Bachelorproject

Rasmus Løvstad

Efterårssemester 2021

Abstract

This project regards an analysis, implementation and comparison of existing efficient solutions to the Approximate Jaccard Similarity Search Problem of estimating the Jaccard Similarity between multiple sets. The findings include both a large theoretical and practical advantage of using advanced methods as presented by Dahlgaard et al.[1] and Knudsen[2] if the user is willing to pre-process the search-corpus before performing the search. This comparison both regards the precision of different methods as well as the run time of an real-world implementation. This is based on an implementation written in a low-level systems language and benchmarked on a regular personal computer with a statistical significance. Furthermore, reflections are made on how the results might scale on more specialized hardware.

Contents

1 Introduction

2 Theory

2.1 Problem Definition

2.1.1 Jaccard Similarity

2.1.2 Similarity Search Problem

2.1.3 Approximate Similarity Search Problem

2.2 Trivial Solution

2.3 MinHash

2.4 Fast Similarity Sketching

2.5 Comparison

3 Methods

3.1 Hypothesis

3.2 Benchmarking

3.3 Expected Results

4 Implementation

4.1 Technology

4.2 Design

4.3 Assumptions

4.4 Challenges

4.5 Correctness

4.6 Benchmarking

5 Results

6 Discussion

7 Conclusion

References

[1] S. Dahlgaard, M. B. T. Knudsen, and M. Thorup, *Fast similarity sketching*, 2017. arXiv: 1704.04370 [cs.DS].

[2] J. B. T. Knudsen, *Fast similarity search with low error probability*, 2021.