STAPI: An Automatic Scraper for Extracting Iterative Title-Text Structure from Web Documents

Nan Zhang[1], Shomir Wilson[1], and Prasenjit Mitra[1,2]
[1]College of Information Sciences and Technology, Penn State University, USA
[2]L3S Research Center, Germany

Motivation

Many formal documents are organized into sections of text, and each section comes with a title. This iterative title-text structure can be valuable for various natural language generation tasks.

However, extracting this structure from web documents is difficult due to **the lack of corpus**, **inconsistent HTML writing styles**, and **the existence of irrelevant texts**.

Inconsistent HTML Writing Styles

Visually similar title-text representations can have completely different HTML structures. Therefore, a structure extraction system should be flexible enough to handle web documents with different writing styles.

Visual Representation	HTML Structure
Data in the Aggregate We may disclose to prospective partners, advertisers and other selected third parties aggregated user statistics data (for example, a statistic indicating that 45% of our users are female) in order to describe our services to these third parties, and for other lawful purposes.	
USE OF COOKIES The Website uses "cookies" to help you personalize your online experience. A cookie is a text file that is placed on your hard disk by a web page server. Cookies cannot be used to run programs or deliver viruses to your computer. Cookies are uniquely assigned to you, and can only be read by a web server in the domain that issued the cookie to you.	
1. You can choose not to receive some types of advertising online, on your satellite TV service or on your wireless device. • **Relevant Advertising:** Opt-out of Relevant Advertising delivered by AT&T here.	 <a> ....

Dataset

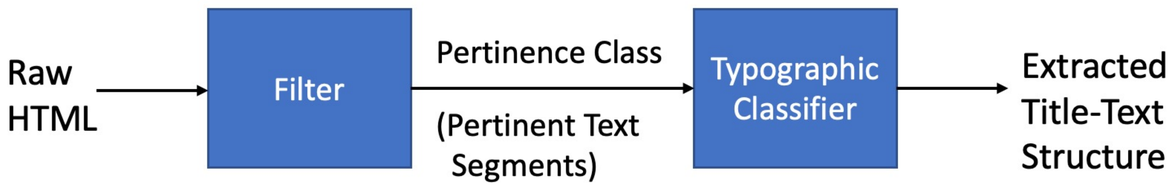
There does not exist a dataset that contains annotated web documents with the label for whether each text segment is a title or prose text. Thus, we annotated three types of web documents: privacy policy (PP), terms of service (TOS), and miscellaneous topics (MISC).

Annotating 291 documents took more than 43 consecutive hours.

The annotation process was completed manually by adding a label to every relevant text piece. Titles must differentiate themselves from prose text by having different visual or syntactic clues.

Type	# Documents	# Pertinence	# Title	# Prose
PP	144	12500	2838	9233
TOS	99	9476	2142	7099
MISC	48	2363	506	1648

Approach



We present STAPI (Section Title And Prose text Identifier), which can automatically extract iterative title-text structure from web documents.

STAPI is a 2-step system. Its first step involves a filter that filters out unrelated content like document headers and footers. Its second step involves a typographic classifier that performs title-text classification.

Training: XGBoost + oversampling on the minority class + one-hot encoding to categorical features.

Morphological characteristics: Number of words, number of punctuation symbols, etc.

Visual clues: HTML tag name, HTML ID, etc.

Spatial information: Relative position.

Semantic feature: Average sentence encoding by BERT.

Results

STAPI is shown to predict unseen data well.

	Weighted Precision	Weighted Recall	Macro Precision	Macro Recall
Filter	0.933	0.933	0.930	0.932
Typographic Classifier	0.968	0.967	0.893	0.872

STAPI outperforms existing baseline models (in terms of weighted F1 score) on the title-text classification task.

	PP	TOS	MISC	All combined
Baseline 1	0.797	0.800	0.733	0.773
Baseline 2	0.854	0.884	0.831	0.869
STAPI	**0.965**	**0.977**	**0.923**	**0.976**

We also examined the necessity of STAPI's filter and evaluated the importance of each feature via an ablation study.

Conclusions

We contribute:

- A **dataset** of web documents that comes with label for whether each text segment is a title or prose text.
- An **automatic tool** that extracts title-text structure and achieves state-of-the-art performance.

The Human Language Technologies Lab at Penn State http://hltlab.org

PennState College of Information Sciences and Technology

Web app

Dataset & source code