Authors: Kyla Kaplan, Elvira Baltasar Jiménez

Date: November 11, 2023

Information Retrieval — Progress Report

We were assigned the following topic for the Information Retrieval final project: "*Choose at least 3 websites that report news from* ***artificial intelligence topics***." Before we proceeded with the project, we first found the sources from which we will be looking to source our artificial intelligence news. Upon doing an initial search, we shortlisted 6 different news channels; some were exclusively about artificial intelligence, and for some, artificial intelligence was just a sub-topic of a larger publication. We weren't sure how this would affect the Scrapy Python library, so we kept even those in our initial choice for the time being.

Continuing further, we took our shortlist, and checked the "*/robots.txt*" route on all these sites, keeping in mind the websites that had 'easier' access than others. It happened to be that for the publications for which artificial intelligence news was just a 'sub-topic' of a larger publication, had certain web scraping guards set up for all their artificial intelligence routes — so that took our list of 6, and shrunk it to 3. Here they are:

- https://news.mit.edu/topic/artificial-intelligence2
- https://www.artificialintelligence-news.com/
- https://venturebeat.com/category/ai/

We initialized our team's GitHub (please see here: https://github.com/kybeka/ir-babes).

Now that we have established our list, we took our 3 websites, and began inspecting their DOM's; along the way, deciding on the information that we do need —

1. Publication sourced from (URL)
2. Title of the article
3. Short description of the article
4. Date of publication
5. Tags/Related Topics (present on some of the websites)
6. Images (we are still figuring out how to import this)

We are expecting that we will need to see more patterns, but at the moment we have begun launching the crawlers locally on our machines, and trialing to see how it works. Our next steps are to be able to aggregate further pages on these websites. Then we will push to our repository the data. After that we will proceed with indexing via PyTerrier.