Advanced Topics in Data Engineering

***Entity Resolution Assignment***

***Supervisor: Mr. Giorgos Alexiou***

A picture containing dark, high

Description automatically generated

**Thursday August 31st, 2023**

**Dimitris Matsanganis, f2822212**

# Abstract

This assignment focuses on the intricate domain of entity resolution and duplicate detection, utilizing advanced methodologies to augment data precision and integrity. Through a structured progression, each task contributes to a comprehensive comprehension of the methodologies' nuances.

Task A: Token Blocking for Block Creation

Task A employs the Token Blocking approach, schema-agnostic in nature, to construct blocks represented as Key-Value (K-V) pairs. All attributes except the identifier (id) contribute to generating distinct Blocking Keys (BKs) from entity attribute values. The id column remains non-participatory in the block index creation process. Lowercasing string attributes during tokenization ensures precise alignment. A function is employed to visualize the BKs in a reader-friendly manner.

Task B: Calculating Possible Comparisons

In Task B, we compute the full spectrum of necessary comparisons to rectify duplicates within the established blocks from Task A. The quantified number of comparisons offers insights into the computational intricacies of the entity resolution process.

Task C: Meta-Blocking Graph with CBS Weighting Scheme

Task C introduces the Meta-Blocking graph, implementing the CBS (Common Block Scheme) Weighting Scheme. Edges with a weight below 2 are pruned, refining the block collection and eliminating unnecessary comparisons. The pruned collection serves as the foundation for recalculating the ultimate number of comparisons, paralleling Task B's methodology.

Task D: Jaccard Similarity Function

Task D encompasses a custom function for computing Jaccard similarity, focusing on the "title" attribute. While not involving practical comparisons, this function serves as a gauge for attribute similarity.

This analysis is executed in Python, specifically version 3.10, using Jupyter Notebook. The assignment amalgamates theoretical understanding with hands-on application, fostering mastery of advanced entity resolution techniques and bolstering data analysis proficiencies.

# Contents

[Abstract 2](#_Toc142506798)

[Contents 3](#_Toc142506799)

[Table of Figures 3](#_Toc142506800)

[Assignment Description 4](#_Toc142506801)

[Case Study Outline 5](#_Toc142506802)

[I. Introduction 7](#_Toc142506803)

[II. Data Preprocessing 8](#_Toc142506804)

[Loading Data 8](#_Toc142506805)

[Data Type Consistency 8](#_Toc142506806)

[Handling Missing Values 8](#_Toc142506807)

[Token Creation from Attributes 8](#_Toc142506808)

[Cleaning Tokens 9](#_Toc142506809)

[Concatenating Attributes 9](#_Toc142506810)

[Final Data Preparation 9](#_Toc142506811)

[Conclusion 9](#_Toc142506812)

[III. Task A: Token Blocking for Block Creation 10](#_Toc142506813)

[Creating Key-Value Pairs for Tokens 11](#_Toc142506814)

[Cleaning the Key-Value Pairs 11](#_Toc142506815)

[Displaying the Key-Value Pairs 11](#_Toc142506816)

[Addressing Limitations in Jupyter Notebook and Solution 12](#_Toc142506817)

[Conclusion 13](#_Toc142506818)

[IV. Task B: Calculating Possible Comparisons 13](#_Toc142506819)

[V. Task C: Meta-Blocking Graph with CBS Weighting Scheme 14](#_Toc142506820)

[VI. Task D: Jaccard Similarity Function 14](#_Toc142506821)

[VII. Conclusion 14](#_Toc142506822)

[VIII. Deliverables 14](#_Toc142506823)

[IX. Appendix 14](#_Toc142506824)

# Table of Figures

[Figure 1: Installation and Setup of Impala via Cloudera QuickStart VM. 5](#_Toc142416818)

[Figure 2: Create the Impala database & the required tables. 9](#_Toc142416819)

[Figure 3: Provided schema. 10](#_Toc142416820)

[Figure 4: Executing the Task B's command variable type error. 10](#_Toc142416821)

[Figure 5: Create the Impala database & the required tables (revisited version). 11](#_Toc142416822)

[Figure 6: Validation procedure of task 3a. 12](#_Toc142416823)

[Figure 7: Executed command successfully adds student record in Impala. 13](#_Toc142416824)

[Figure 8: Execution of Task 3c's query in Impala with Results. 15](#_Toc142416825)

[Figure 9: Execution of Task 3d's query in Impala with Results. 17](#_Toc142416826)

[Figure 10: Results of explain command for demonstrative purposes. 18](#_Toc142416827)

[Figure 11: A sample picture demonstrating the procedure followed for the Task 4. 20](#_Toc142416828)

# Assignment Description

**Task A [30 points]**

Use the Token Blocking (not to be confused with Standard Blocking) method as a schema-agnostic approach 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, except for the id column which should ONLY be used as a reference for the blocking index and should NOT be included in the blocking process (block index creation). Please note that Token Blocking, being schema-agnostic, allows the utilization of all attributes from every entity (except id) for creating the blocks. To ensure accurate matching, it is advised to transform every string to lowercase during the tokens' creation before inserting them into the index. At the end of the creation, use a function to pretty-print the index, displaying the Key-Value pairs in a clear and readable format.

By employing the Token Blocking method in a schema-agnostic manner, you will create blocks that capture the essence of the entities' attributes. Each block's key will represent a distinct Blocking Key (BK) derived from these attribute values, excluding the id column. This approach allows for comprehensive analysis by utilizing all available attributes from each entity. Lowercasing the attribute strings during token creation helps avoid mismatches. Finally, ensure the generated index is presented neatly using a function to pretty-print the Key-Value pairs.

**Task B [25 points]**

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

**Task C [30 points]**

Create a Meta-Blocking graph of the block collection (created in Task 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.

**Task D [15 points]**

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.

# Case Study Outline

In this case study, we embark on a comprehensive exploration of advanced techniques for entity resolution and duplicate detection. In today's data-driven world, the importance of ensuring data accuracy and quality cannot be overstated. Duplicates lurking within datasets can taint analytical outcomes and introduce erroneous insights. The advanced methodologies uncovered in this case study furnish a robust framework to surmount these challenges and elevate data precision to a new level.

The systematic journey unfolds across four distinct tasks, each contributing to the development of a holistic understanding of entity resolution methodologies. These tasks guide us through the intricate landscape of data cleansing and entity consolidation, armed with techniques that stand at the forefront of data refinement. From the art of creating data blocks through Token Blocking, and the quantification of comparisons to efficiently resolve duplicates, to crafting Meta-Blocking graphs and gauging attribute similarity, this case study equips us with the precise tools essential for navigating complex data scenarios.

At the heart of this case study lies the integration of theoretical knowledge with practical implementation. Python 3.10, a powerful and versatile programming language, collaborates seamlessly with the dynamic environment of Jupyter Notebook. Together, they form the vessel for executing and demonstrating the advanced techniques outlined in this study. The ensuing sections of this case study unravel the intricate threads woven into the fabric of entity resolution and duplicate detection, inviting readers to immerse themselves in the richness of these methodologies.

**Assignment Structure:**

**I. Introduction**

The foundation is laid with an exploration of data quality's paramount importance. Within the expansive realm of data management and analysis, the specter of duplicates threatens accuracy. This assignment's core objective revolves around the investigation of advanced methodologies that elevate data integrity through precise identification and resolution of duplicates.

**II. Data Preprocessing**

In the realm of accurate analysis, the importance of data preprocessing cannot be overstated. Our primary objective is to refine and transform the dataset, making it more amenable to token blocking and subsequent examination.

**III. Task A: Token Blocking for Block Creation**

The initial step introduces us to the Token Blocking method—an avant-garde schema-agnostic technique. By leveraging all attributes except the identifier (id), we construct Key-Value (K-V) pairs that encapsulate distinctive Blocking Keys (BKs). These BKs, derived from entity attribute values, become the building blocks for data consolidation. The id column's passive role in block index creation is emphasized, alongside the importance of lowercasing string attributes for precise tokenization. The result is a structured index, elegantly visualized using a dedicated function.

**IV. Task B: Calculating Possible Comparisons**

In Task B, we tackle the computational intricacies entailed in resolving duplicates. The task entails a comprehensive calculation of all requisite comparisons within the established blocks from Task A. By quantifying the magnitude of computations required, we gain insight into the computational complexity inherent in precise entity resolution.

**V. Task C: Constructing a Meta-Blocking Graph**

Task C introduces us to the concept of Meta-Blocking, a powerful strategy that leverages the CBS (Common Block Scheme) Weighting Scheme. Assembling a graph based on the block collection from Task A, we prune edges with weights below 2—a strategic move to streamline the block collection and curtail superfluous comparisons. The pruned collection forms the bedrock for recalculating the definitive number of comparisons.

**VI. Task D: Jaccard Similarity Function**

Task D unveils a custom-crafted Jaccard similarity function, tailor-made for attribute comparison based on the "title" attribute. Though devoid of practical comparisons, the function serves as an invaluable tool for quantifying attribute similarity—a linchpin in entity resolution.

**VII. Conclusion**

As the case study culminates, the accomplishments of each task and the methodologies' intricacies converge. The assignment concludes with a reflective summary, shedding light on the multifaceted contributions of executed techniques to bolstered data quality, encapsulating the essence of accurate entity resolution and duplicate detection.

**VIII. Deliverables**

The assignments and analysis deliverables will be mentioned in this section.

**IX. Appendix**

Delve further into the world of advanced techniques with these supplementary resources. Explore Token Blocking, Meta-Blocking, Jaccard similarity, Python, Jupyter Notebook, and recommended reading for a well-rounded understanding of entity resolution.

# I. Introduction

In the intricate realm of **data management and analysis**, the **cornerstone of success** lies in unwavering **data quality** and unwavering **accuracy**. Amidst this pursuit, an **omnipresent challenge** emerges in the form of **duplicates** residing within datasets. These duplicates possess the **uncanny ability** to distort insights and introduce misinterpretations, demanding the application of nothing short of **robust and advanced techniques** for effective **entity resolution** and meticulous **duplicate detection**.

Within the **purview** of this assignment, the **profound importance** of **data quality** takes center stage. Duplicates, when left unchecked, can inflict **havoc** upon analytical endeavors, casting a **shadow of doubt** upon the **credibility** of outcomes. To address this **critical concern**, this assignment embarks on a journey to unravel a suite of **advanced methodologies** meticulously crafted to not only detect but to **precisely identify** and rectify duplicates.

Our **voyage of exploration** commences with **Task A**, where the innovative **Token Blocking method** is introduced. This **schema-agnostic approach** ingeniously creates **blocks** through the harmonious pairing of **Key-Value (K-V) pairs**. These pairs, stemming from distinctive **Blocking Keys (BKs)**, encapsulate the very **essence** of entity attributes. However, the **id column** stands distinct, reserved solely for referencing within the **blocking index**, thus serving as a marker but not actively participating in the block creation process. An important note to consider is the transformation of **string attributes** to **lowercase** during token creation, ensuring meticulous accuracy and circumventing **mismatches**. The culmination of this process is a **visually appealing** and **intelligible Key-Value pair representation**, presented through a dedicated **function**.

Advancing forward, **Task B** ventures into the domain of **computational complexity**. Here, we undertake the meticulous **quantification** of all **conceivable comparisons**. This calculation shines a **spotlight** on the intricate web of comparisons required for the **precise resolution of duplicates** within the blocks that were methodically constructed in **Task A**.

As our exploration unfolds, **Task C** unveils the realm of **Meta-Blocking**, introducing a thoughtfully constructed **graph**. This graph is birthed from the **block collection** forged in **Task A**, and it harnesses the potency of the **CBS (Common Block Scheme) Weighting Scheme**. Pruning edges with weights below the threshold of **2** stands as a **strategic decision**, streamlining the block collection and minimizing the **quantum of superfluous comparisons**. This conscientious curation forms the **bedrock** for recalculating the ultimate **number of comparisons**.

**Task D**, the final stride, delves into the creation of a function that mirrors the **Jaccard similarity concept**. This custom function, designed to compute similarity based on the attribute “**title**,” emerges as an **indispensable tool** for gauging the **likeness of attributes**. Although this function does not perform **actual comparisons** within this assignment, it is poised to play a **pivotal role** in evaluating similarity across the **expanse of attributes**.

This assignment unfurls within the canvas of **Python 3.10** and the dynamic tapestry of **Jupyter Notebook**. United in purpose, these tools merge **theoretical understanding** with **hands-on application**. As we journey through these tasks, let’s embrace the **opportunity** to **hone our skills**, cultivate a **profound understanding** of **advanced techniques** for entity resolution, and contribute to the realm of **data integrity** and **accuracy**. Through this **comprehensive voyage**, we shall refine our **expertise**, bridging the gap between **theory and practice**, and ultimately elevate our capability to wield **advanced methodologies** for meticulous **entity resolution** and **duplicate detection**.

# II. Data Preprocessing

In this section we are going to mention all the preprocessing procedures needed prior to initialize our analysis. ***We need to point out that for a better understanding you can look up for our Jupyter notebook, where our analysis has taken place*** (*See* [*Deliverables*](#_VII._Deliverables)).

In other words, to ensure accurate analysis, we must first preprocess our data. This phase forms the bedrock of our subsequent analysis, ensuring that the dataset is in the right shape and format.

## Loading Data

To begin with, every analysis starts with data. For our case study, we'll be loading a dataset named ***ER-Data.csv***. This dataset is a treasure trove of columns filled with entity attributes, which will be the core of our analysis.

## Data Type Consistency

In the initial phase, we focus on *converting* all relevant columns to ***string******type***. When dealing with multiple sources, it’s common to encounter a mix of data types. **Ensuring** that all relevant columns have a consistent data type, specifically strings, is crucial. This uniformity guarantees that operations like concatenation proceed without any type-related errors, forming a *stable foundation* for the steps that follow.

## Handling Missing Values

Next, we address the ever-present challenge of *missing values*, often represented as NaN (Not a Number). These gaps, if overlooked, can disrupt the analysis, and leaving them untreated might lead to inconsistencies. By **replacing** these NaN values and the string ‘nan’ with ‘None’, we ensure uniformity across the dataset. This step underscores the importance of a clean dataset, which is vital for subsequent analysis.

## Token Creation from Attributes

Tokens, the atomic units of our dataset, form the core of our analysis. To create these tokens, we undertake the vital step of ***converting all attribute values to lowercase and then splitting them into individual words*** *(as mentioned in the assignment's instructions)*. These tokens subsequently serve as the building blocks for the **blocking keys**, which will be instrumental in the entity resolution process. The transformation into tokens ensures that we have distinct, manageable units that *encapsulate the essence* of the information contained within the attributes.

## Cleaning Tokens

In our pursuit of perfection, we noticed that some tokens may contain extraneous characters, such as commas, especially after the split operation. These characters, though seemingly insignificant, can hinder our analysis. By **iteratively processing** each token and removing such characters, we ensure that our tokens are clean and truly represent unique words or attribute values. Furthermore, we remove the words **none** and **nan** to not be biased on the current analysis. This cleanliness is pivotal for the effectiveness of the subsequent analytical methods.

## Concatenating Attributes

With our tokens immaculately prepared, the next logical step is to *combine* them into a unified list for each entity. By concatenating the tokens derived from authors, venue, year, and title, we create a comprehensive **fingerprint** (unique identifier **id**) for each entity. This *amalgamation* ensures that we encapsulate all available information about an entity, significantly enhancing the chances of accurate matches in the subsequent stages.

## Final Data Preparation

As we edge towards concluding our preprocessing, we undertake two pivotal steps. First, we **discard** the original attribute columns, as they become redundant post token creation. Subsequently, we *reconfigure* the dataframe’s index using the **id** column. These optimizations not only make our data leaner but also streamline its structure, making it more intuitive for subsequent analysis.

## Conclusion

The data preprocessing stage, often understated, is vital in any data analysis process. By meticulously following the aforementioned steps, we have transformed our dataset into an *optimal format*. This preprocessing lays the groundwork for the token blocking method and further analysis, ensuring that each entity in our dataset is represented by a comprehensive set of tokens derived from all its available attributes. This *structured and clean format* is the bedrock upon which the success of the subsequent stages of our analysis is built.

A sample figure of the first ten rows of our dataframe after the whole data preprocessing procedures follows below:

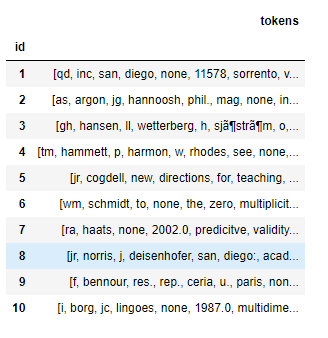


Figure 1: Snapshot of the First 10 Rows of the Final Preprocessed Dataset.

With the completion of the meticulous data preprocessing phase, our dataset is now in an optimal state for in-depth analysis. We shall commence our investigative journey with *Task A*, laying the foundation for subsequent analytical endeavors.

# III. Task A: Token Blocking for Block Creation

In accordance with the task’s directive within this assignment, we embark on the intriguing journey of **Task A** - a pursuit that dives deep into the realm of advanced entity resolution techniques. Employing the dataset presented in the **ER-Data.csv** file, this task harnesses the power of **Token Blocking** as a schema-agnostic methodology to create blocks, subsequently forming **Key-Value (K-V) pairs**.

The foundation of this task resides in the innate need for **data integrity** and the meticulous elimination of duplicates that permeate datasets. While the **Token Blocking** method shares the overarching objective of duplicate detection, it offers a distinct approach. Unlike standard blocking, Token Blocking isn’t confined to specific attributes or schemas, thus encompassing the holistic range of attributes from each entity (excluding the id column). This approach fuels a **comprehensive analysis**, drawing insights from all available attributes.

Within this methodology, the creation of blocks transpires through the ingenious fusion of **Key-Value (K-V) pairs**. Each entry’s key is ingeniously derived from a distinct **Blocking Key (BK)**, which in turn is an embodiment of the entity’s attribute values. Notably, the id column plays a unique role as a mere reference point for the blocking index, abstaining from active participation in the block creation process. A crucial nuance within Token Blocking involves the **transformation of string attributes to lowercase** during tokenization. This tactical maneuver significantly contributes to ensuring accurate matching and precludes potential mismatches that might arise from case discrepancies.

Upon the completion of this process, we culminate with an index that magnificently captures the essence of each entity’s attributes, showcased through the clarity of **Key-Value pairs**. These pairs, the embodiment of the schema-agnostic Token Blocking methodology, illuminate the unique attributes that each entity contributes to the data landscape. Through the thoughtful implementation of a dedicated function, the index is meticulously **pretty-printed**, allowing for clear and discernible observation of the Key-Value pairs.

As we embark on this Task A journey, we stand poised to unravel the intricacies of Token Blocking, where each Key-Value pair serves as a testament to our commitment to data accuracy and duplicate eradication. In parallel, we set the stage for the subsequent tasks, where these meticulously crafted blocks will play a pivotal role in the profound exploration of advanced techniques for entity resolution and duplicate detection.

*Answer of Task A:*

In this segment of our analysis, our primary focus lies in the Token Blocking method, an approach aimed at crafting blocks through Key-Value (K-V) pairs. The objective is clear: discern and record relationships between different entities based on shared tokens, facilitating efficient entity resolution.

## Creating Key-Value Pairs for Tokens

Our analysis begins with the design of a strategic dictionary. Here, each unique token metamorphoses into a key, with corresponding values being the IDs of entities that harbor these tokens: Significance of Multi-entity Keys.

More precisely, ***Significance of Multi-entity Keys*** is one of the foundational pillars of this step was ensuring keys linked to multiple entities. By setting a cut-off at a minimum of 2 entities per key, we underscored the importance of each key in the blocking mechanism.

## Cleaning the Key-Value Pairs

Following the dictionary’s inception, we pivot towards refining it ***Pruning Redundancies***. Certain tokens, while present, don’t contribute meaningfully to our analysis. Specifically, keys resonating with ‘nan’, ‘none’, and blank values were pruned to ensure the cleanliness and relevance of our constructed blocks.

## Displaying the Key-Value Pairs

With the dictionary structured and refined, our next step orbits around its presentation: ***Showcasing Interrelationships.*** We endeavored to present our Key-Value pairs in a manner that was both clear and insightful. Each token, juxtaposed with its associated entities, painted a vivid picture of interlinked relationships, all anchored by shared tokens.

A number on a white background

Description automatically generated

Figure 2: Displaying the Key-Value Pairs.

## Addressing Limitations in Jupyter Notebook and Solution

During our intricate analysis within the Jupyter Notebook environment, we stumbled upon an unexpected challenge. As the volume of data we were processing and visualizing grew, we confronted an ***error related to Jupyter Notebook's IOPub data rate limit***. The precise message was an "IOPub data rate exceeded" warning. This error acts as a safety net, designed to prevent Jupyter's server from overloading the client with vast outputs, which could potentially lead to crashes (*See Figure 3*).

The essence of this error is rooted in the default settings of Jupyter. Our expansive analysis, particularly during the display phases, was generating outputs that surpassed this default threshold. This disruption, while a minor setback, required swift troubleshooting to ensure the continuation of our workflow.

Our remedy was both immediate and effective. By invoking the command

***jupyter notebook --NotebookApp.iopub\_data\_rate\_limit=1e10***,

we were able to elevate the IOPub data rate limit considerably. This adjustment paved the way for our data-intensive tasks to operate seamlessly, eliminating the risk of triggering the aforementioned error and ***giving us the chance to output all the K-V pairs***. On top of thatm we have included a figure titled "IOPub Data Rate Exceeded Error" immediately following this section, offering a snapshot of the error we grappled with.

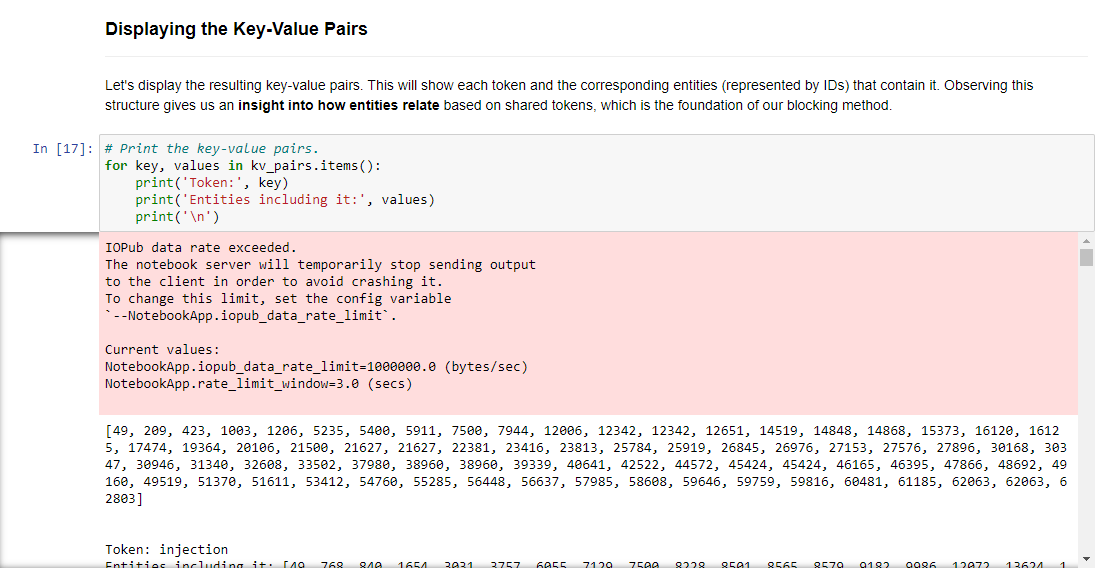


Figure 3: IOPub Data Rate Exceeded Error.

## Conclusion

The culmination of ***Task A*** signifies a blend of methodical planning intertwined with meticulous execution. The Token Blocking method, augmented by our systematic approach, ensures that we’re poised for success in the subsequent segments of our entity resolution process. Each stage, from token identification to visualization, has been engineered to elevate the accuracy and comprehensiveness of our blocks.

# IV. Task B: Calculating Possible Comparisons

# V. Task C: Meta-Blocking Graph with CBS Weighting Scheme

# VI. Task D: Jaccard Similarity Function

# VII. Conclusion

# VIII. Deliverables

# IX. Appendix