Open Secrets: Investigating systemic biases in Wikipedia, the free online encyclopedia

Eamon Ma, Hisbaan Noorani, Rachel Xie, Philip Harker

2021-04-12

1 Introduction

etc.

2 Computational overview

As lazy people, we don't like to do more work than necessary. That is why we were ecstatic to find the Python library `wikitextparser` to do work for us—the most important being extracting links from a given article. Unfortunately, this functionality of the library was not written with performance as its highest priority. We can see experimentally that the library takes on average about 30 seconds on a modern computer to parse through an XML file approximately one million lines long. (Note that though the parser is written for parsing wikitext, in contrast to XML, the mere *presence* of markup does not affect the functionality.) Extrapolating, the 1.3 billion line Wikipedia dump would take about $1300 \cdot 30$ seconds, or under 10 hours.

The authors believed they could do better—and do better we did. Diving into the code for `wikitextparser`, we see the project relies on regular expressions as a primary method of string manipulation. Unfortunately, this tool falls short for large files. Instead, we decided on a hybrid approach: match all the wikilinks using a regular expression, then conduct further processing on each wikilink with string operations. Immediately, we see a significant improvement: on that same million line file, the string operation implementation takes under 1.3 seconds over a 10 run average. This is more than an order of magnitude faster than the library's implementation—about 20-23 times faster, to be precise. This allows us to reduce the projected computing time from 10 hours to about half an hour—quite the improvement!