# Advanced Topics in Data Engineering

\*\*A.\*\* Use the Token Blocking (not to be confused with Standard Blocking) method to create blocks in the form of K-V (Key-value) pairs. The key for every entry will be each distinct Blocking Key (BK) derived from the entities’ attribute values and the values for each BK will be the entities’ ids. Please note that the id column in the data can be used only as reference for the blocking index and it will NOT be used in the blocking process (block index creation). Please also note that you are advised to transform every string to lower case during the tokens’ creation (before you insert it in the index) to avoid mismatches. At the end of the creation use a function to pretty-print the index.

\*\*B.\*\* Compute all the possible comparisons that shall be made to resolve the duplicates within the blocks that were created in Step A. After the computation, please print the final calculated number of comparisons.

\*\*C.\*\* Create a Meta-Blocking graph of the block collection (created in step A) and using the CBS Weighting Scheme (i.e., Number of common blocks that entities in a specific comparison have in common) i) prune (delete) the edges that have weight < 2 ii) re-calculate the final number of comparisons (like in step B) of the new block collection that will be created after the edge pruning.

\*\*D.\*\* Create a function that takes as input two entities and computes their Jaccard similarity based on the attribute title. You are not requested to perform any actual comparisons using this function.