LARGE SCALE ENTITY MATCHING

Maisha Alam Meem, Professor Erik Saule (Advisor)

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Introduction

Large-scale entity matching involves identifying and linking similar records from extensive datasets.

The process is challenging due to:

- Diversity of Data
  (Heterogeneous Nature)
- Data Size
  (Millions of Records)
- Computational Complexity
  (Efficient Processing)

Objective

We need a solution that is:

- ***Robust:***
  Resistant to typos & inconsistencies
- ***Adoptable:***
  Handling Various Data Types
- ***Linear Complexity:***
  Efficiently process millions of records
- ***Scalable***:
  To multiple machines

Methodology

- Zip File Parsing:

- Efficient extraction of data from compressed files
- Seamless handling of large datasets.

```
with gzip.open(file_path, 'rt', encoding='utf-8') as zip_file:
    parsed_papers = 0

    for current_line in zip_file:
        if parsed_papers == papers_count:
            break
```

- Custom Callback Functions:

Implementation of versatile callback functions for:

- Enhanced adaptability
  (Domain-specific Rules and Logic)
- Allowing Fine-tuning
  (Unique data characteristics)

```
def run_callback(paper):
    for callback in callbacks:
        callback(paper)

    paper = None
```

Callback Functions

- K-Shingling:

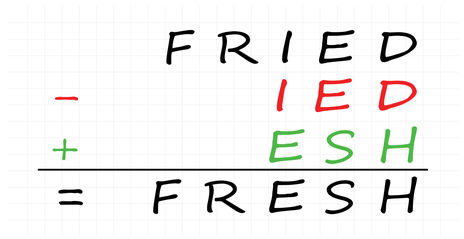
Generating Hash Table of parsed titles which is then used to calculate the Jaccard Similarity Score etc.

```
String 1: "Data Science is Exciting"
String 2: "Data Analysis is Exciting"

Now, let's use k-shingling with k = 3

String 1 Shingles:
"Dat", "ata", "ta ", "a S", " Sc", "Sci", "cie", "ien", "enc",
"nce", "ce ", "e i", " is", "is ", "s E", " Ex", "Exc", "xci",
"cit", "iti", "tin", "ing"

String 2 Shingles:
"Dat", "ata", "ta ", "a A", " An", "Ana", "nal", "aly", "lys",
"ysi", "sis", "is ", "s E", " Ex", "Exc", "xci", "cit", "iti",
"tin", "ing"

Jaccard Similarity Calculation:
Intersection of Shingles:
{"Dat", "ata", "ta ", "is ", "s E", " Ex", "Exc", "xci", "cit",
"iti", "tin", "ing"}

Union of Shingles:
{"Dat", "ata", "ta ", "a S", " Sc", "Sci", "cie", "ien", "enc",
"nce", "ce ", "e i", " is", "is ", "s E", " Ex", "Exc", "xci",
"cit", "iti", "tin", "ing", "a A", " An", "Ana", "nal", "aly",
"lys", "ysi", "sis"}

Jaccard Similarity Coefficient:
Jaccard Similarity = (Intersection Size) / (Union Size)
Jaccard Similarity = 12 / 26 ≈ 0.4615
```

- Levenshtein Distance:

Calculating Levenshtein distance among parsed titles which gives the similarity score.

- String 1: "Fried"
- String 2: "Fresh"



- Levenshtein Distance between "Fresh" and "Fried" is 4.

Conclusion

This project represents a significant step forward in a journey of data science. Through the utilization of advanced Python techniques, including zip file parsing and custom callback functions, this solution will provide a valuable contribution to the field of entity matching and data quality assurance.

Output:

```
SUMMARY
----------------------------------------
Simulation Duration: 0.65 seconds
DBLP Papers Parsed: 100
MAG Papers Parsed: 10
Total Papers Parsed: 110

DBLP hash table generated.
MAG hash table generated.

----------------------------------------
Similarity: DBLP vs DBLP
----------------------------------------
Title of Paper ID "tr/lzi/DagAnnRep2017" is similar to 1 papers with average score of 0.9615384615384616.
Title of Paper ID "tr/lzi/DagAnnRep2012" is similar to 2 papers with average score of 0.9615384615384616.
Title of Paper ID "tr/lzi/DagAnnRep2020" is similar to 3 papers with average score of 0.9245283018867925.
Title of Paper ID "tr/lzi/DagAnnRep2019" is similar to 4 papers with average score of 0.9522859216255443.
Title of Paper ID "tr/lzi/DagAnnRep2014" is similar to 5 papers with average score of 0.9541364296081276.
Title of Paper ID "tr/lzi/DagAnnRep2018" is similar to 6 papers with average score of 0.9553701015965167.
Title of Paper ID "tr/lzi/DagAnnRep2016" is similar to 7 papers with average score of 0.9562512958739374.
Title of Paper ID "tr/lzi/DagAnnRep2022" is similar to 8 papers with average score of 0.9291545718432511.
Title of Paper ID "tr/lzi/DagAnnRep2015" is similar to 9 papers with average score of 0.9533139816158686.
Title of Paper ID "tr/lzi/DagAnnRep2021" is similar to 10 papers with average score of 0.9319303338171263.
Titles having similarity score >= 0.5: 10
```

Dataset:

- Zip file of *Microsoft Academic Graph (MAG Dataset)* - TXT format
- Zip file of *Digital Bibliography & Library Project (DBLP Dataset)* - XML format