Project Report

Luojie Xiang, Junchao Yan

1 Hadoop Setup

Describes how we set up our 6 node hadoop cluster.

2 Data Cleaning

Describes the steps we take to do data cleaning.

remove content within ¡code¿
remove tags `"<[^>]*>"`
Remove punctuation `"[!@#$%^&*()-\_:;"'?/.,<>]"`
Remove new lines `"\n" to ""`
Remove multiple white spaces `"\s+" to " "`
Remove words shorter than length 3
Turn all letters to lowercase