# Homework 1: Finding Similar Items: Textually Similar Documents

ID2222 Data Mining

Márk Antal Csizmadia
macs@kth.se

November 2021

# 1 Introduction

In this document, methods and results of finding similar items in the context of finding similar documents is discussed. The discussion is accompanied by the exploration of the architecture of the code base submitted alongside with this report. Firstly, the methods used in the code are reviewed, and then some results are shown and discussed.

The accompanying code base is written in Python and PySpark. Follow the `README.md` file on how to setup and run the code. The classes and methods are factored into the `classes.py` file and data reading and plotting are included in the `utils.py` file. The demonstration of the use of the code can be found in the Jupyter Notebook titled `main.ipynb`.

# 2 Data

The data used in the code is a collection of BBC news articles, which is publicly available here [1]. The dataset includes 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005. The data set also includes 5 class labels for the topics of the articles such as business, entertainment, politics, sport, and tech. In the code, only some of the articles are used for sensible running times. Specifically, 20 business articles are used to find similar ones amongst them. The data is downloaded within the `main.ipynb` notebook, and is read with the `DataReader` class located in the `utils.py` file.

# 3 Methods

Similar documents are found through the pipeline of shingling, MinHashing, and Locally Sensitive Hashing (LSH). Reprehensibility is ensured by fixing the Python environment variable `PYTHONHASHSEED` so that the built-in Python `hash()` function yields consistent results, and by passing the `seed` variable to the `MinHashing` class constructor as minhashing uses the `numpy.random.randint()` function.

In shingling, the text of the articles is split into chunks of characters of length `k`, and the k-shingles are then hashed to be represented by integers for saving space in memory. K-Shingling is implemented in the `Shingling` class in `classes.py`. In the notebook, `k_shingles_rdd` includes the set of hashed k-shingles per document for each document, and `k_shingles_all_rdd` includes all of the unique hashed k-shingles from across all of the documents used in the exercise. In the example setting, there are `n_shingles` = 37936 unique hashed `k`-shingles found across the 20 documents when k=10.

Using the `Vectorizer` class in `classes.py`, the hashed k-shingles are transformed into a Boolean vector in which 1s represent hashed k-shingles from across all of the documents (`k_shingles_all_rdd`) that are present in a given document, and 0s represent the opposite. Therefore, each vector per document in `k_shingles_vectorized_rdd` is of length `n_shingles`=37936, filled with 1s and 0s. The vectorized hashed k-shingle representation are used in subsequent steps.

The first method of fining similar documents is via computing the Jaccard similarity of each pair of vectorized hashed k-shingle representation, which is implemented with the `CompareSets` class in `classes.py`. The Jaccard similarity is the intersection over the union of the sets of hashed k-shingles per the two documents. In the notebook, `js_rdd` includes the pairwise similarities of the hashed k-shingles of each pair of documents. Jaccard similarities of the hashed k-shingles are visualized with the use the `make_heatmap` function in `utils.py`.

To reduce the in-memory size of the text representations, MinHashing is used to hash multiple times the hashed k-shingles of each text yielding a MinHashing signature of uniform size for each document. The MinHashing is implemented in the `MinHashing` class in `classes.py`. MinHasing take as arguments the size of the signatures, or equivalently, the number of hash functions, (`n_signatures`), which in the example is 100, and a seed value for reproduciblity. MinHashing uses hash functions of the form h(x) = (ax + b) % c where c is set to the prime number (`prime_modulo`) after the integer representing the number of unique hashed k-shingles across all documents (`n_shingles`) to avoid hash collision. In the notebook, `signatures_rdd` contains the MinHashing signatures for each of the 20 documents.

The second method of finding similar documents is by computing the similarities of MinHashing signatures for each pair of documents. This is implemented in the `CompareSignatures` class in `classes.py`. The similarity of two signature is defined as the ratio of the signatures having the same value to the size of the signature vectors. In the notebook, the `signature_similarities_rdd` include the pairwise similarities of signatures per document. The similarities of signatures are visualized with the use the `make_heatmap` function in `utils.py`.

The third and final method of finding similar documents is by the Locally Sensitive Hashing (LSH) algorithm. LSH is implemented in the `LSH` class in `classes.py`. The LSH splits the signatures of size `n_signatures` into `n_bands` number of bands each having `LSH.n_rows_per_band` number of signature entries. LSH hashes the bands of each signature in `n_buckets` number of buckets per band. If the same band of two signature hashed to the same bucket equal or more times than `LSH.hash_to_n_min` times (set to 1), the two signatures are marked as similar candidates. In the example in the notebook, `n_signatures` = 100, `n_bands` = 5, `LSH.n_rows_per_band` = 20, and `n_buckets` = 20. This yields a similarity threshold (for flagging twp documents as similar candidates) of `(1 / n_bands)` ^(1 / LSH.n_rows_per_band) = 0.9226. In the notebook, `candidates_rdd` includes the similar candidates per each document.

## 4 Results

The Jaccard similarities of hashed k-shingles are shown in Figure 1. The similarities are defined on a scale of 0-1, 1 being identical. The similarities are shown on a logarithmic scale. The numbers around the heatmap are the document identifiers of the 20 documents used in the code. For the document ids refer to Table 1.

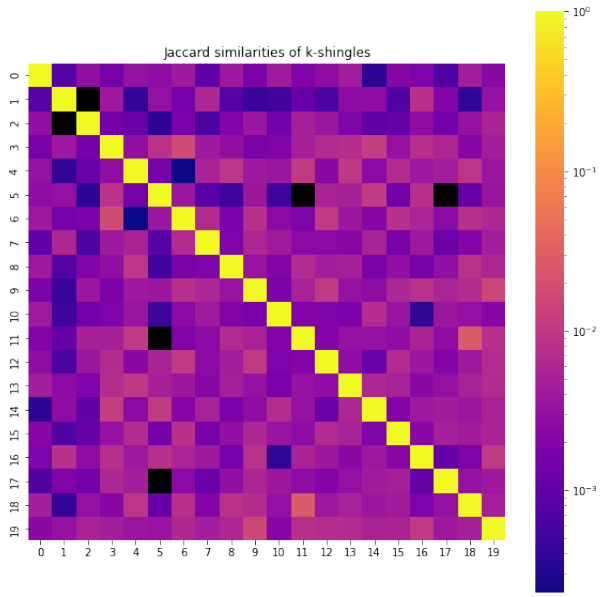


Figure 1: Jaccard similarities of hashed k-shingles

The MinHash signature similarities are shown in Figure 2. The similarities are defined on a scale of 0-1, 1 being identical. The similarities are shown on a logarithmic scale. The numbers around the heatmap are the document identifiers of the 20 documents used in the code. For the document ids refer to Table 1.

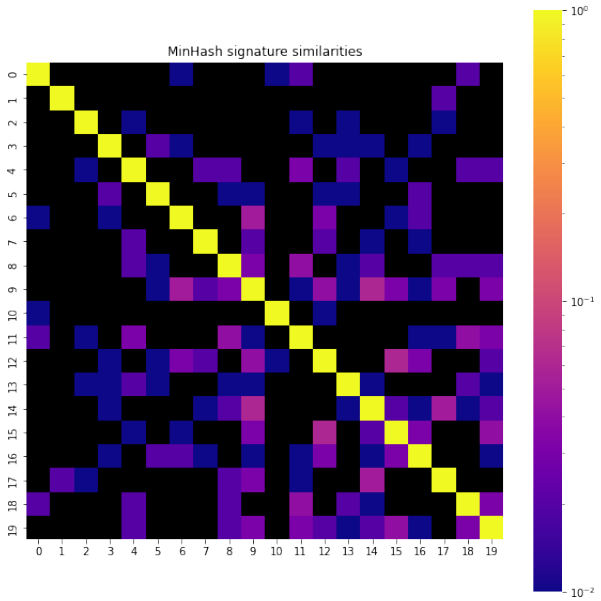


Figure 2: MinHash signature similarities

The similar candidates proposed by LSH are shown in Table 1. The candidate column includes a list of candidates proposed to be similar for each document.

| id | topic | path | candidates |
|---|---|---|---|
| 0 | business | data/bbc/business/489.txt | [0, 4, 11, 13, 14, 17, 18] |
| 1 | business | data/bbc/business/194.txt | [1, 4, 8] |
| 2 | business | data/bbc/business/461.txt | [2, 5, 7, 16, 17] |
| 3 | business | data/bbc/business/007.txt | [3, 7, 11, 12, 17] |
| 4 | business | data/bbc/business/145.txt | [0, 1, 4, 14, 17] |
| 5 | business | data/bbc/business/463.txt | [2, 5, 12, 13, 14] |
| 6 | business | data/bbc/business/294.txt | [6, 9, 10, 15, 16, 18] |
| 7 | business | data/bbc/business/505.txt | [2, 3, 7, 18] |
| 8 | business | data/bbc/business/348.txt | [1, 8, 10, 12, 15, 19] |
| 9 | business | data/bbc/business/159.txt | [6, 9, 10, 11, 15, 19] |
| 10 | business | data/bbc/business/123.txt | [6, 8, 9, 10, 15, 19] |
| 11 | business | data/bbc/business/077.txt | [0, 3, 9, 11, 17, 18] |
| 12 | business | data/bbc/business/002.txt | [3, 5, 8, 12, 13] |
| 13 | business | data/bbc/business/509.txt | [0, 5, 12, 13, 19] |
| 14 | business | data/bbc/business/370.txt | [0, 4, 5, 14, 17] |
| 15 | business | data/bbc/business/050.txt | [6, 8, 9, 10, 15, 19] |
| 16 | business | data/bbc/business/072.txt | [2, 6, 16, 18] |
| 17 | business | data/bbc/business/011.txt | [0, 2, 3, 4, 11, 14, 17] |
| 18 | business | data/bbc/business/314.txt | [0, 6, 7, 11, 16, 18] |
| 19 | business | data/bbc/business/052.txt | [8, 9, 10, 13, 15, 19] |

Table 1: LSH candidates

## 5 Discussion

Example similar document pairs picked up the Jaccard similarity of hashed k-shingles include the documents with id 11 and 18 as one pair, the documents with id 3 and 6, as shown in Figure 1. The documents with id 11 and 18 have common topics such Russia, the oil giant Yukos, a company called Yugansk, a key figure called Mikhail Khodorkovsky, and the finances of the aforementioned companies. See the texts of the documents in the appendix. The similarity of these articles are also shown in the similarity of their MinHashing signature in Figure 2. They are also proposed similar candidates in Table 1 based on LSH.

Another similar pair of documents is the ones with id 15 and 19 with common topics of growth rates, but in different contexts. This similarity is picked up by all of the Jaccard similarity of hashed k-shingles, the similarities between the MinHashed signatures, and LSH.

Nevertheless, another similar pair of documents is the ones with id 3 and 6 that have the common topics of job losses and creation in the US. This similarity is well picked up by the Jaccard similarities of the hashed k-shingles, less by the similarities of the MinHashed signatures, and not at all by LSH.

## 6 Conclusion

The efficiency of the pipelined algorithm for finding similar documents showed good potential but further tuning of the hyperparameters would likely yield better and more convincing results.

## 7 Appendix

Example Documents:

**ID 11:**

China had role in Yukos split-up China lent Russia $6bn (£3.2bn) to help the Russian government renationalise the key Yuganskneftegas unit of oil group Yukos, it has been revealed. The Kremlin said on Tuesday that the $6bn which Russian state bank VEB lent state-owned Rosneft to help buy Yugansk in turn came from Chinese banks. The revelation came as the Russian government said Rosneft had signed a long-term oil supply deal with China. The deal sees Rosneft receive $6bn in credits from Chinaś CNPC. According to Russian newspaper Vedomosti, these credits would be used to pay off the loans Rosneft received to finance the purchase of Yugansk. Reports said CNPC had been offered 20% of Yugansk in return for providing finance but the company opted for a long-term oil supply deal instead. Analysts said one factor that might have influenced the Chinese decision was the possibility of litigation from Yukos, Yuganskś former owner, if CNPC had become a shareholder. Rosneft and VEB declined to comment. "The two companies [Rosneft and CNPC] have agreed on the pre-payment for long-term deliveries," said Russian oil official Sergei Oganesyan. "There is nothing unusual that the pre-payment is for five to six years." The announcements help to explain how Rosneft, a medium-sized, indebted, and relatively unknown firm, was able to finance its surprise purchase of Yugansk. Yugansk was sold for 9.3$bninanauctionlastyeartohelpYukospayoffpartofa$27bn bill in unpaid taxes and fines. The embattled Russian oil giant had previously filed for bankruptcy protection in a US court in an attempt to prevent the forced sale of its main production arm. But Yugansk was sold to a little known shell company which in turn was bought by Rosneft. Yukos claims its downfall was punishment for the political ambitions of its founder Mikhail Khodorkovsky. Once the countryś richest man, Mr Khodorkovsky is on trial for fraud and tax evasion. The deal between Rosneft and CNPC is seen as part of Chinaś desire to secure long-term oil supplies to feed its booming economy. Chinaś thirst for products such as crude oil, copper and steel has helped pushed global commodity prices to record levels. "Clearly the Chinese are trying to get some leverage [in Russia]," said Dmitry Lukashov, an analyst at brokerage Aton. "They understand property rights in Russia are not the most important rights, and they are more interested in guaranteeing supplies." "If the price of oil is fixed under the deal, which is unlikely, it could be very profitable for the Chinese," Mr Lukashov continued. "And Rosneft is in desperate need of cash, so itś a good deal for them too."

**ID 18:**

Yukos unit fetches $9bn at auction A little-known Russian company has bought the main production unit of oil giant Yukos at auction in Moscow. Baikal Finance Group outbid favourite Gazprom, the state-controlled gas monopoly, to buy Yuganskneftegas. Baikal paid 260.75bn roubles ($9.37bn: £4.8bn) for Yugansk - nowhere near the $27bn Russia says Yukos owes in taxes. Yukos reacted immediately by repeating its view that the auction was illegal in international and Russian law, and said Baikal had bought itself trouble. "The company considers that the victor of todays auction has bought itself a serious $9bn headache," said Yukos spokesman Alexander Shadrin. He said the company would continue to make "every lawful move" to protect tens of thousands of shareholders in Yukos from "this forcible and illegitimate removal of their property". Meanwhile, Tim Osborne, head of Yukos main shareholdersgroup Menatep, said that Yukos may have to declare itself bankrupt, and that legal action would be taken, outside Russia, against the auction winners. Reports from Russia say Baikal has paid a deposit of nearly $1.7bn from a Sberbank (Savings Bank) account to the Russian Federal Property Fund, for Yugansk. The sale came despite a restraining order issued by a US court dealing with the firms bankruptcy application for Chapter 11 protection. Yukos has always insisted the auction was state-sponsored theft but Russian authorities argued they were imposing the law, trying to recover billions in unpaid taxes. There were originally four registered bidders, and with its close ties to the Kremlin, state-backed gas monopoly Gazprom had been seen as favourite. But just two companies turned up for the auction, Gazprom and the unknown Baikal Finance Group, named after a large freshwater lake in Siberia. And, according to Tass news agency, Gazprom did not make a single bid, leaving the way open for Baikal, which paid above the auction start price of 246.75bn roubles. Mystery firm Baikal Finance Group is officially registered in the central Russian region of Tver, but many analysts believe it may be linked to Gazprom. Kaha Kiknavelidze, analyst at Troika Dialog, said: "I think a decision that Yugansk should end up with Gazprom was taken a long time ago. So the main question was how to structure this transaction. "I would not exclude that the structure of the deal has slightly changed and Gazprom now has a partner. "I would also not exclude that Baikal will decline to pay in 14 days, that are given by law, and Gazprom is then recognised as the winner. This would give Gazprom an extra 14 days to accumulate the needed funds. "Another surprise was that the winner paid a significant premium above the starting price." However, Gazprom has announced it is not linked to Baikal in any way. And Paul Collison, chief analyst at Brunswick UBS, said: "I see no plausible explanation for the theory that Baikal was representing competing interests. "Yugansk will most likely end up with Gazprom but could still end up with the government. There is still potential for surprises." Yugansk is at the heart of Yukos - pumping close to a million barrels of oil a day. The unit was seized by the government which claims the oil giant owes more than $27bn in taxes and fines. Yukos says those tax demands are exorbitant, and had sought refuge in US courts. The US bankruptcy courts initial order on Thursday - to temporarily block the sale - in response to Yukos filing for Chapter 11 bankruptcy protection, was upheld in a second ruling on Saturday. The protection, if recognised by the Russian authorities, would have allowed Yukoscurrent management to retain control of the business and block the sale of any company assets. Yukos has said the sale amounts to expropriation - punishment for the political ambitions of its founder, Mikhail Khodorkovsky. Mr Khodorkovsky is now in jail, on separate fraud charges. But President Vladimir Putin has described the affair as a crackdown on corruption - and the BBCs Sarah Rainsford in Moscow says most Russians believe the destruction of Yukos is now inevitable. Hours before the auction lawyers for Menatep, a group through which Mr Khodorkovsky and his associates control Yukos, said they would take legal action in other countries. Menatep lawyers, who were excluded from observing the auction, said they would retaliate by seeking injunctions in foreign courts to impound Russian oil and gas exports.

## References

[1] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proc. 23rd International Conference on Machine learning (ICML'06)*, ACM Press, 2006, pp. 377–384.