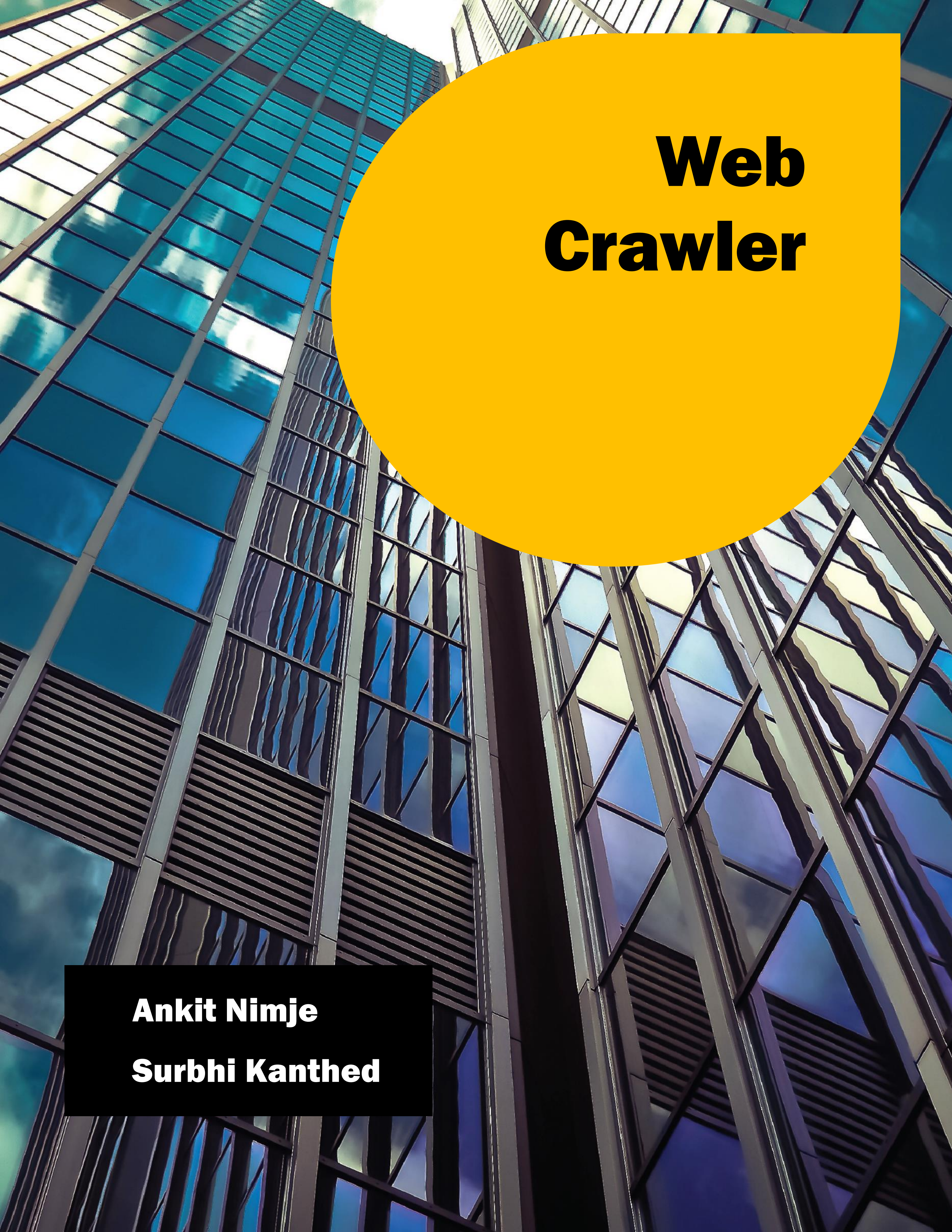

Web Crawler

Ankit Nimje

Surbhi Kanthed

Problem Statement:

Today we are dealing with enormous amount of structured as well as unstructured data. To retrieve information manually become very time consuming and tedious task. In today's era there are lot of competitor and to remain updated about each every competitor's would be impossible without some automated search.

Introduction to Web Crawler:

According to Wikipedia, "A **Web crawler**, sometimes called a **spider** or **spiderbot** and often shortened to **crawler**, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web Indexing (*web spidering*)." Web crawlers are basically used in search engines such as Google, Yahoo, etc. and acts as a backend support for such large search engine giants. They are also used in large database systems with very large datasets such as insurance companies, medical history, university systems, etc.

A web crawler is a program or application or robot which browses through Internet in progressive or automated manner. It starts with a list of websites to visit (also called as seeds) and identifies all hyperlinks in the current website and adds them to list of URL's (also called as crawl frontier). Thus, gradually progressing to create an index of terms which will be used for future query searches.

Objective:

We plan to develop an application or web application based on web crawlers which will successfully crawl through World Wide Web and create an Index of most relevant terms for records. Talking about search engines, best search engine is the one which provides most relevant results with passed query. Our goal is to develop an application which will scan through all recorded indexed terms and find most relevant site which matches our query.

References:

1] INTRODUCTION OF THE WEB CRAWLING FOR SEARCH ENGINE (ShodhGanga)

2] https://en.wikipedia.org/wiki/Web_crawler

3] https://blog.datafiniti.co/what-is-web-crawling-9184d019e094