Large Scale Entity Matching for theAdvisor

Davis Spradling, UNC Charlotte
Dr. Erik Saule, College of Computing and Informatics

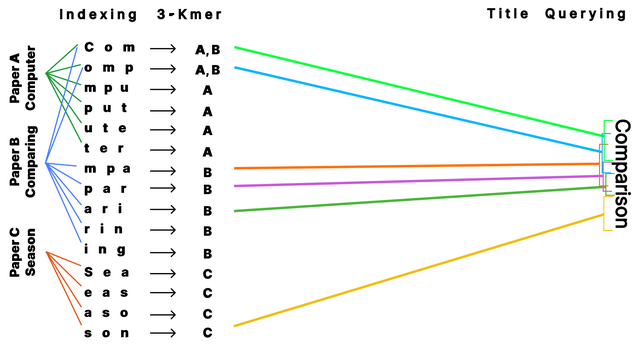
Introduction

- To analyze academic data, need good metadata
- No complete data available
- DBLP (Computer Science Bibliography)
- MAG (Microsoft Academic Graph)
- Merge these datasets through a common key (title)

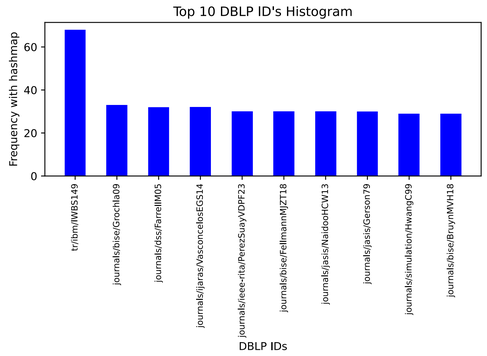
The Problem
- DBLP contains 6,000,000 papers
- MAG contains 217,000,000 papers
- Naive approach of linear searching would take 41,000 years if each comparison took a millisecond

Two Phase Method

- **Phase 1:** K-Mer Hashing: Candidate generation with k-mer hashing



Candidate matching process demonstration



Candidate matching results

- **Phase 2:** Levenshtein Distance (Future Work)
  - Take the top number of candidates through K-Mer hashing and run them through Levenshtein algorithm.

Conclusions

- Preliminary results indicate a promising method for merging dataset attributes to augment the metadata of DBLP with citation information from MAG.
- With shorter titles matching becomes harder/less likely therefore we will need to implement varying k-mer lengths for varying title lengths to optimize algorithms efficiency.

.

Future Work

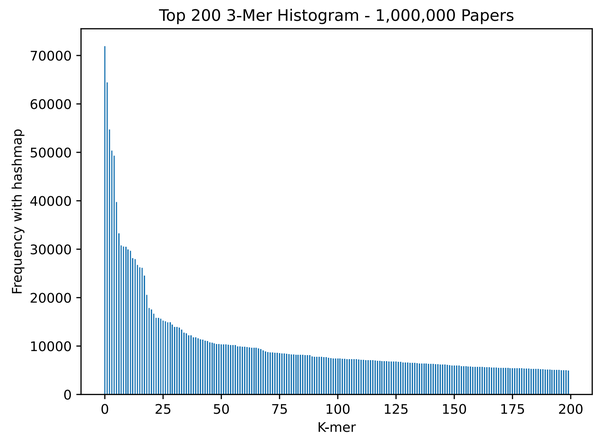
- Ongoing trials to fine-tune algorithm parameters, such as k-mer size and the number of top mers to remove.
- Progress from k-mer hashing (phase 1) to the application of Levenshtein Distance (phase 2) using candidates from k-mer hashing.
- After algorithm is complete turn into a full stack application that is available to the public.

On-Going Trial Data

K-Mer

Nums Removed	3	4	5	6	7
0	100% | 12.72	100% | 9.77	100% | 10.59	100% | 10.17	100% | 8.37
20	100% | 10.97	100% | 8.26	100% | 9.24	100% | 8.30	100% | 6.61
50	100% | 8.10	100% | 6.84	100% | 7.80	100% | 6.95	100% | 5.87
100	100% | 6.48	83.33% | 5.52	100% | 5.95	100% | 5.69	100% | 4.87
500	100% | 2.20	83.33% | 1.78	100% | 2.11	90% | 2.07	100% | 1.99
1000	100% | 0.69	66.67% | 0.51	100% | 0.69	90% | 0.72	100% | 0.71
5000	42.86% | 0.013	33.33% | 0.006	20% | 0.004	40% | 0.007	33.3% | 0.007

Average accuracy/ query time in seconds of querying DBLP dataset, removing top 40 repeating k-mers and performing Levenshtein on top 10 candidates



Top 200 K-Mers of 1,000,000 Papers

References

- Onur Küçüktunç, Érik Saule, Kaya, K., & Çatalyürek, Ü. V. (2013). *TheAdvisor*. https://doi.org/10.1145/2467696.2467752
- Brihadiswaran, G. (2020, July 2). *Bioinformatics 1: K-mer Counting*. The Startup. https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29