575 | 13464 |
| 16 | 2 | 0.25 | 2153195 |
| Signature size = 24 | | | |
| 1 | 24 | 1 | 9453 |
| 2 | 12 | 0.9438743127 | 9453 |
| 3 | 8 | 0.8716855429 | 9453 |
| 4 | 6 | 0.793700526 | 9461 |
| 6 | 4 | 0.6389431042 | 11348 |
| 8 | 3 | 0.5 | 52275 |
| 12 | 2 | 0.2886751346 | 1188618 |
| Signature size = 20 | | | |
| 1 | 20 | 1 | 9453 |
| 2 | 10 | 0.9330329915 | 9454 |
| 4 | 5 | 0.7578582833 | 9735 |
| 5 | 4 | 0.668740305 | 15145 |
| 10 | 2 | 0.316227766 | 1828744 |
| Signature size = 16 | | | |
| 1 | 16 | 1 | 9453 |
| 2 | 8 | 0.9170040432 | 9453 |
| 4 | 4 | 0.7071067812 | 14097 |
| 8 | 2 | 0.3535533906 | 1581314 |
| Signature size = 8 | | | |
| 1 | 8 | 1 | 9453 |
| 2 | 4 | 0.8408964153 | 9834 |
| 4 | 2 | 0.5 | 304860 |

(Getting 9453 matches for a similarity threshold of 1 is due to the fact that 138 documents from the dataset are empty strings.)

Histograms:

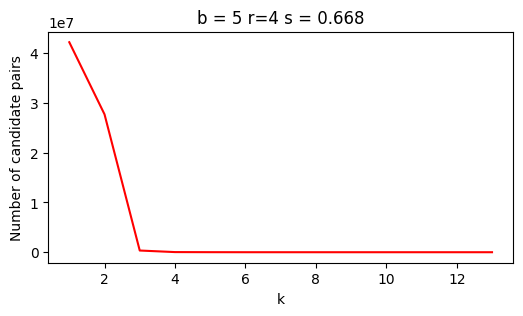


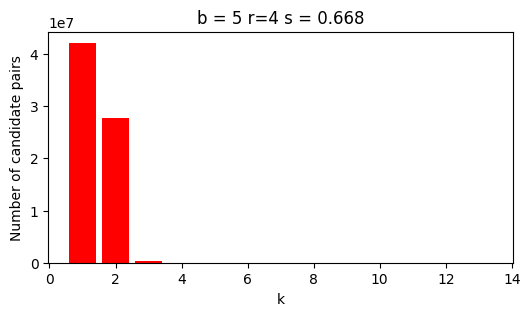
X axis contains the similarity threshold and the y axis contains the number of candidate pairs. While plotting the graphs extremely low thresholds, with b=signature size have been avoided as the graph became indecipherable, as the number of candidate pairs shoot up at the very large similarity.

**Varying the shingle size**

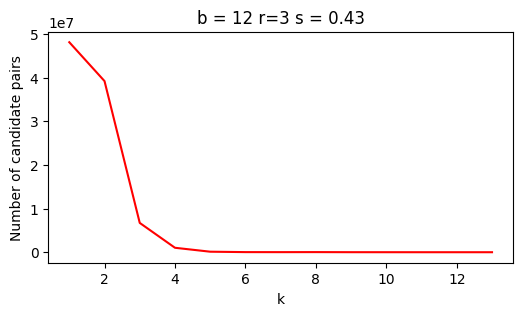
As noted above, we experimented with shingle sizes 1 to 13.

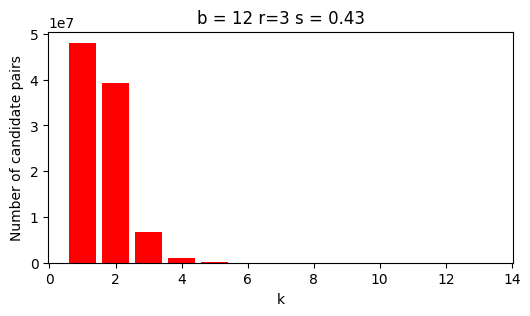
| **b=5, r=4 (signature size =20), s=0.668** | |
| --- | --- |
| **k** | No of pairs |
| 1 | 42134930 |
| 2 | 27673177 |
| 3 | 350720 |
| 4 | 36202 |
| 5 | 18295 |
| 6 | 10513 |
| 7 | 9877 |
| 8 | 9512 |
| 9 | 9498 |
| 10 | 9536 |
| 11 | 9698 |
| 12 | 9609 |
| 13 | 9751 |





| **b=12,r=3 (signature size= 36), s= 0.43** | |
| --- | --- |
| **k** | **No of pairs** |
| 1 | 48096981 |
| 2 | 39197532 |
| 3 | 6722417 |
| 4 | 1028542 |
| 5 | 128999 |
| 6 | 34731 |
| 7 | 27795 |
| 8 | 36777 |
| 9 | 17775 |
| 10 | 13454 |
| 11 | 10373 |
| 12 | 9977 |
| 13 | 10207 |





| **b=4, r=9, signature size=36, s = 0.85** | |
| --- | --- |
| **k** | **No of pairs** |
| 1 | 27408860 |
| 2 | 586208 |
| 3 | 9824 |
| 4 | 9464 |
| 5 | 9453 |
| 6 | 9454 |
| 7 | 9453 |
| 8 | 9454 |
| 9 | 9453 |
| 10 | 9453 |
| 11 | 9591 |
| 12 | 9591 |
| 13 | 9730 |

