Homework 1

Finding Similar Items: Textually Similar Documents

Badai Kesuma Hamid Dimyati

November 2020

1. Code commentary

The implementation of the task is done in Python mainly using numpy, pandas, and nltk libraries. The application finds the similarity of 100 documents from *"A Corpus of Plagiarised Short Answers"* data set[1]. This application is capable of extracting pairs of similar documents from a data set presented in that corpus. Documents containing similar texts are shown in the hash similarity score and signature similarity score. Jaccard similarity is used as a similarity measure to find identical pairs.

Within the code (`lsh.py`), we created a class called `shingling` which requires the parameter `k` as the number of words in one shingle. This class gives an output in the form of Jaccard similarity of each pair of documents. Apart from this matrix, we also save a Characteristic matrix within the class object. The next one, we created a class called `minhashing` which processes the Characteristic matrix into smaller matrices, Signature matrix and saves it into the class object. This class also requests parameter `k` as the number of signatures. The output of this class would be the Signature score for each pair of documents. Lastly, we implemented a class called `lsh` to run the Locality Sensitive Hashing (LSH) by processing the Signature matrix to get a set of candidates that meet the `threshold` parameter. Apart from `threshold`, this class also requires parameter `bands` as number of bands and `rows` as number of rows in each band.

[1] https://ir.shef.ac.uk/cloughie/resources/plagiarism_corpus.html

2. How to run

```
git clone https://github.com/hamiddimyati/id2222-data-mining-advanced.git
cd id2222-data-mining-advanced/assignment-1
python3 lsh.py
```

We could also modify all the parameters through the python code, such as `k`, `k_signature`, `bands`, `rows`, and `threshold`.

3. Results

In order to see the results, we have to look at the hash similarity score and signatures similarity score from each pair of documents. In hash similarity score, we see the number of similar shingles divided by the number of total shingles which both documents do not equal to zero. In signatures similarity score, we see the number of similar signatures divided by the number of all signatures.

From the result of Figure 1 we found that many pairs of documents have a similarity score of zero in both hash similarity score and signatures similarity score, which means that the pair of documents are completely different. From Figure 2 we see the pair of documents which have higher scores in the similarity, which indicate that the documents are plagiarized. The shingles length and the number of signatures affect the output of the similarity scores. Both hash similarity scores and signatures similarity scores give almost the same results which mean the approximation in signatures give good results compared to the exact computation in the hash, which was what we expected.

Figure 3 and Figure 4 are the results from the application (`lsh.py`), which are the most similar documents based on Signature similarity score and the sample of pair of candidates based on LSH, respectively.

	Doc1	Doc2	Jaccard_Score
0	Doc0	Doc1	0.000000
1	Doc0	Doc2	0.000000
2	Doc0	Doc3	0.000000
3	Doc0	Doc4	0.000000
4	Doc0	Doc5	0.000000
5	Doc0	Doc6	0.000000
6	Doc0	Doc7	0.000000
7	Doc0	Doc8	0.000000
8	Doc0	Doc9	0.000000
9	Doc0	Doc10	0.000000
10	Doc0	Doc11	0.000000
11	Doc0	Doc12	0.000000
12	Doc0	Doc13	0.000000
13	Doc0	Doc14	0.000000
14	Doc0	Doc15	0.000000
15	Doc0	Doc16	0.000000
16	Doc0	Doc17	0.000000
17	Doc0	Doc18	0.000000
18	Doc0	Doc19	0.000000
19	Doc0	Doc20	0.000000
20	Doc0	Doc21	0.000000
21	Doc0	Doc22	0.000000
22	Doc0	Doc23	0.000000
23	Doc0	Doc24	0.000000
24	Doc0	Doc25	0.000000
25	Doc0	Doc26	0.000000
26	Doc0	Doc27	0.000000
27	Doc0	Doc28	0.000000
28	Doc0	Doc29	0.000000
29	Doc0	Doc30	0.000000
...	...	...	...

	Doc1	Doc2	Jaccard_Score
...	...	...	...
9870	Doc99	Doc69	0.172794
9871	Doc99	Doc70	0.000000
9872	Doc99	Doc71	0.000000
9873	Doc99	Doc72	0.000000
9874	Doc99	Doc73	0.000000
9875	Doc99	Doc74	0.018613
9876	Doc99	Doc75	0.000000
9877	Doc99	Doc76	0.000000
9878	Doc99	Doc77	0.000000
9879	Doc99	Doc78	0.000000
9880	Doc99	Doc79	0.481481
9881	Doc99	Doc80	0.000000
9882	Doc99	Doc81	0.000000
9883	Doc99	Doc82	0.000000
9884	Doc99	Doc83	0.000000
9885	Doc99	Doc84	0.241379
9886	Doc99	Doc85	0.000000
9887	Doc99	Doc86	0.000000
9888	Doc99	Doc87	0.000000
9889	Doc99	Doc88	0.000000
9890	Doc99	Doc89	0.079023
9891	Doc99	Doc90	0.000000
9892	Doc99	Doc91	0.000000
9893	Doc99	Doc92	0.000000
9894	Doc99	Doc93	0.000000
9895	Doc99	Doc94	0.002725
9896	Doc99	Doc95	0.000000
9897	Doc99	Doc96	0.000000
9898	Doc99	Doc97	0.000000
9899	Doc99	Doc98	0.000000

9900 rows × 3 columns

	Doc1	Doc2	Signature_Score
0	Doc0	Doc1	0.00
1	Doc0	Doc2	0.00
2	Doc0	Doc3	0.00
3	Doc0	Doc4	0.00
4	Doc0	Doc5	0.00
5	Doc0	Doc6	0.00
6	Doc0	Doc7	0.00
7	Doc0	Doc8	0.00
8	Doc0	Doc9	0.00
9	Doc0	Doc10	0.00
10	Doc0	Doc11	0.00
11	Doc0	Doc12	0.00
12	Doc0	Doc13	0.00
13	Doc0	Doc14	0.00
14	Doc0	Doc15	0.00
15	Doc0	Doc16	0.00
16	Doc0	Doc17	0.00
17	Doc0	Doc18	0.00
18	Doc0	Doc19	0.00
19	Doc0	Doc20	0.00
20	Doc0	Doc21	0.00
21	Doc0	Doc22	0.00
22	Doc0	Doc23	0.00
23	Doc0	Doc24	0.00
24	Doc0	Doc25	0.00
25	Doc0	Doc26	0.00
26	Doc0	Doc27	0.00
27	Doc0	Doc28	0.00
28	Doc0	Doc29	0.00
29	Doc0	Doc30	0.00
...	...	...	...

	Doc1	Doc2	Signature_Score
...	...	...	...
9870	Doc99	Doc69	0.75
9871	Doc99	Doc70	0.00
9872	Doc99	Doc71	0.00
9873	Doc99	Doc72	0.00
9874	Doc99	Doc73	0.00
9875	Doc99	Doc74	0.00
9876	Doc99	Doc75	0.00
9877	Doc99	Doc76	0.00
9878	Doc99	Doc77	0.00
9879	Doc99	Doc78	0.00
9880	Doc99	Doc79	0.50
9881	Doc99	Doc80	0.00
9882	Doc99	Doc81	0.00
9883	Doc99	Doc82	0.00
9884	Doc99	Doc83	0.00
9885	Doc99	Doc84	0.50
9886	Doc99	Doc85	0.00
9887	Doc99	Doc86	0.00
9888	Doc99	Doc87	0.00
9889	Doc99	Doc88	0.00
9890	Doc99	Doc89	0.00
9891	Doc99	Doc90	0.00
9892	Doc99	Doc91	0.00
9893	Doc99	Doc92	0.00
9894	Doc99	Doc93	0.00
9895	Doc99	Doc94	0.00
9896	Doc99	Doc95	0.00
9897	Doc99	Doc96	0.00
9898	Doc99	Doc97	0.00
9899	Doc99	Doc98	0.00

9900 rows × 3 columns

Figure 1. The output of hash similarity scores on each pair of documents (left) and signature similarity scores on each pair of documents (right)

	Doc1	Doc2	Jaccard_Score	Signature_Score
9425	Doc95	Doc20	0.878689	1.00
2074	Doc20	Doc95	0.878689	1.00
9485	Doc95	Doc80	0.790625	0.75
8014	Doc80	Doc95	0.790625	0.75
7940	Doc80	Doc20	0.683891	0.75
2059	Doc20	Doc80	0.683891	0.75
9765	Doc98	Doc63	0.619048	1.00
6334	Doc63	Doc98	0.619048	1.00
9785	Doc98	Doc83	0.493438	0.75
8314	Doc83	Doc98	0.493438	0.75
7919	Doc79	Doc99	0.481481	0.50
9880	Doc99	Doc79	0.481481	0.50
8280	Doc83	Doc63	0.459948	0.75
6319	Doc63	Doc83	0.459948	0.75
6855	Doc69	Doc24	0.366460	0.25
2444	Doc24	Doc69	0.366460	0.25
4949	Doc49	Doc99	0.357394	0.50
9850	Doc99	Doc49	0.357394	0.50
2049	Doc20	Doc70	0.341216	0.00
6950	Doc70	Doc20	0.341216	0.00
7009	Doc70	Doc80	0.341216	0.25
7990	Doc80	Doc70	0.341216	0.25
9610	Doc97	Doc7	0.340050	0.25
789	Doc7	Doc97	0.340050	0.25
9475	Doc95	Doc70	0.320635	0.00
7024	Doc70	Doc95	0.320635	0.00
5029	Doc50	Doc80	0.316667	0.50
7970	Doc80	Doc50	0.316667	0.50
5044	Doc50	Doc95	0.300792	0.50
9455	Doc95	Doc50	0.300792	0.50

Figure 2. The output of similarity scores > 0.3

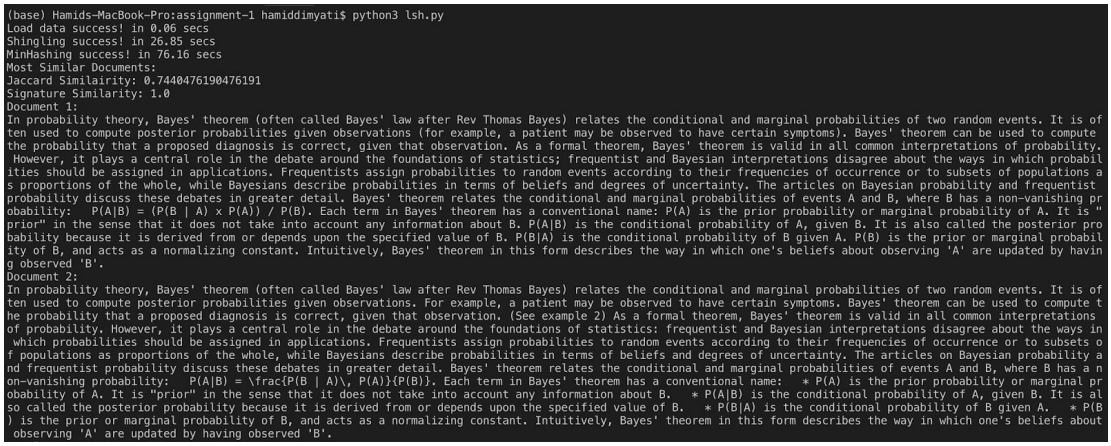

Figure 3. Output of minhashing class as the most similar documents based on Signature similarity score

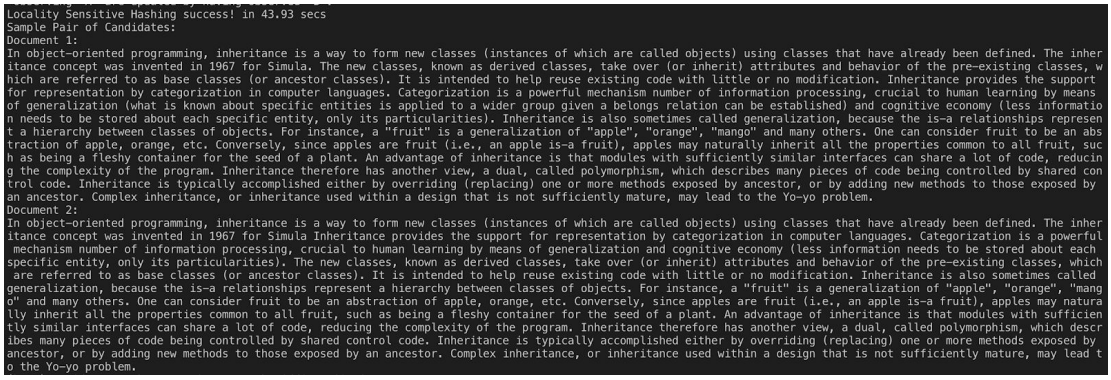

Figure 4. Output of lsh class as the sample of pair of candidates to be further checked

References

[1] Clough, P., & Stevenson, M. (2011). Developing a corpus of plagiarised short answers. Language resources and evaluation, 45(1), 5-24.