Similar Items
Ukraine Conflict

Martin Ranieri
student no. 967158
martinvincente.ranieri@studenti.unimi.it

November 2022

In this project a research of nearly identical tweets was performed. Tweets were related to the 2022 Russian invasion of Ukraine. A locality-sensitive hashing over minhash signatures was implemented through a Map Reduce architecture.

Data

The dataset was downloaded from Kaggle [bwandowando, 2022]. It collects tweets related to the ongoing Ukraine vs Russia conflict. Only raw texts were used, metadata of tweets were not considered inside the analysis. Before going through the analysis:

- case sensitiveness was removed
- hyperlinks were removed
- punctuations, hashtag (#) and citing (@) markers were ignored
- duplicates were removed

This cleaning steps were applied through the RDD system. At the end, about 600 of thousands of tweets were left.

Scalability

Due to the high magnitude of data, a scalable implementation was deployed. Apache Spark was used as baseline framework, and PySpark python module was used to interact with the system. The list of tweets was distributed through a Spark RDD (Resilient Distributed Dataset). This distributed system would help to parallelize the workload in a distributed file over multiple nodes.

Algorithm

Following Rajaraman and Ullman [2011], a minhashing signature was computed for each tweet. Then, to find the candidate pairs of similar tweets, a local sensitive hashing (LSH) was applied.

Shingles Each tweet was represented as a set of shingles (the shingle set that describe a given string contains all its possible substrings of a prefixed length). At first, the length of shingles was set to 9 characters, but since tweet was likely to be short, the length was reduced to 5 characters to improve performance.
Then each shingle was hashed to a number that range from 0 to $2^{32} - 1$ (i.e. a 4-byte binary string) to reduce the amount of memory used, as we know that collision would be rare.

Minhashing The characteristic matrix was built considering each tweet string as a document: columns corresponded to document index numbers, rows corresponded to hashed shingles, and values were set to 1 when the document at the given columns actually contained the shingle at the given rows, 0 otherwise. To avoid sparsity, only coordinates of the matrix were stored in the RDD (as a tuple list) if they were pointing to strictly positive values.
120 linear congruential generators modulo 2^{32} were drawn. A linear congruential generator is basically a function described as $Y = (aX + c) \mod m$. By choosing wisely values for a and b, this function is expected to generate a pseudo random Y by using X as seed number (see Wikipedia [2022]). Setting m equal to 2^{32}, X shares the same value domain of Y ($\mathbf{N}$ form 0 to $2^{32} - 1$), so those generators were used as permutation function to simulate permutations of rows and therefore compute minhashes (see 3.3.5 [2011]). So, for each document, 120 minhashes were computed and concatenated to construct its minhash signature.

LSH As described in 3.4.1 [2011], LSH was applied to identify similar documents (tweets). Minhash functions (and therefore their minhashes) were clustered in 12 bands (all of them with cardinality equal to 10), thus the similarity

threshold was estimated to be **0.78**. An hash function was applied on each band chunk of all minhash signatures to bucket together those ones that could potentially be similar. Hash function for bucketing was the same for all band, but bands did not share bucket array.

Results

An adjacent list of tweets fallen in same buckets was computed, then connected components of tweets were extracted and printed out. Here below, an example of component.

1 9 3 days of full-scale #Russia's war on #Ukraine. Information on #Russian invasion. Losses of the Russian armed forces in Ukraine, September 4. https://t.co/9WNE3RuwcZ
1 9 6 days of full-scale #Russia's war on #Ukraine. Information on #Russian invasion. Losses of the Russian armed forces in Ukraine, September 7. https://t.co/QQCZQxFtOB
2 0 0 days of full-scale #russia's war on #Ukraine. Information on #russian invasion. Losses of the russian armed forces in Ukraine, September 11. https://t.co/fFRhiFBINh

As one can be seen, the three tweets are nearly identical.

References

bwandowando. Ukraine conflict twitter dataset, February 2022. URL www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows. [Online; accessed 2022-Sep-22].

Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

Wikipedia. Linear congruential generator — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Linear%20congruential%20generator&oldid=1109285432, 2022. [Online; accessed 2022-Oct-05].

Declaration

"I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study."