Exploring Information Retrieval Techniques Through Programming Assignment 1

João Vítor Fernandes Dias
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
joaovitorfd2000@ufmg.br

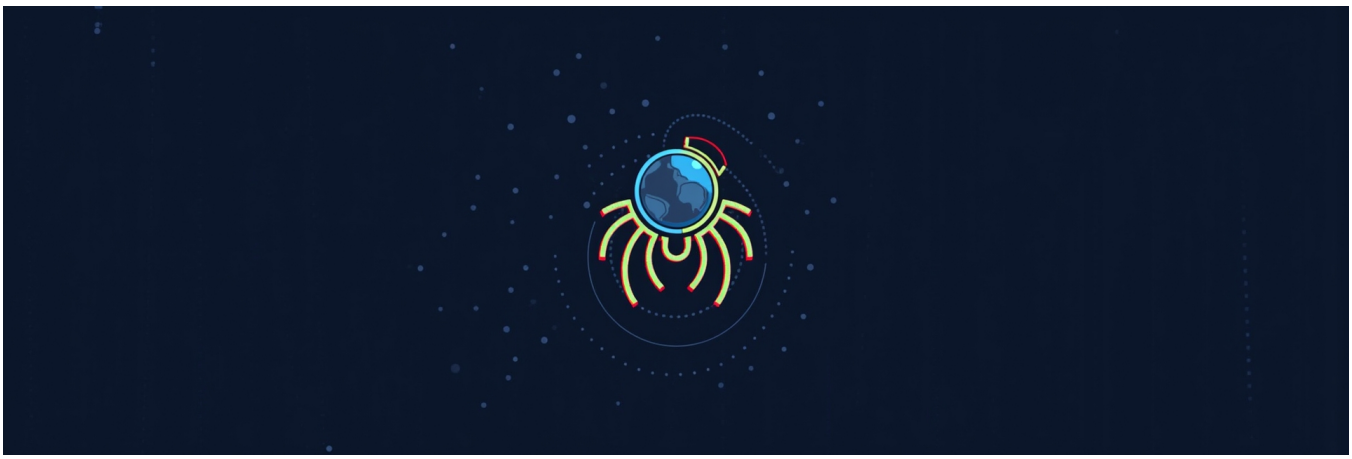

Figure 1: Web Crawler Logo

Abstract

This article is summary of the implementation of the first programming assignment of the Information Retrieval course at the Federal University of Minas Gerais (UFMG). The assignment consists of programming a web crawler using Python 3 to scrape 100.000 unique pages efficiently, respecting certaing policies. The crawler must be distributed, using the multiprocessing library to take advantage of multiple CPU cores. After crawling, the pages must be stored into a WARC file, which is a standard format for archiving web pages, and zipped to save space.

CCS Concepts

• **Information systems** → **Information retrieval**; *Information extraction*; *Retrieval effectiveness*; *Retrieval efficiency*; *Distributed retrieval*; Data structures; • **Social and professional topics** → Acceptable use policy restrictions; Student assessment.

Keywords

Information Retrieval, Web Crawler, Python, WARC, Multiprocessing

ACM Reference Format:
João Vítor Fernandes Dias. 2025. Exploring Information Retrieval Techniques Through Programming Assignment 1. In *Proceedings of Information Retrieval (IR '25).* ACM, New York, NY, USA, 1 page. https://doi.org/XXXXXXX.XXXXXXX

IR '25, April 01–28, 2025, Belo Horizonte, MG

ACM ISBN 978-1-4503-XXXX-X/2025/04
https://doi.org/XXXXXXX.XXXXXXX

Received 31 March 2025; revised 28 April 2025