Xtrax Data-Sources Design Document

Overview

Inspiration from OpenAI’s GPT-3

An interesting time to use these new API’s to explore machine reading and machine understanding of technical content.

General Framework

OAI-PMH Implementation – Implement a OAI Client using the http-kit async API.

Data-Sources

arXiv Data-Source

Strategy:

* Use the s3cmd utility and s3 data sources to boot strap the primary content – however, I don’t believe that that data supplies the BibText/DOI links and/or the references or citations data. So, additional sources are needed.

Metadata sources

* **OAI-PMH** - arXiv supports the OAI protocol for metadata harvesting (OAI-PMH) to provide access to metadata for all articles, updated daily with new articles. This is the preferred way to bulk-download or keep an up-to-date copy of arXiv metadata.
* **API** - arXiv supports real-time programmatic access to metadata and our search engine via the arXiv API. Results are returned using the Atom XML format for easy integration with web services and toolkits.
* **RSS** - arXiv provides RSS feeds of new updates each day. These are intended primarily for human consumption but do use well defined XML formats and thus might be useful to machine applications.

A manual operational Robot:

We ask that users intent on harvesting use the dedicated site <https://export.arxiv.org> for these purposes, which contains an up-to-date copy of the corpus and is specifically set aside for programmatic access. This will mitigate impact on readers who are using the main site interactively.