Вариант 4: Веб-робот.

Для того, чтобы найти все ссылки на странице, запускается код на питоне. В ***generate*** берется стартовый набор ссылок (ссылки, на которые можно перейти со страницы https://ru.wikipedia.org/wiki/Linux, всего их 280), далее задается рандомное число от 0 до 280 и открывается выбранный сайт, далее берутся все ссылки с него и записываются в итоговый сгенерированный файл. (Лучше не генерировать файлы таким образом, потому что будет много ссылок, и в дальнейшем будет очень долго работать).
Функция ***map*** разбивает весь файл на отдельные файлы (по 60 строк), затем каждой ссылке дописывает в качестве значения 0 -- значит данную ссылку мы еще не посещали. Затем все файлы сливаются в один.
На этапе ***reduce*** сначала происходит внешняя сортировка. Указанный файл считывается по 50 строк в файл, каждый из полученных файлов сортируется (добавляла строки в сет). Далее сливаем все отсортированные файлы в один, используя сет, таким образом получим отсортированный файл со всеми данными. Далее данный файл разбиваем на файлы по ключу. Так как ключей может быть много, и много потоков запустится одновременно, то разбиваем по 60 ключей, применяем к ним reduce, ждем завершения, и только тогда продолжаем обрабатывать остальной файл. После этого имеем файлы, разбитые по ключу. К каждому файлу применяем reduce. Он для данной в файле убирает все повторения и ставит 0 или 1, в зависимости от того, была ли данная ссылка посещена. Далее все такие файлы с уникальными ссылками сливаются в один файл. Теперь к этому файлу можно применить скрипт питона и перейти по каждой уникальной ссылке, если она еще не была посещена. Все полученные после перехода на сайт ссылки записываются в отдельные файлы, затем эти файлы объединяются. Получили новый файл на глубине захода +1. Но в данном файле ссылки могут повторяться. Поэтому применяем к нему всю операцию reduce. Так продолжаем до нужной глубины захода. Получаем итоговый файл с ключом -- url страниц и значением 0 или 1. Проходимся по файлу и заменяем в каждой строке значение на “”.

Как запускать:
./generate filename.txt
./main map ./map input.txt in.txt 2
./main reduce ./reduce in.txt output.txt 2

(в файле input.txt одна строка, в полученном файле output.txt 321 строка. Если добавить больше строк во входной файл, то будет долго работать (попадались сайты, где очень много ссылок, даже один такой сайт работал больше минуты), также была проблема, что не все полученные ссылки нормально открываются).