Data Mining - Homework 1

Robert-Andrei Damian and Alice De Schutter

Finding Similar Items: Textually Similar Documents

November 12, 2021

Task

Implement the stages of finding textually similar documents based on Jaccard similarity using the shingling, minhashing, and locality-sensitive hashing (LSH) techniques and corresponding algorithms. The implementation can be done using any big data processing framework, such as Apache Spark, Apache Flink, or no framework, e.g., in Java, Python, etc. To test and evaluate the implementation, write a program that uses your implementation to find similar documents in a corpus of 5-10 or more documents such as web pages or emails.

1 Detailed information

Apache Spark was used. The task was divided into different subsections (see below). Each subsection deals with one of the stages of finding textually similar documents.

- **Subsection 1: Computes k-shingles for each document**, where k has default value *shingle_size = 10.* For each document, k-shingles were created and shingle doubles were removed. Subsection 1 outputs a three column Spark DataFrame. Each row corresponds to a document. The column named "shingles" contains shingles for each document.
- **Subsection 2: "Cross-compares" documents based on Jaccard similarity**. Subsection 2 outputs a two-column Spark DataFrame. Each row represents a pair of documents. The column called "similarity" contains the Jaccard similarity values.
- **Subsection 3: Builds MinHash signatures and "cross-compares" documents based on the signatures**. This was performed twice:
  1. With the function *shingles_to_signatures()* and using the public LSH class MinHashLSH().
  2. With the function *shingles_to_signatures_from_scratch()* and using our own hash functions. The parameter *signature_reducing_factor* is used to control the size of the signatures (the size is equal to the total number of shingles divided by the parameter). The *signature_reducing_factor* parameter has default value 100.

  Subsection 3 outputs four Spark DataFrames (two for 1. and two for 2.). The first DataFrames show hashed shingles and signatures for all documents. The second Dataframes show "similarity" based on signatures between pairs of documents (as a fraction of components in which they agree).
- **Subsection 4: Finds candidate pairs of signatures agreeing on at least fraction t of their components,** where t has default value lsh_threshold = 0.8. The optimal number of rows per band is found with a binary search. Subsection 4 outputs a two-column Spark DataFrame where rows represent document pairs for which at least a fraction t of their components agree. The column "similarity" represents similarity based on signatures found with our own hash functions (see Subsection 3.2.)

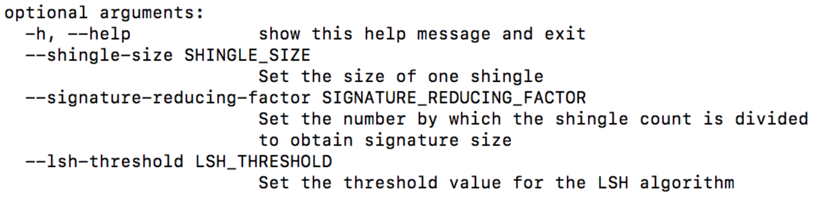
For readability, subsections were separated by descriptive comments in the code. The same part of the code was used to "cross-compare" documents for Subsection 2 and 3 ("Compare hashes / signatures").

Found in the "documents" folder, a corpus of 9 documents was used to test and evaluate the implementation. Documents were purposely chosen to be similar with respect to other documents in the corpus (i.e. "ndv_macron_merkel.txt" and "thelocalfr_macron_merkel.txt").

Instructions on how to build and run the program

- pip install pyspark
- spark-submit similar_items.py

Information about optional command line parameters:



```
optional arguments:
  -h, --help            show this help message and exit
  --shingle-size SHINGLE_SIZE
                        Set the size of one shingle
  --signature-reducing-factor SIGNATURE_REDUCING_FACTOR
                        Set the number by which the shingle count is divided
                        to obtain signature size
  --lsh-threshold LSH_THRESHOLD
                        Set the threshold value for the LSH algorithm
```

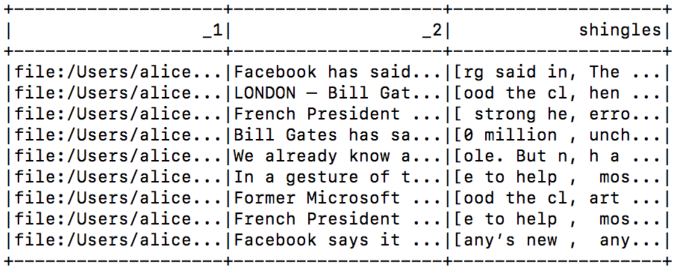
Figure 1: Optional arguments

2 Results

The results shown below are found with default values for command line parameters.

- *shingle_size = 10*
- *signature_reducing_factor = 100*
- lsh_threshold = 0.8

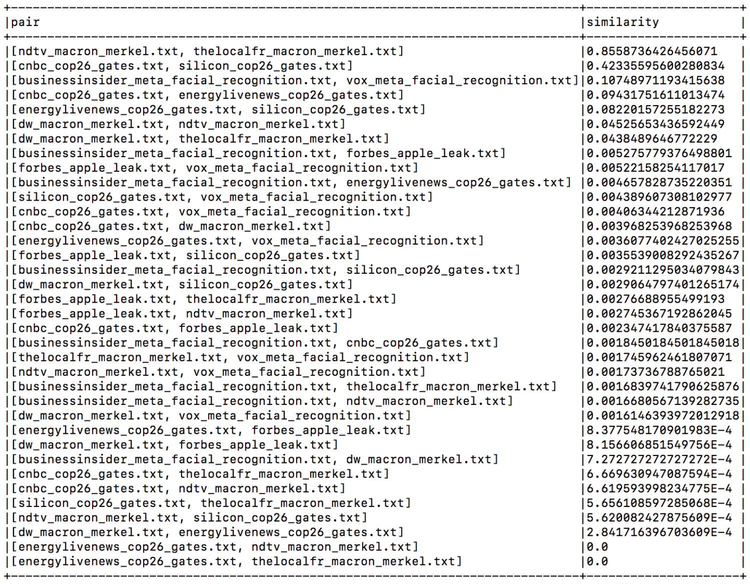
- **Subsection 1:**



```
+--------------------+--------------------+--------------------+
|                  _1|                  _2|            shingles|
+--------------------+--------------------+--------------------+
|file:/Users/alice...|Facebook has said...|[rg said in, The ...|
|file:/Users/alice...|LONDON – Bill Gat...|[ood the cl, hen ...|
|file:/Users/alice...|French President ...|[ strong he, erro...|
|file:/Users/alice...|Bill Gates has sa...|[0 million , unch...|
|file:/Users/alice...|We already know a...|[ole. But n, h a ...|
|file:/Users/alice...|In a gesture of t...|[e to help ,  mos...|
|file:/Users/alice...|Former Microsoft ...|[ood the cl, art ...|
|file:/Users/alice...|French President ...|[e to help ,  mos...|
|file:/Users/alice...|Facebook says it ...|[any's new ,  any...|
+--------------------+--------------------+--------------------+
```

Figure 2: Spark DataFrame output for Subsection 1

- **Subsection 2:**

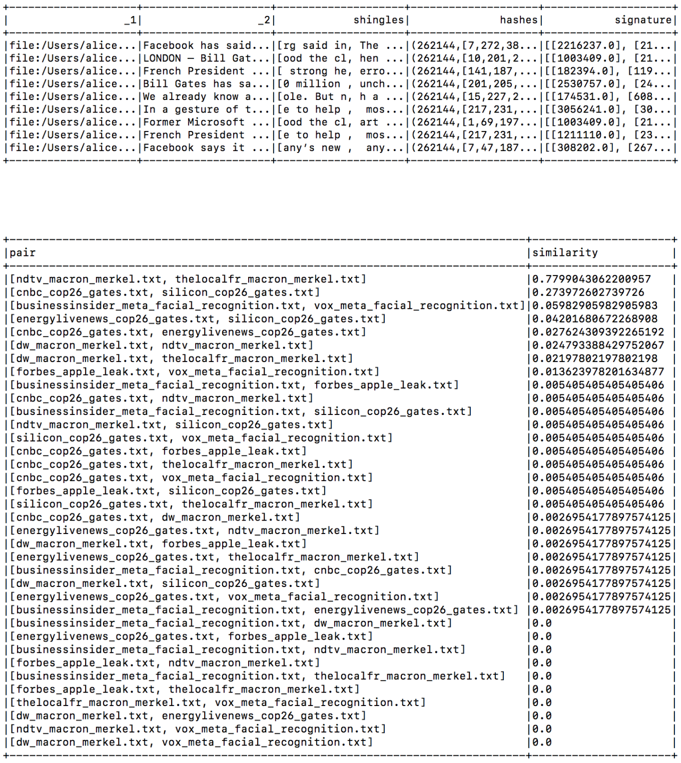


```
+------------------------------------------------------------------------------+---------------------+
|pair                                                                          |similarity           |
+------------------------------------------------------------------------------+---------------------+
|[ndtv_macron_merkel.txt, thelocalfr_macron_merkel.txt]                        |0.8558736426456071   |
|[cnbc_cop26_gates.txt, silicon_cop26_gates.txt]                               |0.42335595600280834  |
|[businessinsider_meta_facial_recognition.txt, vox_meta_facial_recognition.txt]|0.10748971193415638  |
|[cnbc_cop26_gates.txt, energylivenews_cop26_gates.txt]                        |0.09431751611013474  |
|[energylivenews_cop26_gates.txt, silicon_cop26_gates.txt]                     |0.08220157255182273  |
|[dw_macron_merkel.txt, ndtv_macron_merkel.txt]                                |0.04525653436592449  |
|[dw_macron_merkel.txt, thelocalfr_macron_merkel.txt]                          |0.0438489646772229   |
|[businessinsider_meta_facial_recognition.txt, forbes_apple_leak.txt]          |0.005275779376498801 |
|[forbes_apple_leak.txt, vox_meta_facial_recognition.txt]                      |0.00522158254117017  |
|[businessinsider_meta_facial_recognition.txt, energylivenews_cop26_gates.txt] |0.004657828735220351 |
|[silicon_cop26_gates.txt, vox_meta_facial_recognition.txt]                    |0.004389607308102977 |
|[cnbc_cop26_gates.txt, vox_meta_facial_recognition.txt]                       |0.00406344212871936  |
|[cnbc_cop26_gates.txt, dw_macron_merkel.txt]                                  |0.003968253968253968 |
|[energylivenews_cop26_gates.txt, vox_meta_facial_recognition.txt]             |0.0036077402427025255|
|[forbes_apple_leak.txt, silicon_cop26_gates.txt]                              |0.0035539008292435267|
|[businessinsider_meta_facial_recognition.txt, silicon_cop26_gates.txt]        |0.0029211295034079843|
|[dw_macron_merkel.txt, silicon_cop26_gates.txt]                               |0.0029064797401265174|
|[forbes_apple_leak.txt, thelocalfr_macron_merkel.txt]                         |0.00276688955499193  |
|[forbes_apple_leak.txt, ndtv_macron_merkel.txt]                               |0.002745367192862045 |
|[cnbc_cop26_gates.txt, forbes_apple_leak.txt]                                 |0.002347417840375587 |
|[businessinsider_meta_facial_recognition.txt, cnbc_cop26_gates.txt]           |0.0018450184501845018|
|[thelocalfr_macron_merkel.txt, vox_meta_facial_recognition.txt]               |0.001745962461807071 |
|[ndtv_macron_merkel.txt, vox_meta_facial_recognition.txt]                     |0.00173736788765021  |
|[businessinsider_meta_facial_recognition.txt, thelocalfr_macron_merkel.txt]   |0.0016839741790625876|
|[businessinsider_meta_facial_recognition.txt, ndtv_macron_merkel.txt]         |0.0016680567139282735|
|[dw_macron_merkel.txt, vox_meta_facial_recognition.txt]                       |0.0016146393972012918|
|[energylivenews_cop26_gates.txt, forbes_apple_leak.txt]                       |8.377548170901983E-4 |
|[dw_macron_merkel.txt, forbes_apple_leak.txt]                                 |8.156606851549756E-4 |
|[businessinsider_meta_facial_recognition.txt, dw_macron_merkel.txt]           |7.272727272727272E-4 |
|[cnbc_cop26_gates.txt, thelocalfr_macron_merkel.txt]                          |6.669630947087594E-4 |
|[cnbc_cop26_gates.txt, ndtv_macron_merkel.txt]                                |6.619593998234775E-4 |
|[silicon_cop26_gates.txt, thelocalfr_macron_merkel.txt]                       |5.656108597285068E-4 |
|[ndtv_macron_merkel.txt, silicon_cop26_gates.txt]                             |5.620082427875609E-4 |
|[dw_macron_merkel.txt, energylivenews_cop26_gates.txt]                        |2.841716396703609E-4 |
|[energylivenews_cop26_gates.txt, ndtv_macron_merkel.txt]                      |0.0                  |
|[energylivenews_cop26_gates.txt, thelocalfr_macron_merkel.txt]                |0.0                  |
+------------------------------------------------------------------------------+---------------------+
```

Figure 3: Spark DataFrame output for Subsection 2

- **Subsection 3:**

1. Output associated with 1.

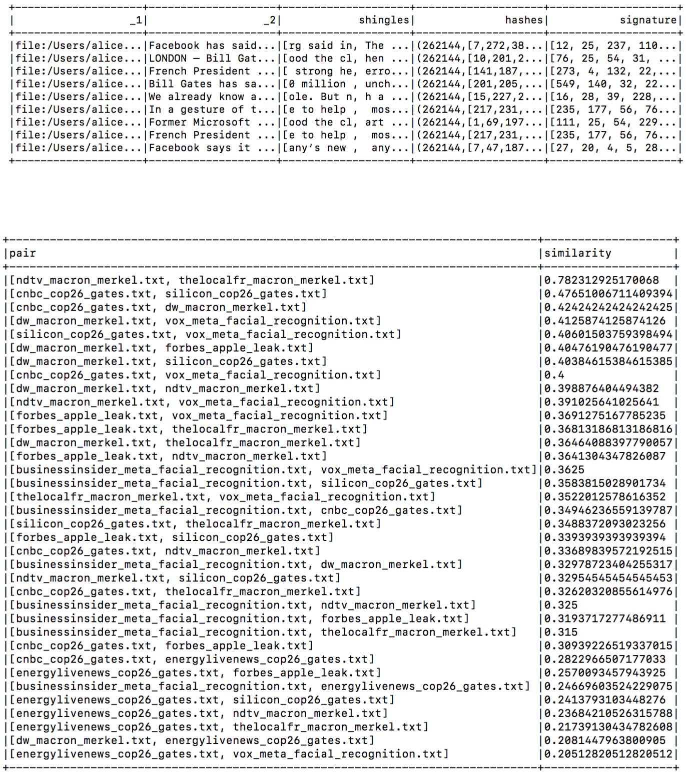


```
+--------------------+--------------------+--------------------+--------------------+--------------------+
|                  _1|                  _2|            shingles|              hashes|           signature|
+--------------------+--------------------+--------------------+--------------------+--------------------+
|file:/Users/alice...|Facebook has said...|[rg said in, The ...|(262144,[7,272,38...|[[2216237.0], [21...|
|file:/Users/alice...|LONDON — Bill Gat...|[ood the cl, hen ...|(262144,[10,201,2...|[[1003409.0], [21...|
|file:/Users/alice...|French President ...|[ strong he, erro...|(262144,[141,187,...|[[182394.0], [119...|
|file:/Users/alice...|Bill Gates has sa...|[0 million , unch...|(262144,[201,205,...|[[2530757.0], [24...|
|file:/Users/alice...|We already know a...|[ole. But n, h a ...|(262144,[15,227,2...|[[174531.0], [608...|
|file:/Users/alice...|In a gesture of t...|[e to help ,  mos...|(262144,[217,231,...|[[3056241.0], [30...|
|file:/Users/alice...|Former Microsoft ...|[ood the cl, art ...|(262144,[1,69,197...|[[1003409.0], [21...|
|file:/Users/alice...|French President ...|[e to help ,  mos...|(262144,[217,231,...|[[1211110.0], [23...|
|file:/Users/alice...|Facebook says it ...|[any’s new ,  any...|(262144,[7,47,187...|[[308202.0], [267...|
+--------------------+--------------------+--------------------+--------------------+--------------------+
```

```
+------------------------------------------------------------------------------+---------------------+
|pair                                                                          |similarity           |
+------------------------------------------------------------------------------+---------------------+
|[ndtv_macron_merkel.txt, thelocalfr_macron_merkel.txt]                        |0.7799043062200957   |
|[cnbc_cop26_gates.txt, silicon_cop26_gates.txt]                               |0.273972602739726    |
|[businessinsider_meta_facial_recognition.txt, vox_meta_facial_recognition.txt]|0.05982905982905983  |
|[energylivenews_cop26_gates.txt, silicon_cop26_gates.txt]                     |0.04201680672268908  |
|[cnbc_cop26_gates.txt, energylivenews_cop26_gates.txt]                        |0.027624309392265192 |
|[dw_macron_merkel.txt, ndtv_macron_merkel.txt]                                |0.024793388429752067 |
|[dw_macron_merkel.txt, thelocalfr_macron_merkel.txt]                          |0.02197802197802198  |
|[forbes_apple_leak.txt, vox_meta_facial_recognition.txt]                      |0.013623978201634877 |
|[businessinsider_meta_facial_recognition.txt, forbes_apple_leak.txt]          |0.005405405405405406 |
|[cnbc_cop26_gates.txt, ndtv_macron_merkel.txt]                                |0.005405405405405406 |
|[businessinsider_meta_facial_recognition.txt, silicon_cop26_gates.txt]        |0.005405405405405406 |
|[ndtv_macron_merkel.txt, silicon_cop26_gates.txt]                             |0.005405405405405406 |
|[silicon_cop26_gates.txt, vox_meta_facial_recognition.txt]                    |0.005405405405405406 |
|[cnbc_cop26_gates.txt, forbes_apple_leak.txt]                                 |0.005405405405405406 |
|[cnbc_cop26_gates.txt, thelocalfr_macron_merkel.txt]                          |0.005405405405405406 |
|[cnbc_cop26_gates.txt, vox_meta_facial_recognition.txt]                       |0.005405405405405406 |
|[forbes_apple_leak.txt, silicon_cop26_gates.txt]                              |0.005405405405405406 |
|[silicon_cop26_gates.txt, thelocalfr_macron_merkel.txt]                       |0.005405405405405406 |
|[cnbc_cop26_gates.txt, dw_macron_merkel.txt]                                  |0.0026954177897574125|
|[energylivenews_cop26_gates.txt, ndtv_macron_merkel.txt]                      |0.0026954177897574125|
|[dw_macron_merkel.txt, forbes_apple_leak.txt]                                 |0.0026954177897574125|
|[energylivenews_cop26_gates.txt, thelocalfr_macron_merkel.txt]                |0.0026954177897574125|
|[businessinsider_meta_facial_recognition.txt, cnbc_cop26_gates.txt]           |0.0026954177897574125|
|[dw_macron_merkel.txt, silicon_cop26_gates.txt]                               |0.0026954177897574125|
|[energylivenews_cop26_gates.txt, vox_meta_facial_recognition.txt]             |0.0026954177897574125|
|[businessinsider_meta_facial_recognition.txt, energylivenews_cop26_gates.txt] |0.0026954177897574125|
|[businessinsider_meta_facial_recognition.txt, dw_macron_merkel.txt]           |0.0                  |
|[energylivenews_cop26_gates.txt, forbes_apple_leak.txt]                       |0.0                  |
|[businessinsider_meta_facial_recognition.txt, ndtv_macron_merkel.txt]         |0.0                  |
|[forbes_apple_leak.txt, ndtv_macron_merkel.txt]                               |0.0                  |
|[businessinsider_meta_facial_recognition.txt, thelocalfr_macron_merkel.txt]   |0.0                  |
|[forbes_apple_leak.txt, thelocalfr_macron_merkel.txt]                         |0.0                  |
|[thelocalfr_macron_merkel.txt, vox_meta_facial_recognition.txt]               |0.0                  |
|[dw_macron_merkel.txt, energylivenews_cop26_gates.txt]                        |0.0                  |
|[ndtv_macron_merkel.txt, vox_meta_facial_recognition.txt]                     |0.0                  |
|[dw_macron_merkel.txt, vox_meta_facial_recognition.txt]                       |0.0                  |
+------------------------------------------------------------------------------+---------------------+
```

Figure 4: Spark DataFrames computed with *shingles_to_signatures()*

2. Output associated with 2.

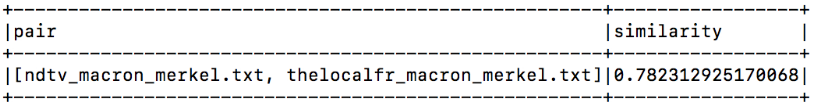


```
+--------------------+--------------------+--------------------+--------------------+--------------------+
|                  _1|                  _2|            shingles|              hashes|           signature|
+--------------------+--------------------+--------------------+--------------------+--------------------+
|file:/Users/alice...|Facebook has said...|[rg said in, The ...|(262144,[7,272,38...|[12, 25, 237, 110...|
|file:/Users/alice...|LONDON — Bill Gat...|[ood the cl, hen ...|(262144,[10,201,2...|[76, 25, 54, 31, ...|
|file:/Users/alice...|French President ...|[ strong he, erro...|(262144,[141,187,...|[273, 4, 132, 22,...|
|file:/Users/alice...|Bill Gates has sa...|[0 million , unch...|(262144,[201,205,...|[549, 140, 32, 22...|
|file:/Users/alice...|We already know a...|[ole. But n, h a ...|(262144,[15,227,2...|[16, 28, 39, 228,...|
|file:/Users/alice...|In a gesture of t...|[e to help ,  mos...|(262144,[217,231,...|[235, 177, 56, 76...|
|file:/Users/alice...|Former Microsoft ...|[ood the cl, art ...|(262144,[1,69,197...|[111, 25, 54, 229...|
|file:/Users/alice...|French President ...|[e to help ,  mos...|(262144,[217,231,...|[235, 177, 56, 76...|
|file:/Users/alice...|Facebook says it ...|[any’s new ,  any...|(262144,[7,47,187...|[27, 20, 4, 5, 28...|
+--------------------+--------------------+--------------------+--------------------+--------------------+
```

```
+------------------------------------------------------------------------------+-------------------+
|pair                                                                          |similarity         |
+------------------------------------------------------------------------------+-------------------+
|[ndtv_macron_merkel.txt, thelocalfr_macron_merkel.txt]                        |0.782312925170068  |
|[cnbc_cop26_gates.txt, silicon_cop26_gates.txt]                               |0.4765100671409394 |
|[cnbc_cop26_gates.txt, dw_macron_merkel.txt]                                  |0.42424242424242425|
|[dw_macron_merkel.txt, vox_meta_facial_recognition.txt]                       |0.4125874125874126 |
|[silicon_cop26_gates.txt, vox_meta_facial_recognition.txt]                    |0.40601503759398494|
|[dw_macron_merkel.txt, forbes_apple_leak.txt]                                 |0.40476190476190477|
|[dw_macron_merkel.txt, silicon_cop26_gates.txt]                               |0.40384615384615385|
|[cnbc_cop26_gates.txt, vox_meta_facial_recognition.txt]                       |0.4                |
|[dw_macron_merkel.txt, ndtv_macron_merkel.txt]                                |0.398876404494382  |
|[ndtv_macron_merkel.txt, vox_meta_facial_recognition.txt]                     |0.391025641025641  |
|[forbes_apple_leak.txt, vox_meta_facial_recognition.txt]                      |0.3691275167785235 |
|[forbes_apple_leak.txt, thelocalfr_macron_merkel.txt]                         |0.36813186813186816|
|[dw_macron_merkel.txt, thelocalfr_macron_merkel.txt]                          |0.36464088397790057|
|[forbes_apple_leak.txt, ndtv_macron_merkel.txt]                               |0.3641304347826087 |
|[businessinsider_meta_facial_recognition.txt, vox_meta_facial_recognition.txt]|0.3625             |
|[businessinsider_meta_facial_recognition.txt, silicon_cop26_gates.txt]        |0.3583815028901734 |
|[thelocalfr_macron_merkel.txt, vox_meta_facial_recognition.txt]               |0.3522012578616352 |
|[businessinsider_meta_facial_recognition.txt, cnbc_cop26_gates.txt]           |0.34946236559139787|
|[silicon_cop26_gates.txt, thelocalfr_macron_merkel.txt]                       |0.3488372093023256 |
|[forbes_apple_leak.txt, silicon_cop26_gates.txt]                              |0.3393939393939394 |
|[cnbc_cop26_gates.txt, ndtv_macron_merkel.txt]                                |0.33689839572192515|
|[businessinsider_meta_facial_recognition.txt, dw_macron_merkel.txt]           |0.32978723404255317|
|[ndtv_macron_merkel.txt, silicon_cop26_gates.txt]                             |0.32954545454545453|
|[cnbc_cop26_gates.txt, thelocalfr_macron_merkel.txt]                          |0.32620320855614976|
|[businessinsider_meta_facial_recognition.txt, ndtv_macron_merkel.txt]         |0.325              |
|[businessinsider_meta_facial_recognition.txt, forbes_apple_leak.txt]          |0.3193717277486911 |
|[businessinsider_meta_facial_recognition.txt, thelocalfr_macron_merkel.txt]   |0.315              |
|[cnbc_cop26_gates.txt, forbes_apple_leak.txt]                                 |0.30939226519337015|
|[cnbc_cop26_gates.txt, energylivenews_cop26_gates.txt]                        |0.2822966507177033 |
|[energylivenews_cop26_gates.txt, forbes_apple_leak.txt]                       |0.2570093457943925 |
|[businessinsider_meta_facial_recognition.txt, energylivenews_cop26_gates.txt] |0.24669603524229075|
|[energylivenews_cop26_gates.txt, silicon_cop26_gates.txt]                     |0.2413793103448276 |
|[energylivenews_cop26_gates.txt, ndtv_macron_merkel.txt]                      |0.23684210526315788|
|[energylivenews_cop26_gates.txt, thelocalfr_macron_merkel.txt]                |0.21739130434782608|
|[dw_macron_merkel.txt, energylivenews_cop26_gates.txt]                        |0.2081447963800905 |
|[energylivenews_cop26_gates.txt, vox_meta_facial_recognition.txt]             |0.20512820512820512|
+------------------------------------------------------------------------------+-------------------+
```

Figure 5: Spark DataFrames computed with *shingles_to_signatures_from_scratch()*

- **Subsection 4:**



```
+------------------------------------------------------+-----------------+
|pair                                                  |similarity       |
+------------------------------------------------------+-----------------+
|[ndtv_macron_merkel.txt, thelocalfr_macron_merkel.txt]|0.782312925170068|
+------------------------------------------------------+-----------------+
```

Figure 6: Spark DataFrame output for Subsection 3