

Master Thesis Defense

Advancing and Benchmarking Large Scale Content Extraction from the Web

Presented by:
Sanket Gupta

Date:
13.12.2022

Introduction

Introduction | Sections of a web page

makeawebsitehub.com

Make A Website Start A Blog Reviews Resources Blog About


Home - Business - Websites With Ads - Websites That Effectively Use Ads To Monetize Their Content

Websites With Ads - Websites That Effectively Use Ads To Monetize Their Content

by Jamie Spencer

Websites With Ads

Sites That Effectively Use Ads To Monetize Their Content



In the early days of the Internet, ads were seen as an eyesore for websites. Not only did they affect the look of web pages, but they also had a knack for decreasing performance. They often turned websites into clunky digital spaces that took forever to load.

A lot has changed since then. Today, digital advertisements are an important part of the puzzle. Web users have grown accustomed to seeing them and website owners rely on them to make a profit. They're an important monetization tool that many businesses rely on to stay afloat, although there are plenty of free advertising opportunities online.

While the placement of ads may seem arbitrary, there's a lot of careful strategies involved. There's a fine line between affecting the user experience and implementing

Over 1 Million Monthly Readers

Follow Us

Join our email list here.
Get fresh content straight to your inbox
from Makeawebsitehub.com

Special Hosting Offer For My Readers! - \$2.75 p/m + Free Domain

Exclusive Hosting Offer...

Make your side hustle idea come alive!

Create your own site!
Hosted WordPress blog
+ Free domain (.com)
+ 24/7 Support
+ Free SSL certificate
+ One Click Install

Let's go!

A Huge List Of The Most Subscribed YouTube Channels In The World

The Fastest WordPress Hosting Providers 2022 - If You're Not Fast...You're Last!!!

Blog Vs Vlog - Which Is Better?

Best Month To Month Web Hosting Deals - 2022

Header

Promotion

Ad

Links

Leave a Comment

Name *

Email *

☐ Save my name, email, and website in this browser for the next time I comment.

Post Comment

Feedback

Best of the Blog

Online Business Ideas that Work

Best Affiliate Marketing Platforms

Best Blogging Sites

Create A Professional Email Address

Affiliate Marketing Glossary

Free Web Hosting

Product Comparisons

Cool Blog Examples

Latest Hosting Reviews

Best VPS Hosting

Best Magento Hosting

Best Cloud Hosting

Best Dedicated Hosting

Cheap VPS Hosting

Hosting Coupons

Best UK Web Hosting

Best Web Hosting For Australia

Site Builder Reviews

Shopify Review

Weebly Review

Wix Review

Site Links

Affiliate Disclaimer

Cheat Sheets

Black Friday / Cyber Monday

Privacy & Cookie Policy

Footer

Reference: <https://makeawebsitehub.com/websites-with-ads/>

3

Introduction | Sections of a web page

Boilerplate

Any text that can be reused across multiple web pages or the text that is not important to the reader, or benefit the business process.

- **Header:** Navigation menu with links to internal sites.
- **Footer:** Copyright notice, business information
- **Advertisement:** Monetize content based on user search history.

Main Text

Section interesting or important to the user containing relevant information such as new article, sports events, informative post.

Section of a webpage that can be utilized optimally for text mining or driving business processes.

Introduction | Content extraction

Importance of Content Extraction:

- Displaying only relevant information on smaller devices.
- Improving search engine results.
- Eliminating possible boilerplate from unstructured data can be of immense value before employing Natural Language Processing.
- Reduced cost storage.

Introduction | Content extraction

Manual Content Extraction

- A process where reader visually examines a webpage, identify main text and extracts it.
- Less efficient.
- Expensive if done at scale.

Automated Content Extraction

Using software or any other automated technique.

- Scalable.
- Efficient.

Introduction | Content Extractors

Content Extractors perform content extraction from a web page.

Simple Extractors

Removes minimal to no boilerplate and extracts the entire body tag text.

Main Extractors

Removes all possible boilerplate from the web page and deliver main content.

Introduction | Content extractors

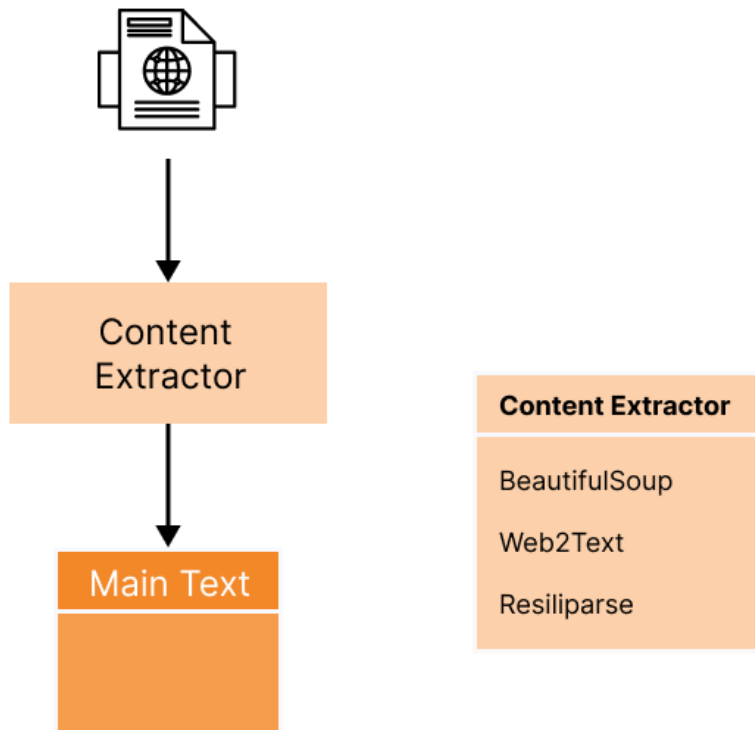
- Benchmarked 15 Open-Source Extractors.
- 9 Main Extractors and 6 Simple Extractors.

Contribution

Contributions

- Collected and Analyzed 15 Content Extractors.
- Collection of labelled web pages for benchmarking extractors.
- Exploratory data analysis on the labelled web pages.
 - Quality checks on the gold standard.
- Designed an Ensemble Content Extractor model.
- Benchmarked content extractors using 2800 labelled web pages.

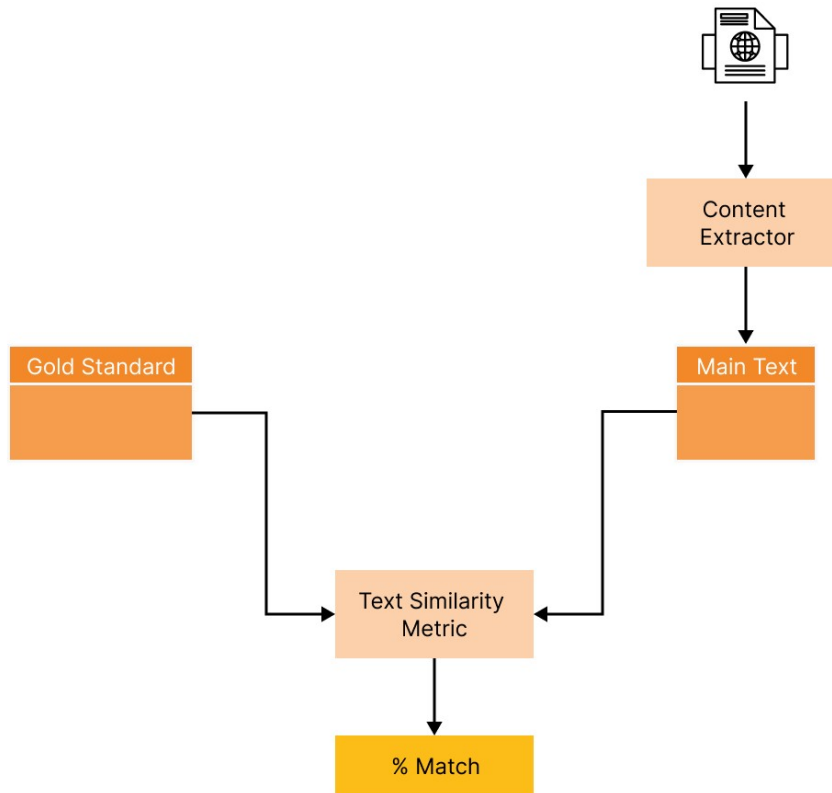
Introduction | Why should we Benchmark Content Extractors?



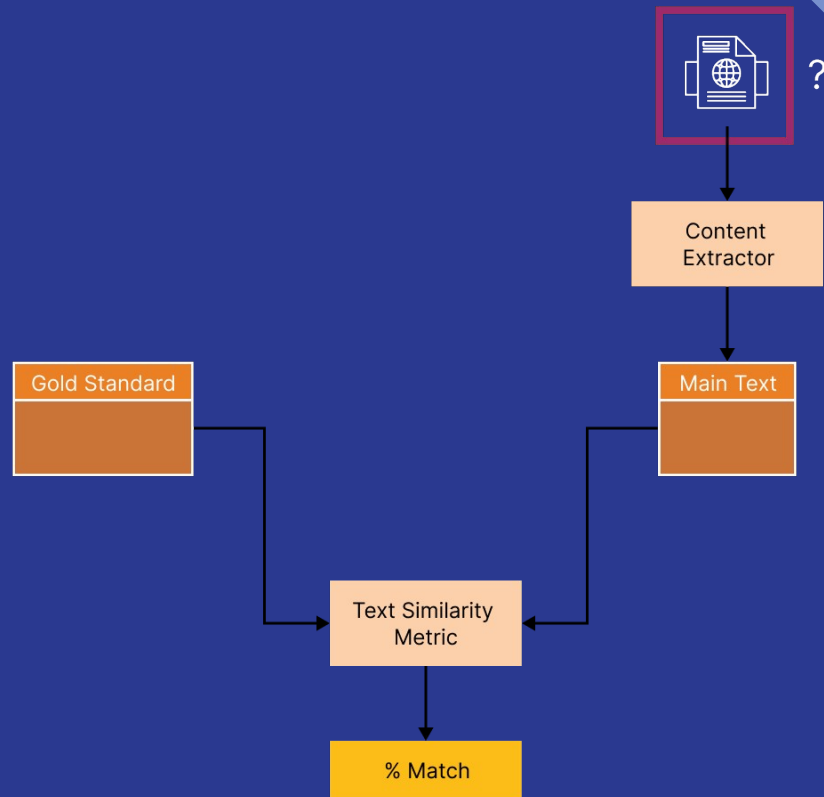
Introduction | Benchmark Content Extractors

To Benchmark Content Extractors

- A set of webpages and its gold standard is required.
 - The gold standard of a webpage is the main text of a webpage that has been checked and corrected to evaluate content extractors.
- Text Similarity Metrics to compare amount of similarity between the gold standard and content extractor output for a webpage.



Data Collection and Preparation



Data Collection and Preparation | Collection of Webpages and Gold Standard

- Researched online for labelled web pages that can be utilized for benchmarking.
- A total of 3114 labelled web pages from five different sources were collected.

Source	Data Size
Readability	115
Article Extraction Benchmark	180
Content Extraction via Text Density Ratio	700
CleanEval	737
Dragnet	1381

Data Collection and Preparation | Quality checks on Gold Standard

- **Text duplication.**
- Concatenation of words without spaces between two tags.
- Incorrect linking of gold standard to its web page.
- Web page lacks content, yet the gold standard contains it.
- **Adding alt attribute text of media tag to gold standard.**

Data Collection and Preparation | Quality checks on Gold Standard

Text Duplication

Web page

Fortune states that the Copyright Royalty Board in Washington D.C. supposed to rule on Thursday about a request by the National Music Publishers' Association to increase royalty rates. An increased royalty rate is wanted for all online music vendors, which includes iTunes.



Gold Standard

Fortune states that the Copyright Royalty Board in Washington D.C. supposed to rule on Thursday about a request by the National Music Publishers' Association to increase royalty rates. An increased royalty rate is wanted for all online music vendors, which includes iTunes.

This increase will raise on average increase prices from 9 cents to 15

Over Possible Royalty Spikes
Posted on September 30th, 2008 in iPod Touch |

Fortune states that the Copyright Royalty Board in Washington D.C. supposed to rule on Thursday about a request by the National Music Publishers' Association to increase royalty rates. An increased royalty rate is wanted for all online music vendors, which includes iTunes.

Data Collection and Preparation | Quantify Gold Standard Errors

Adding alt attribute text of images to the gold standard

Webpage Markup

```
<div class="c_author">
  <div class="normal">
    <a href="http://govhater.newsvine.com/">
      <img class="comment_author_avatar"
        alt="Comment author avatar">
    </a>
    GovHater
  </div>
</div>
...
...
<p>Free market economy is what did it. When the companies
could freely compete, prices came down.
</p>
<p>Bwaaahahahahahaha! That's the funniest thing
I've heard in a long time...</p>
```

Gold Standard

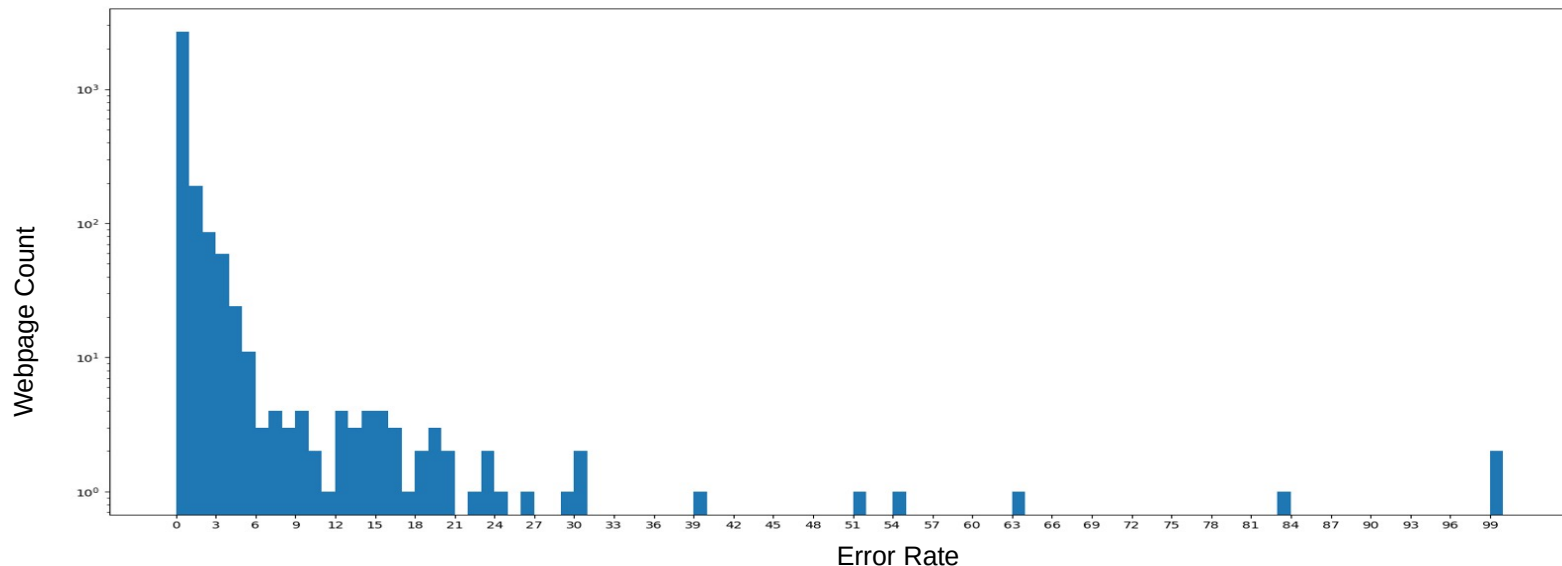
Tue Jul 24, 2012 1:38 PM EDT **Comment author avatar** GovHater willowbrook "Free market economy is what did it. When the companies could freely compete, prices came down. "Bwaaahahahahahaha! That's the funniest thing I've heard in a long time..."

Data Collection and Preparation | Quantify Gold Standard Errors

- The frequency of any alphanumeric token in a gold standard \leq frequency in the webpage.

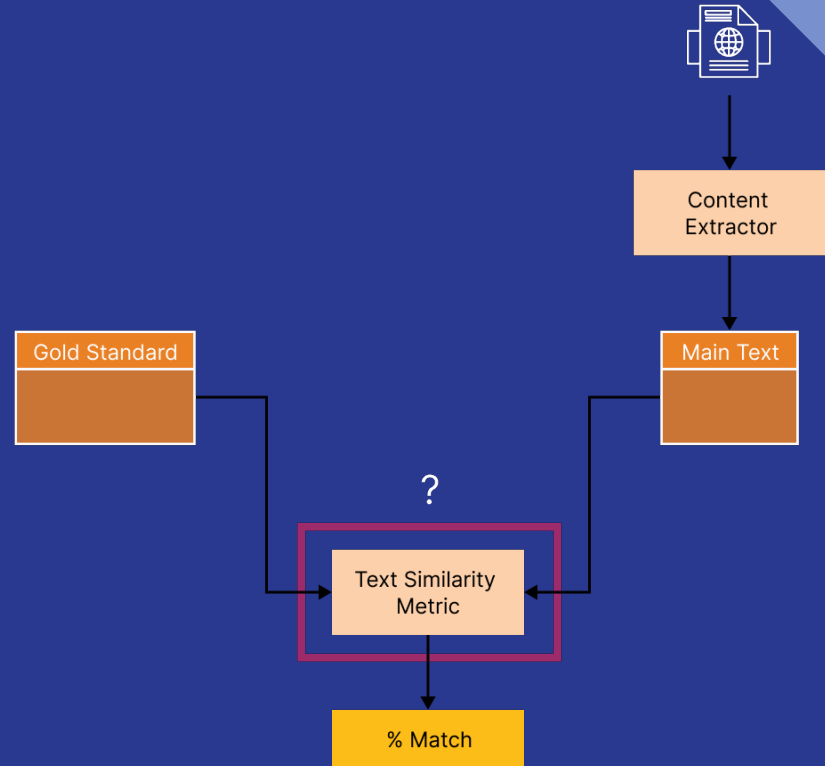
$$Error(webpage) = \frac{\text{number of tokens where } F^{\text{gold standard}} > F^{\text{webpage}}}{\text{number of tokens in the webpage}}$$

Data Collection and Preparation | Quantify Gold Standard Errors



- All the webpages, with error rate $> 2\%$ were discarded to limit the effect of errors on benchmarking to a minimal.
- 3114 webpages were reduced to 2800 webpages.

Text Similarity Metrics



Text Similarity Metrics

- Bag of Words (unigram, 4-grams)
- **RougeLSum**
- Levenshtein Edit Distance
- Jaccard Index

Text Similarity Metrics | RougeLSum Text Similarity

- Takes advantage of Longest Common Subsequence (LCS).
- Measured as Precision, Recall and F1.

Example:

Gold Standard (s1)

The bus is not parked on the highway.

Predicted Text (s2)

A Red bus is on the road.

LCS(s1, s2)

bus is on the

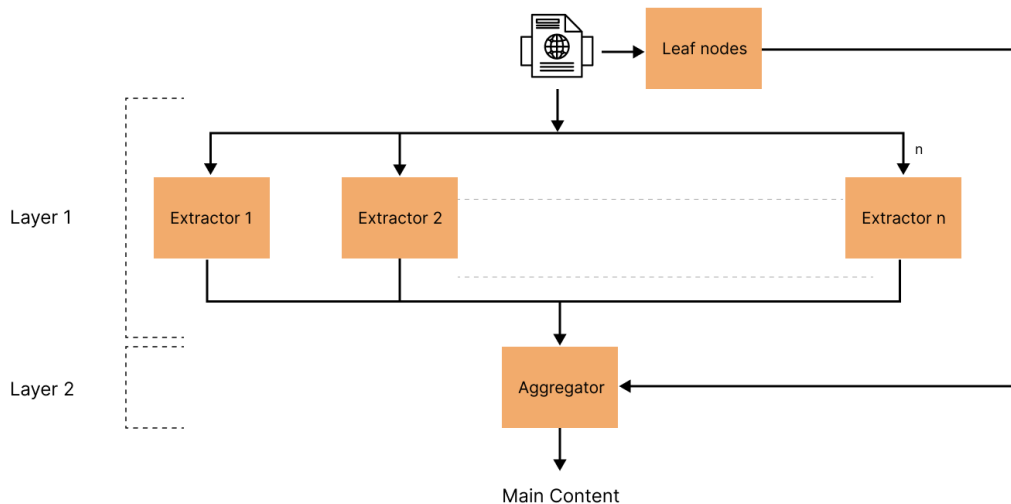
$$Precision : \frac{|LCS(s1, s2)|}{|s2|} = \frac{4}{7}$$

$$Recall : \frac{|LCS(s1, s2)|}{|s1|} = \frac{4}{8}$$

$$F1 = \frac{Precision * Recall}{Precision + Recall} = \frac{4}{15}$$

Ensemble Model Content Extractor

Ensemble Model Content Extractor



Aggregator

Threshold

- Count number of extractors output where leaf node text is subset of it's extraction output.
- If $\text{Count} > \text{Threshold}$, add the text to the final main content.

Machine Learning

- Extract features from layer 1 for the leaf node.
- Feed features into the ml classifier and predict if leaf node should be the subset of the final main content.

Ensemble Model Extractor

Threshold Based Aggregator

- thresh_x_y
 - x: Threshold value in layer 2
 - y: number of extractors in layer 1
- thresh_2_5, thresh_3_4, thresh_4_5, thresh_4_8, thresh_6_8

y	extractors
5	resiliparse, body text extraction, trafilatatura, readability, go domdistiller
8	resiliparse, body text extraction, trafilatatura, readability, go domdistiller, boilerpipe, justext, goose3

Machine Learning based Aggregator

- Binary Classifier: Random Forest Classifier
- ml_y:
 - y: number of extractors in layer 1
- ml_13

ML Classifier	Training Data		Validation Data	
	10 Fold Cross-Validation F1 Score		F1 Score	AUC Score
	Mean	Std deviation		
Logistic Regression	0.859	0.0013	0.820	0.924
Random Forest Classifier	0.872	0.00152	0.8355	0.925

Benchmarks

Benchmarks | Source data

RougeLSum F1 Score Mean

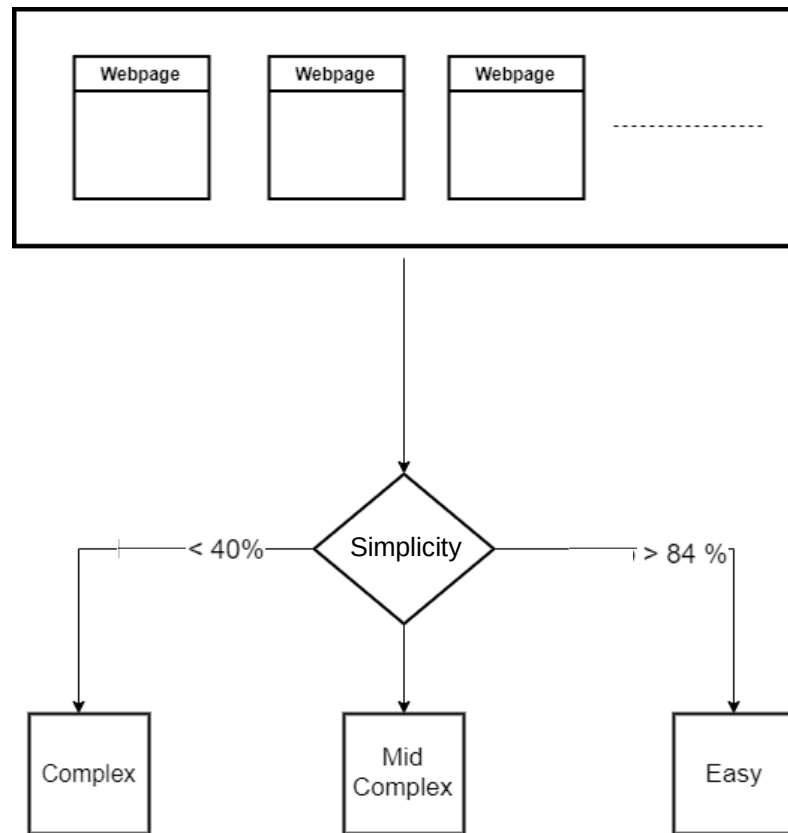
Source Dataset	CETD	RDB	AEB	DG	CE
Extractor					
boilerpipe	0.873	0.828	0.848	0.787	0.819
bs4	0.787	0.770	0.627	0.606	0.893
bte	0.936	0.836	0.820	0.813	0.899
go_domdistiller	0.911	0.929	0.914	0.800	0.877
goose3	0.876	0.762	0.877	0.775	0.796
html_text	0.787	0.770	0.627	0.606	0.893
html2text	0.786	0.773	0.633	0.604	0.885
inscriptis	0.792	0.788	0.642	0.621	0.890
justext	0.868	0.536	0.726	0.789	0.892
lxmlCleaner	0.803	0.844	0.750	0.662	0.871
ml_13	0.805	0.829	0.733	0.671	0.879
readability	0.926	0.869	0.914	0.820	0.881
resiliparse	0.906	0.926	0.938	0.739	0.891
thresh_2_5	0.937	0.907	0.906	0.831	0.910
thresh_3_5	0.921	0.895	0.950	0.826	0.903
thresh_4_5	0.895	0.863	0.950	0.784	0.847
thresh_4_8	0.915	0.896	0.928	0.823	0.904
thresh_6_8	0.886	0.798	0.919	0.788	0.847
trafilatura	0.909	0.935	0.940	0.843	0.841
web2text	0.926	0.966	0.815	0.830	0.932
xpath_text	0.641	0.600	0.339	0.441	0.825

Rank

Source Dataset	CETD	RDB	AEB	DG	CE
Extractor					
boilerpipe	13	13	11	11	20
bs4	18	18	19	19	6
bte	2	11	12	7	5
go_domdistiller	7	3	8	8	14
goose3	12	19	10	13	21
html_text	19	17	20	18	7
html2text	20	16	18	20	11
inscriptis	17	15	17	17	10
justext	14	21	16	9	8
lxmlCleaner	16	10	14	16	15
ml_13	15	12	15	15	13
readability	4	8	7	6	12
resiliparse	9	4	4	14	9
thresh_2_5	1	5	9	2	2
thresh_3_5	5	7	1	4	4
thresh_4_5	10	9	2	12	17
thresh_4_8	6	6	5	5	3
thresh_6_8	11	14	6	10	16
trafilatura	8	2	3	1	18
web2text	3	1	13	3	1
xpath_text	21	20	21	21	19

Benchmarks | Group by Simplicity of Web pages

$$Simplicity(webpage) = \frac{\text{number of tokens in gold standard}}{\text{number of tokens in the webpage}}$$



DataSet Source	25 th percentile	50 th percentile	75 th percentile
Article Extraction Benchmark	33.534	48.838	69.398
Content Extraction via Text Density	53.742	64.190	80.212
CleanEval	79.019	91.708	97.213
Readability	43.116	62.205	83.637
Dragnet	28.704	46.528	67.046
All Sources Combined (Mean)	40.816	62.640	84.078

Benchmarks | Simplicity based Grouping

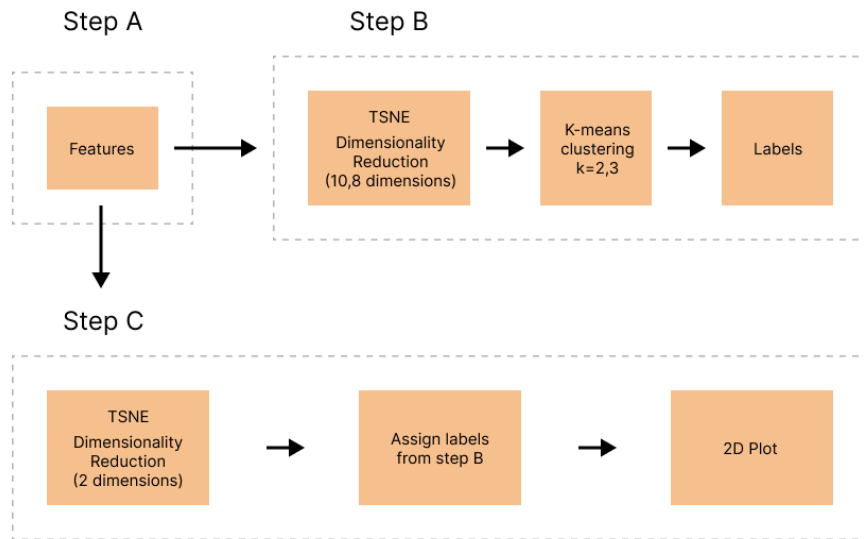
RougeLSum F1 Score Mean

	Complex	Mid-Complex	Easy
Extractor			
boilerpipe	0.769	0.866	0.773
bs4	0.382	0.759	0.961
bte	0.687	0.905	0.941
go_domdistiller	0.797	0.883	0.849
goose3	0.778	0.840	0.774
html_text	0.382	0.759	0.961
html2text	0.386	0.754	0.958
inscriptis	0.402	0.767	0.961
justext	0.659	0.870	0.857
lxmlCleaner	0.473	0.791	0.949
ml_13	0.477	0.797	0.951
readability	0.828	0.883	0.862
resiliparse	0.748	0.846	0.873
thresh_2_5	0.815	0.899	0.904
thresh_3_5	0.841	0.892	0.872
thresh_4_5	0.789	0.866	0.822
thresh_4_8	0.830	0.892	0.866
thresh_6_8	0.781	0.865	0.812
trafilatura	0.827	0.889	0.863
web2text	0.709	0.921	0.957
xpath_text	0.227	0.566	0.910

Rank

	Complex	Mid-Complex	Easy
Extractor			
boilerpipe	10	11	21
bs4	20	18	1
bte	13	2	8
go_domdistiller	6	7	17
goose3	9	14	20
html_text	19	19	2
html2text	18	20	4
inscriptis	17	17	3
justext	14	9	16
lxmlCleaner	16	16	7
ml_13	15	15	6
readability	3	8	15
resiliparse	11	13	11
thresh_2_5	5	3	10
thresh_3_5	1	5	12
thresh_4_5	7	10	18
thresh_4_8	2	4	13
thresh_6_8	8	12	19
trafilatura	4	6	14
web2text	12	1	5
xpath_text	21	21	9

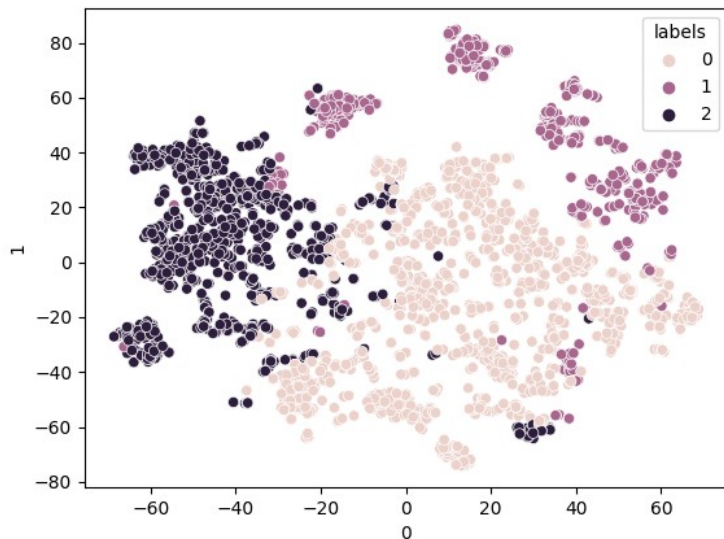
Benchmarks | Group Web pages by K-Means Cluster



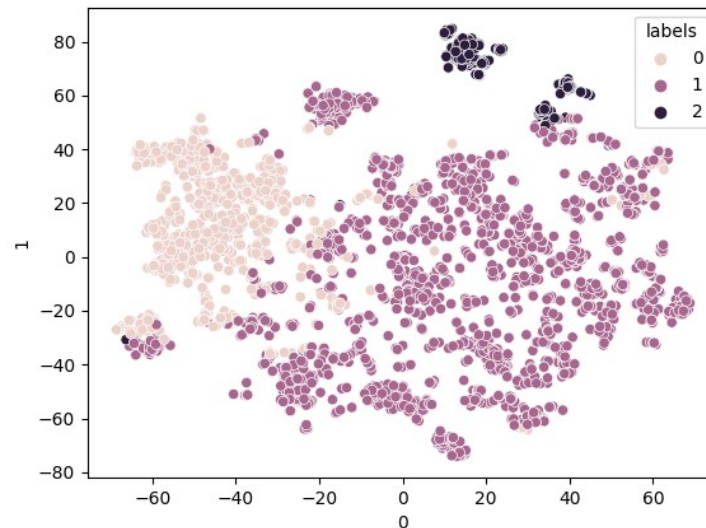
Features: Most Frequently used HTML Elements

tag	Description
h1	Most commonly used to mark up a web page title
h2	Sub-level heading
h3	Sub-level heading
h4	Sub-level heading
h5	Sub-level heading
h6	Sub-level heading
p	Represents a paragraph
ul	Unordered list
a	Defines a hyperlink
div	Division or a Section
br	create a line or break
strong	Define text with bold characters
em	Define emphasized text

Benchmarks | Group Web pages by K-Means Cluster



Dimensionality Reduction=8
K-mean=3



No Dimensionality Reduction
K-mean=3

Benchmarks | K Means cluster Algorithm based Grouping

RougeLSum F1 Score Mean

	Label 0	Label 1	Label 2
Extractor			
boilerpipe	0.833	0.813	0.796
bs4	0.636	0.726	0.853
bte	0.841	0.860	0.893
go_domdistiller	0.850	0.852	0.861
goose3	0.817	0.826	0.778
html_text	0.636	0.726	0.853
html2text	0.634	0.726	0.848
inscriptis	0.650	0.728	0.857
justext	0.803	0.806	0.839
lxmlCleaner	0.688	0.764	0.856
ml_13	0.692	0.766	0.863
readability	0.858	0.874	0.868
resiliparse	0.788	0.870	0.872
thresh_2_5	0.865	0.894	0.895
thresh_3_5	0.868	0.885	0.878
thresh_4_5	0.829	0.847	0.839
thresh_4_8	0.863	0.874	0.881
thresh_6_8	0.832	0.832	0.828
trafilatura	0.865	0.881	0.860
web2text	0.856	0.866	0.924
xpath_text	0.446	0.597	0.767

Rank

	Label 0	Label 1	Label 2
Extractor			
boilerpipe	9	13	19
bs4	19	19	13
bte	8	8	3
go_domdistiller	7	9	9
goose3	12	12	20
html_text	18	20	14
html2text	20	18	15
inscriptis	17	17	11
justext	13	14	17
lxmlCleaner	16	16	12
ml_13	15	15	8
readability	5	5	7
resiliparse	14	6	6
thresh_2_5	3	1	2
thresh_3_5	1	2	5
thresh_4_5	11	10	16
thresh_4_8	4	4	4
thresh_6_8	10	11	18
trafilatura	2	3	10
web2text	6	7	1
xpath_text	21	21	21

Benchmarks | Average performance in each grouping

	Source	Simplicity	K-Means
Extractor			
thresh_3_5	1	2	3
thresh_2_5	2	1	1
trafilatura	3	5	5
web2text	4	4	2
thresh_4_8	5	3	4
go_domdistiller	6	8	8
readability	7	6	6
resiliparse	8	10	9
thresh_4_5	9	9	10
bte	10	7	7
thresh_6_8	11	11	11
boilerpipe	12	12	13
goose3	13	13	14
lxmlCleaner	14	16	16
ml_13	15	15	15
justext	16	14	12
inscriptis	17	17	17
bs4	18	18	18
html_text	19	19	19
html2text	20	20	20
xpath_text	21	21	21

Summary

Summary

- The thesis demonstrated optimal methods of performing content extraction from web pages.
- Choice of an ideal Content Extractor depends on the properties of a web page.
- It is not always beneficial to choose a Main Extractor over a Simple Extractor.
 - Simple Extractor performs better than Main Extractor on the web pages with lesser boilerplate text.
- Performance of Threshold based aggregator was higher than machine learning based aggregator in ensemble model content extractor.



Thank You!