Senior Project, 2018, Fall

Web Page Archiving and Content Analysis

Student: Adam Tahoun, Florida International University
Mentor: Dr. Mark Finlayson, Florida International University
Professor: Dr. Masoud Sadjadi, Florida International University

Problem

- Temporary web content
- Filtering and parsing is time consuming and difficult
- Content may need translation
- Content not available for offline access and manipulation
- No solution exists

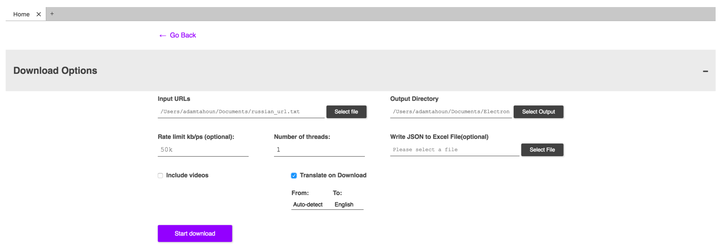
Solution

- Custom file format
- Improved GUI for viewing and interfacing with custom file format
- Introduced deep learning translation API
- Introduced machine learning approach to content extraction
- Cross-platform

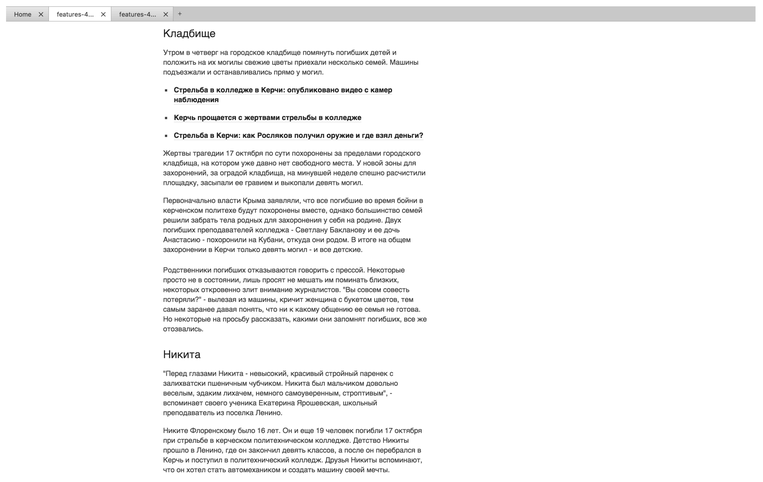
Implementation

- Python2.7 and wget for the downloading CLI
- Electron, HTML5, CSS3, and JS for the GUI
- Dragnet for the content extraction
- PyTorch and OpenNMT for the translation

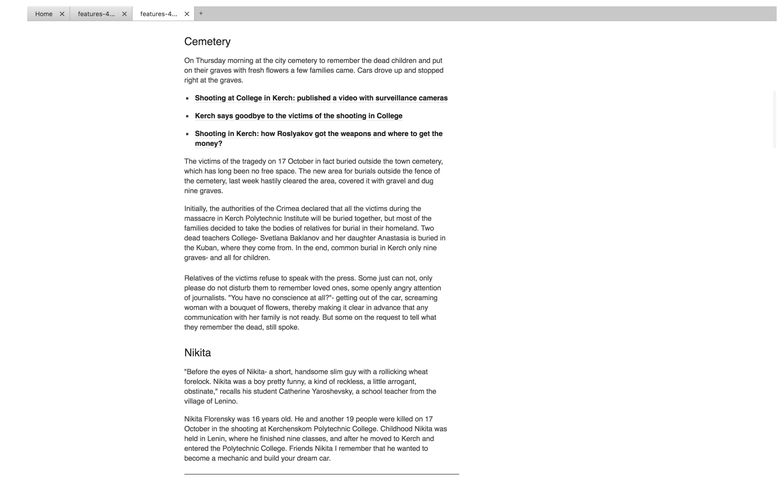
Screenshots



Download Screen



Original Article



Translated Article

Current System

- Previous projects were incomplete
- Existing partial solutions have been deprecated

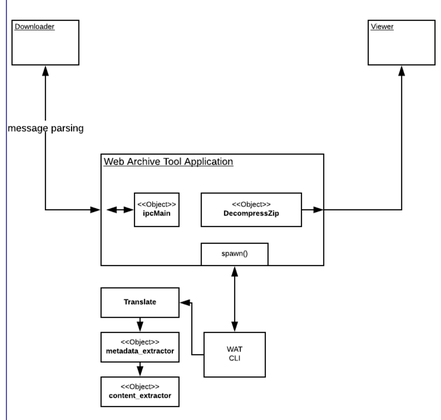
Requirements

- Batch download large sets of pages
- Encapsulate sets of pages in a session
- Continue session if download was interrupted
- Archive web page in format that allows easy programmatic access
- Encapsulate this archive in a single file
- Ability to easily view web archive
- Identify, translate, and extract the main textual content of the article
- Identify and extract the metadata of the article
- Cross-platform

Verification

- Unit and integration testing
- Tested on 5GB of articles
- Tested on macOS, Windows, and Linux
- Content Extraction: 90% accuracy
- In-house Translation: 70% accuracy

System Design

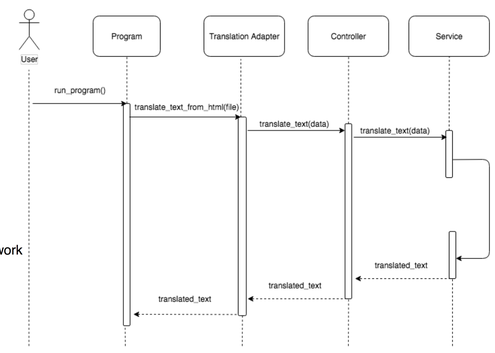


Overview of System

End-to-end Translation Algorithm

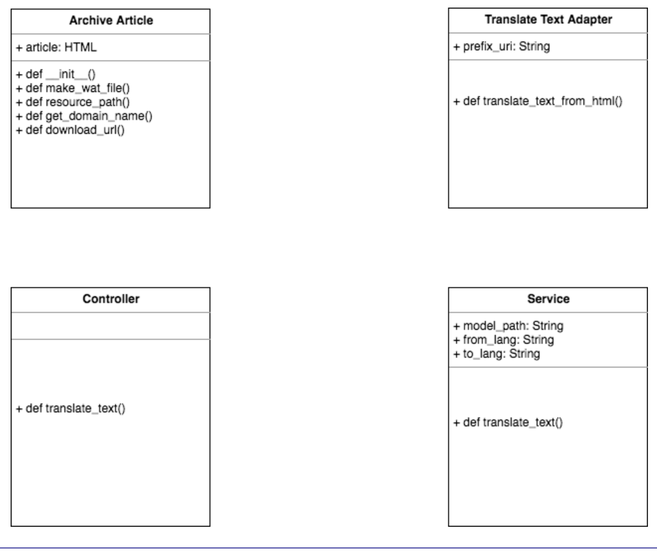
1. Traverse DOM of webpage and create list of text from visible divs in HTML.
2. For each text element, tokenize and encode the text into vector representation.
3. For each vector, predict its translated vector representation using a recurrent neural network with LSTM cells.
4. Decode the vector into its word representation of desired translated text.
5. Traverse the DOM and replace text nodes with the translated version.

End-to-end Translation Algorithm



Overview of End-to-end Translation

Object Design



Summary

Our work allows users to:

- Download and archive batches of web pages
- Translate and extract important information from the web pages
- View these web pages on any platform

Acknowledgment

The material presented in this poster is based upon the work supported by Mark Finlayson, Andres Cremisini, and Pedro Torres-Carrasquillo. I am thankful for the help that I received from my group member, Oscar Martinez.

Follow us @FIUSCIS