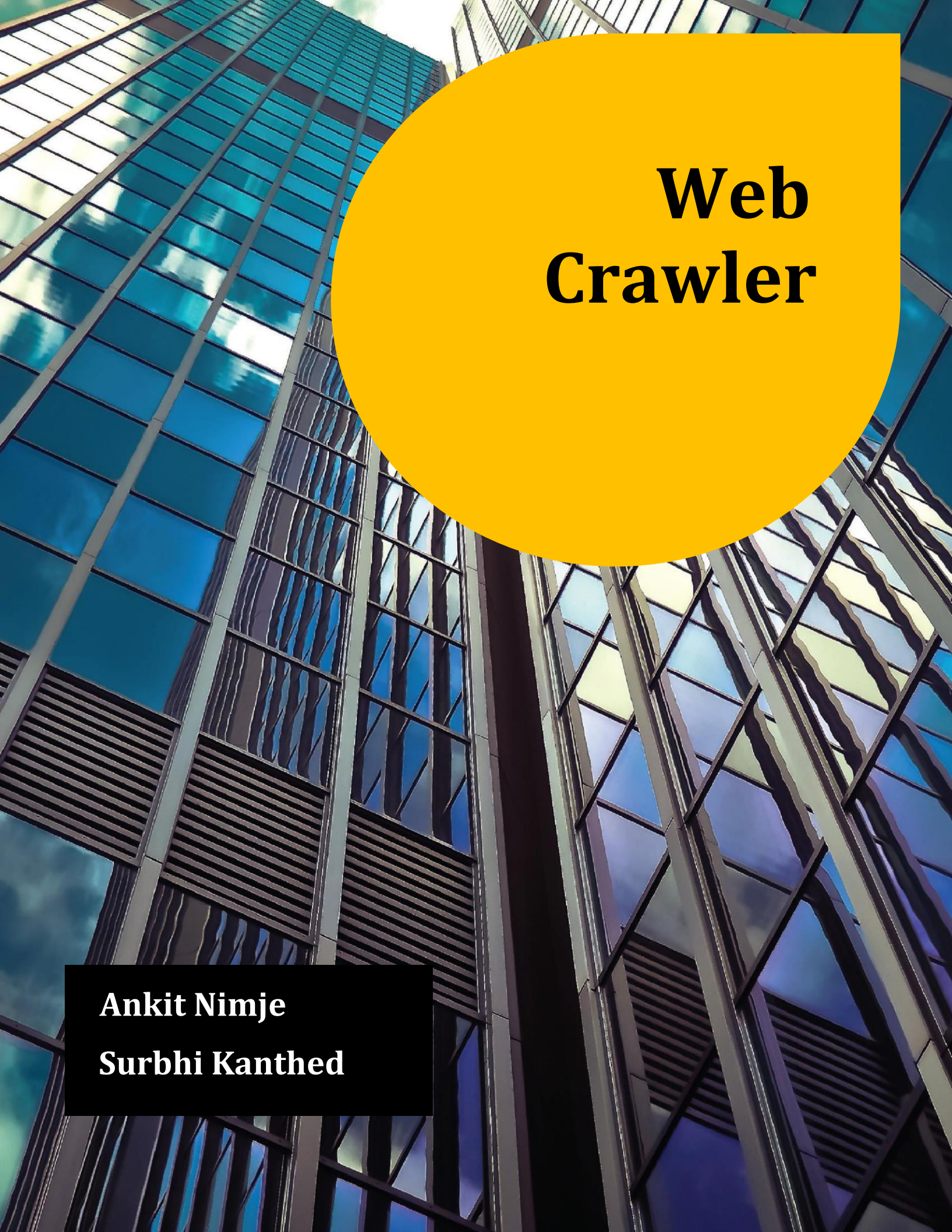

Web Crawler

Ankit Nimje

Surbhi Kanthed

We successfully managed to develop a web crawler after carefully studying various sources and came up with two techniques Depth First Search (DFS) and Breadth First Search (BFS).

In DFS, like a traditional DFS technique, starts with first frontier and goes as far as it can go to the depth before backtracking to next same level node. In other words, if we provide a single website as frontier, it will fetch all hyperlinks and crawl very first link found by it. In the same manner our crawler will go to the depth and keep crawling until out of links or if termination condition exists. There is a drawback of traversing using this technique as a crawler can go off topic (out of focus) after series of links and provide unwanted results.

BFS technique as well works like traditional BFS graph algorithm, where it starts with root node (frontier) and traverse through all the neighbor nodes before moving to next level. To explain, for example it starts with one frontier and fetches all links and traverse through all those links and at the same time adding hyperlinks it found in those crawled links to our seed list. Using this technique, there is a small probability of losing focus and going off topic.

In coming week, we will be working on focused crawler which collects webpages that satisfy certain properties for example searching for ‘baseball’ or pages with large page rank.

References:

1] S. Chakrabarti, M. Berg, and B. Dom,” Focused crawling: a new approach to topic specific Web resource discovery”, Computer Networks, vol. 31, pp. 1623-1640, 1999.

2] M. Najork and J. Wiener,” Breadth-first search crawling yields high-quality pages”, in Proc. 10thInt. World Wide Web Conf., pp. 114-118, 2001.