Algorithms

Team11-credit

Algorithm followed to detect spam sites is calculating the frequency of each word as a percentage of the whole document ,if the percentage of any word exceeded a specified allowed percentage the page is marked as spam.
For the seeds we chose the pages with the most out degree as they refer to many sites and will probably bring diverse topics.
When crawling new urls are checked that they are not self referential and are appended with a slash at the end to avoid visiting the same page once for url with slash and once for the one with no slash.Also a limit is set for fetching urls from the same domain as the current page.