EECS 405 Project Proposal

for

Improving Efficiency of String Similarity Searches through Clustered Pruning

Prepared by

James Fitzpatrick
Kyle Patterson

Version 1.0
March 18, 2014

Revision History

Name	Date	Reasons for Change	Version
James Fitzpatrick	3/17/14	Initial Proposal	

Contents

1. Introduction

2. Goals

1. Research and gain a thorough understanding of the research papers
2. Implement the B^{ed} Tree and Top-K Algorithms
3. Build clustered datasets based on the Longest Common Subsequence on:
    - Movie Titles
    - Publication Titles
4. Test the effects on the performance of the proposed clustering techniques
5. Formally present findings through a final presentation

3. Division of Labor

James Fitzpatrick will be implementing the Top-K Algorithim as documented in the papers below. James will also create an agnostic test harness to assess the differences between the different seardh algorithms and the different clustering techniques on the datasets.

Kyle Patterson will be implementing the B^{ed} Tree Algorithm and will be developing the clustering algorithms for the datasets.

4. Papers Identified