CS 453: Project 2 - Data Preparation

Project Report

Creepy - Data Cleanser

Travis Hall
trvs.hll@gmail.com

Brittany Thompson
miller317@gmail.com

Bhadresh Patel
bhadresh@wsu.edu

Washington State University Vancouver
October 03, 2010

Abstract

The main goal of this project is to design and implement data preparation or data cleansing step of the search engine. We already collected small collection of pages via the web crawler. We will be removing all unnecessary portions in the document collection through a combined process of (i) tag-stripping, (ii) tokenizing, (iii) stopping, and (iv) stemming. Additionally, we will be preparing a document graph for link analysis.

1 Overview

2 XML Document Graph

3 XML Validator

4 Tag-stripping

5 Tokenizing

6 Stopping

7 Stemming

8 Roles

Travis Hall Link analysis graph and XML validator.

Brittany Miller Stopping and stemming.

Bhadresh Patel Tag-stripping and tokenizing.

9 Test Environment

For testing/production purpose, we set up a machine instance on Amazon EC2. The instance id of the machine is `i-4fb4ec25`. The source code is checked out at `/home/ubuntu/creepy/`.