ACM SIGMOD PROGRAMMING CONTEST 2022

Team: QaisHousien (3rd place)
Member: Qais Abou Housien qaisabouhousien@gmail.com
Advisor: Tomer Sagi tsagi@is.haifa.ac.il
Data Management Lab @ University of Haifa

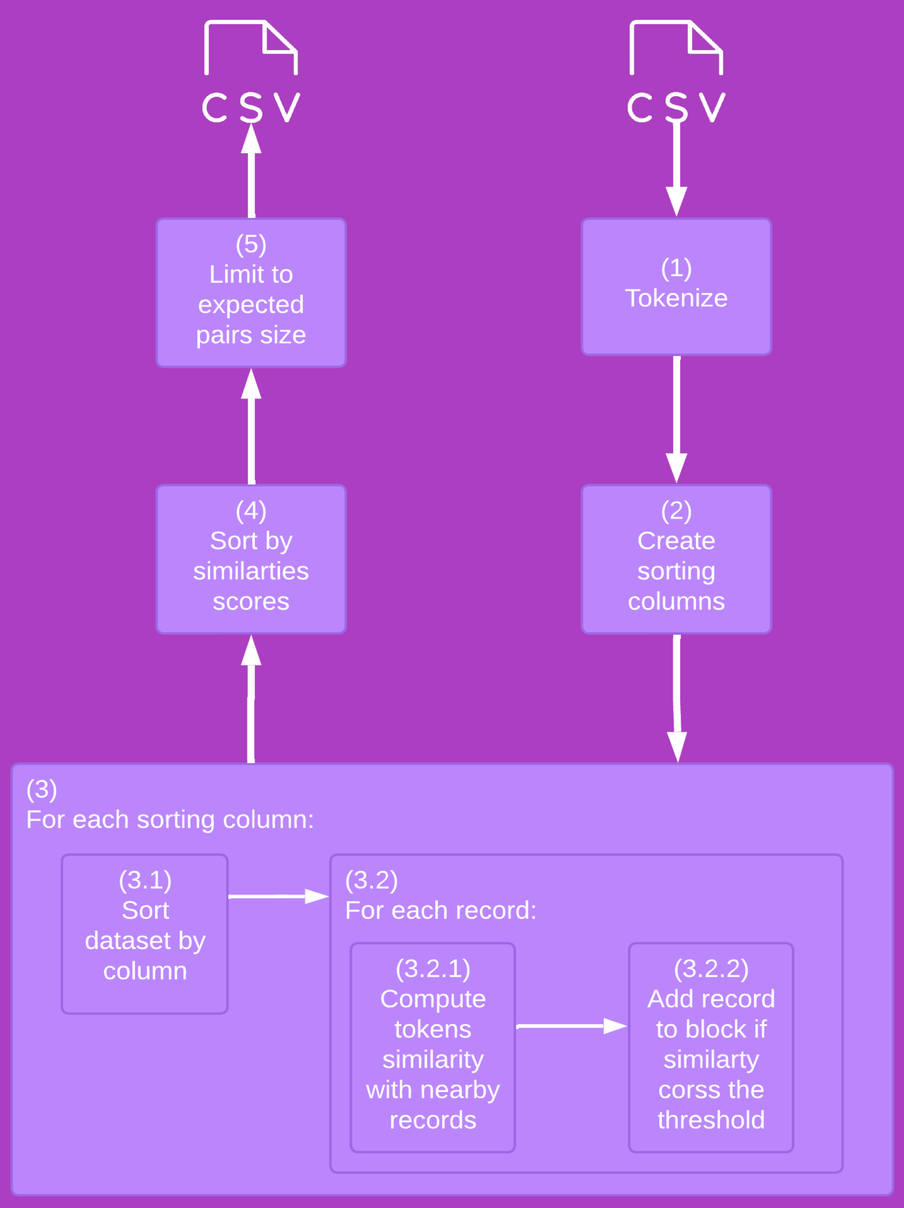
Contest Overview

- Task: Perform blocking in a limited time to generate a candidate set that contains pairs for matching
- Datasets:

#	Name	Num of rows	Num of blocking pairs
1	Notebook	1,000,000	1,000,000
2	Altosight	1,000,000	2,000,000

- Measurement: Average Recall

Solution Overview



Tokenization

- Made using the Spacy library
- Stop words are discarded
- Only nouns and pronouns are used
- Lowercasing
- Token lemmatization

Sorted Neighborhood

- The core solution is based on the sorted neighborhood algorithm
- Similarity between two records is computed using the Jaccard similarity between the two sets of tokens
- Two different variations of the algorithm were used based on the input dataset
- The first one used an adaptive window size to adjust the size based on how many records are being detected (dataset 1)
- The second one introduced the idea of *tolerance*, which is defined as the allowed number of pairs that are discarded before moving to the next record (dataset 2)

Results

#	Recall
1	0.726
2	0.301

Average recall over the 2 datasets: **0.514**