Text Extraction from the Web via Text-to-Tag Ratio

Tim Weninger and William H. Hsu

Department of Computing and Information Sciences
Kansas State University, Manhattan KS

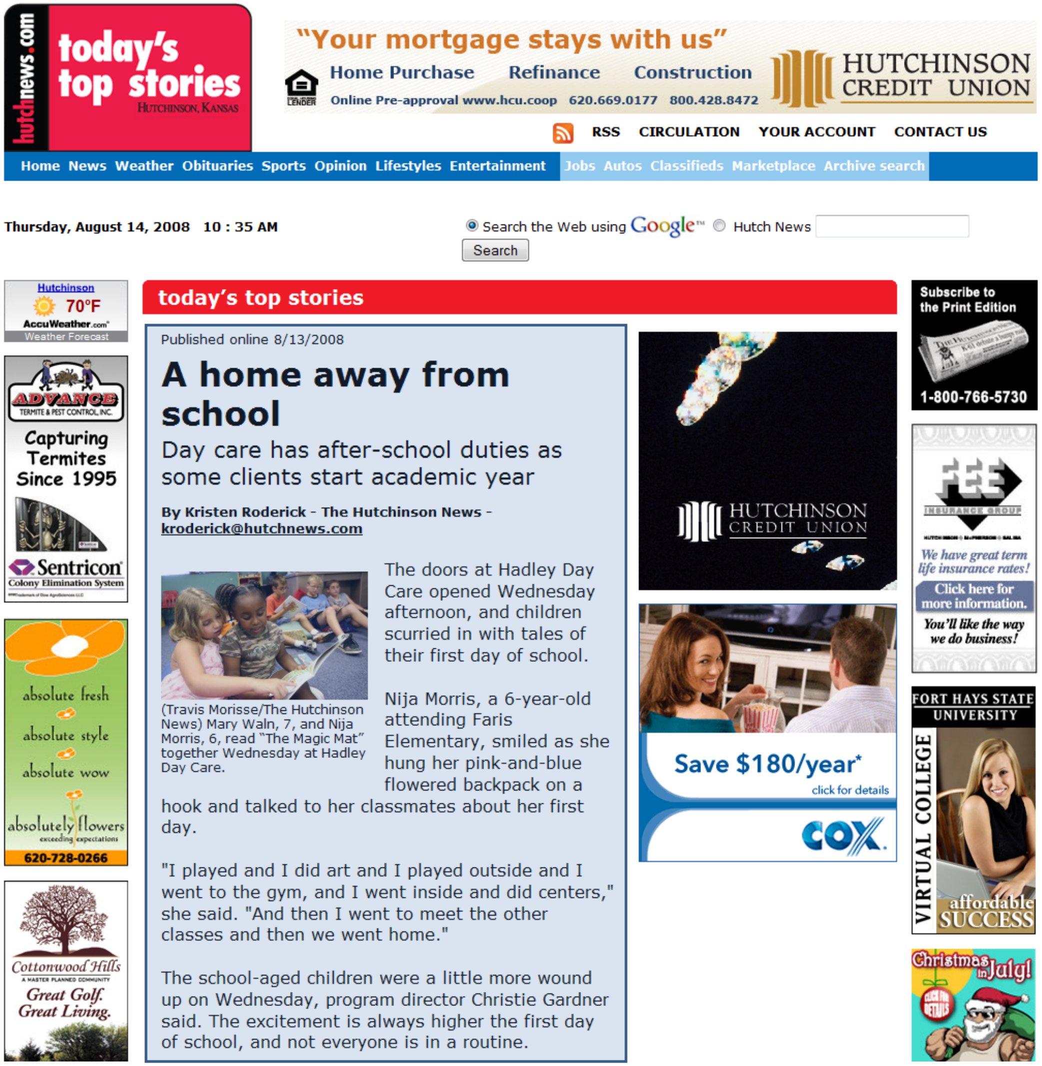
DEXA 2008 Workshop on Text-based Information Retrieval

Turin, Italy

Outline

- Introduction
  - Motivation
  - Related Work
- The Text-to-Tag Ratio
  - Heuristic
  - Worst Case
- Methodology
  - Pre-processing
  - Computing clusters
- Results
  - Evaluation Metrics
  - Results
- Conclusions and Future Work

Introduction – Motivation [1]


hutchnews.com
today's top stories
Hutchinson, Kansas
"Your mortgage stays with us"
Home Purchase Refinance Construction
Online Pre-approval www.hcu.coop 620.669.0177 800.428.8472
HUTCHINSON CREDIT UNION
RSS CIRCULATION YOUR ACCOUNT CONTACT US
Home News Weather Obituaries Sports Opinion Lifestyles Entertainment Jobs Autos Classifieds Marketplace Archive search
Search
Hutchinson 70°F AccuWeather.com Weather Forecast
ADVANCE Termite & Pest Control, Inc.
Capturing Termites Since 1995
Sentricon Colony Elimination System
absolute fresh
absolute style
absolute wow
absolutely flowers
exceeding expectations
620-728-0266
Cottonwood Hills
A Master Planned Community
Great Golf. Great Living.
HUTCHINSON CREDIT UNION
Save $180/year*
click for details
COX
Subscribe to the Print Edition
1-800-766-5730
FEE Insurance Group
We have great term life insurance rates!
Click here for more information.
You'll like the way we do business!
FORT HAYS STATE UNIVERSITY
VIRTUAL COLLEGE
affordable SUCCESS
Christmas in July!
Click for details

Thursday, August 14, 2008 10 : 35 AM

● Search the Web using Google™ ○ Hutch News

today's top stories

Published online 8/13/2008

A home away from school

Day care has after-school duties as some clients start academic year

By Kristen Roderick - The Hutchinson News - kroderick@hutchnews.com

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

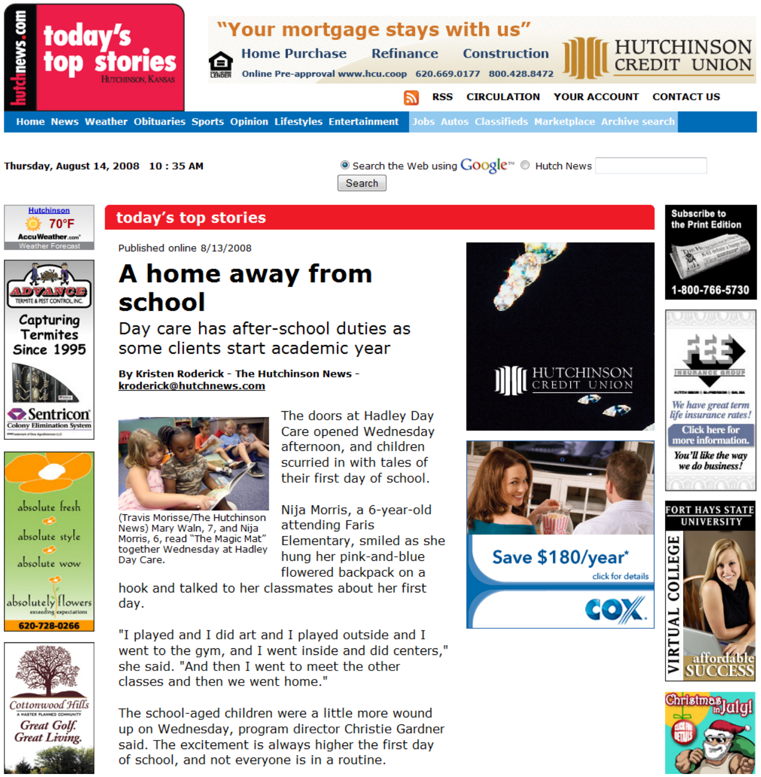
"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.

Taken from The Hutchinson News on 8/14/2008

- Problem:
  - Too much *junk* in a web page
- Goal:
  - Extract only the content of a page

Introduction – Motivation [2] – Example



Thursday, August 14, 2008 10 : 35 AM

today's top stories

Published online 8/13/2008

A home away from school

Day care has after-school duties as some clients start academic year

By Kristen Roderick - The Hutchinson News - kroderick@hutchnews.com

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.

Rendered HTML Document

Published online 8/13/2008

A home away from school
Day care has after-school duties as some clients start academic year
By Kristen Roderick - The Hutchinson News - kroderick@hutchnews.com

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

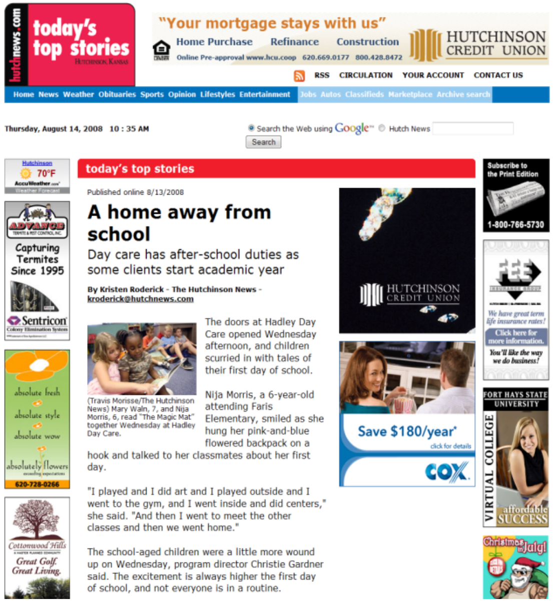
"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.

Text content of the document

Related Work [1]

- Naïve Approach
  - Remove all HTML tags


hutchnews.com
today's top stories
Hutchinson, Kansas
"Your mortgage stays with us"
Home Purchase Refinance Construction
Online Pre-approval www.hcu.coop 620.669.0177 800.428.8472
HUTCHINSON CREDIT UNION
RSS CIRCULATION YOUR ACCOUNT CONTACT US
Home News Weather Obituaries Sports Opinion Lifestyles Entertainment Jobs Autos Classifieds Marketplace Archive search
Search the Web using Google™ Hutch News
Search
Hutchinson 70°F AccuWeather.com Weather Forecast
ADVANCE Termite & Pest Control, Inc.
Capturing Termites Since 1995
Sentricon Colony Elimination System
absolute fresh absolute style absolute wow absolutely flowers
620-728-0266
Cottonwood Hills
Great Golf. Great Living.
HUTCHINSON CREDIT UNION
Save $180/year*
click for details
COX
Subscribe to the Print Edition
1-800-766-5730
FEE
We have great term life insurance rates!
Click here for more information.
You'll like the way we do business!
FORT HAYS STATE UNIVERSITY
VIRTUAL COLLEGE
affordable SUCCESS
Christmas in July!

Thursday, August 14, 2008 10 : 35 AM

today's top stories

Published online 8/13/2008

A home away from school

Day care has after-school duties as some clients start academic year

By Kristen Roderick - The Hutchinson News - kroderick@hutchnews.com

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.

Original, Rendered HTML Document

RSS
CIRCULATION
YOUR ACCOUNT
CONTACT US
Home
News
Top Stories
Local/Regional
News
Briefs
Education
Ask Hutch
Business
Public Record
Special sections
Videos
Photos
Slideshows
Forums
Weather
Obituaries
Sports
Latest Sports
NJCAA
Tournament
Opinion
Editorials
Letters to the
Editor
Columnists
Lifestyles
Food
Health

Religion
Outdoors
Life's Little
Moments
Weddings
Engagements
Anniversaries
Coming Events
Et cetera
Entertainment
Preview
TV listings
Jobs
Autos
Classifieds
Marketplace
Archive search
Thursday, August 14,
2008 10 : 35 AM
Search the Web using Hutch
News
Weather Forecast
today's top stories

Published online 8/13/2008

A home away from school
Day care has after-school
duties as some clients start
academic year
By Kristen Roderick - The
Hutchinson News -

kroderick@hutchnews.com

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.
The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

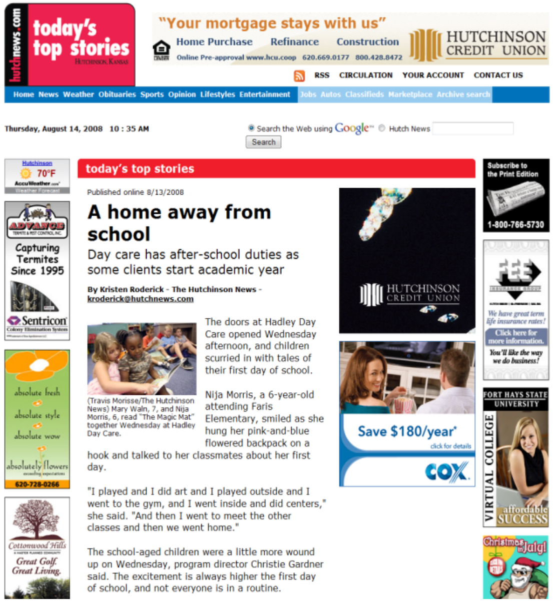
Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

All Text of the Document

Related Work [2]

- Tag Approach
  - Use HTML tags as clues for content
  - Problem: Style-sheets


hutchnews.com
today's top stories
Hutchinson, Kansas
"Your mortgage stays with us"
Home Purchase
Refinance
Construction
Online Pre-approval www.hcu.coop 620.669.0177 800.428.8472
HUTCHINSON CREDIT UNION
RSS
CIRCULATION
YOUR ACCOUNT
CONTACT US
Home News Weather Obituaries Sports Opinion Lifestyles Entertainment Jobs Autos Classifieds Marketplace Archive search
Search the Web using Google
Hutch News
Search
Hutchinson
70°F
AccuWeather.com
Weather Forecast
ADVANCE Termite & Pest Control, Inc.
Capturing Termites Since 1995
Sentricon Colony Elimination System
absolute fresh
absolute style
absolute wow
absolutely flowers
620-728-0266
Cottonwood Hills
Great Golf. Great Living.
Subscribe to the Print Edition
1-800-766-5730
We have great term life insurance rates!
Click here for more information.
You'll like the way we do business!
Save $180/year*
click for details
COX
FORT HAYS STATE UNIVERSITY
VIRTUAL COLLEGE
affordable SUCCESS
Christmas in July!

Thursday, August 14, 2008 10 : 35 AM

today's top stories

Published online 8/13/2008

A home away from school

Day care has after-school duties as some clients start academic year

By Kristen Roderick - The Hutchinson News - kroderick@hutchnews.com

(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

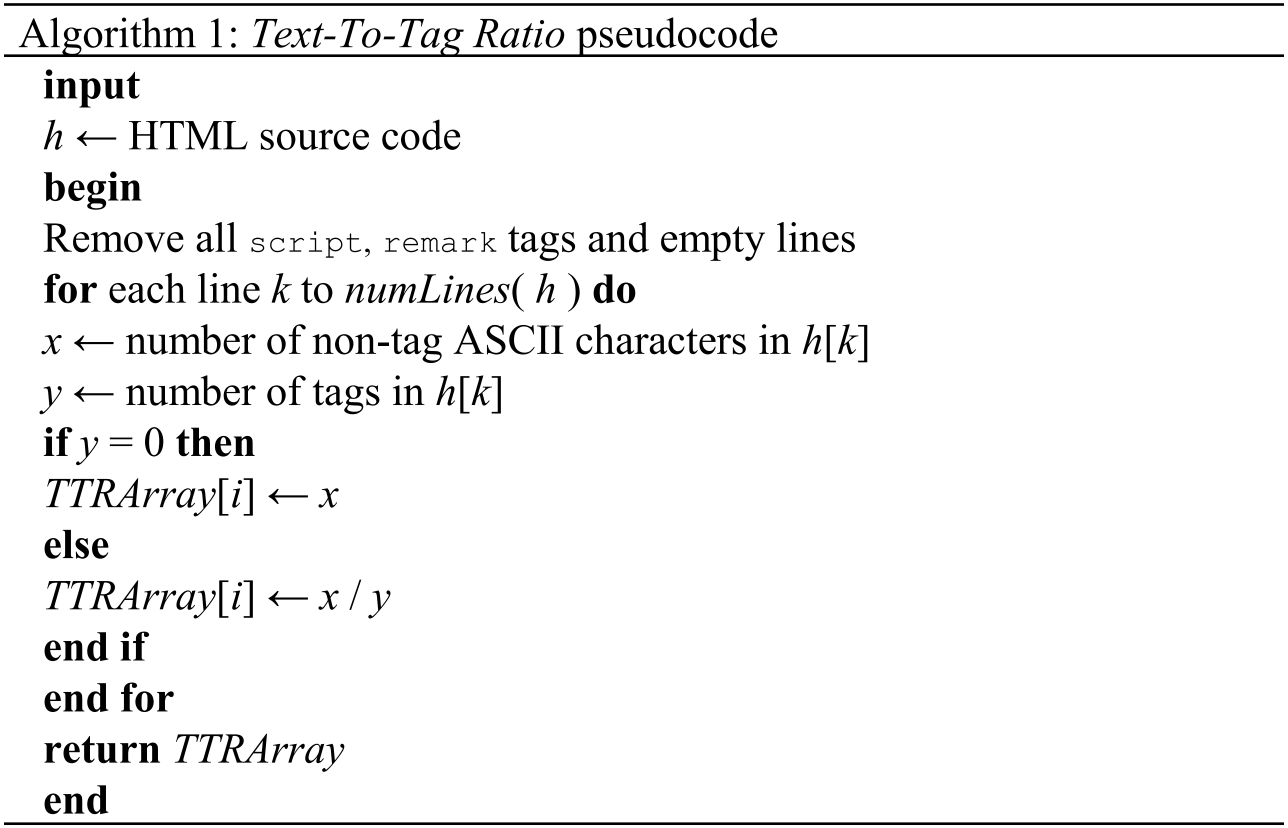
"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.

Original, Rendered HTML Document

```
<div>
        <div>
        </div>
        <div>
                <div>
                        Eat at Joes
                </div>
        </div>
        <div>
                <div>
                        <div>
The doors at Hadley Day Care opened Wednesday
afternoon, and children scurried in with tales of their
first day of school.
                        </div>
                        <div>
Nija Morris, a 6-year-old attending Faris Elementary,
smiled as she hung her pink-and-blue flowered
backpack on a hook and talked to her classmates about
her first day.
                        </div>
                </div>
        </div>
</div>
```

Text-to-Tag Ratio [1]

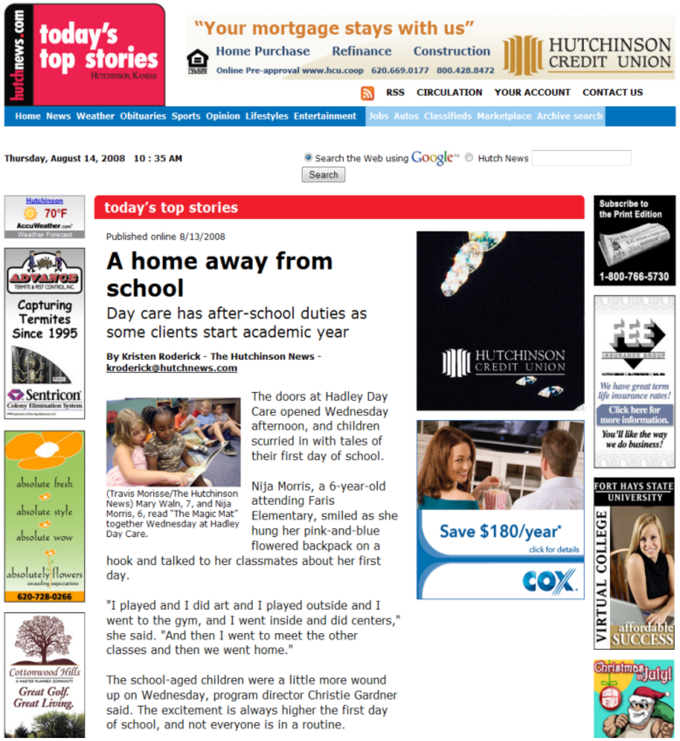


Algorithm 1: *Text-To-Tag Ratio* pseudocode

```
input
$h \leftarrow$ HTML source code
begin
Remove all script, remark tags and empty lines
for each line $k$ to $numLines(h)$ do
$x \leftarrow$ number of non-tag ASCII characters in $h[k]$
$y \leftarrow$ number of tags in $h[k]$
if $y = 0$ then
$TTRArray[i] \leftarrow x$
else
$TTRArray[i] \leftarrow x / y$
end if
end for
return $TTRArray$
end
```

Text-to-Tag Ratio [2]

- Example


hutchnews.com
today's top stories
Hutchinson, Kansas
"Your mortgage stays with us"
Home Purchase Refinance Construction
Online Pre-approval www.hcu.coop 620.669.0177 800.428.8472
HUTCHINSON CREDIT UNION
RSS CIRCULATION YOUR ACCOUNT CONTACT US
Home News Weather Obituaries Sports Opinion Lifestyles Entertainment Jobs Autos Classifieds Marketplace Archive search
○ Search the Web using Google™ ○ Hutch News
Search
Hutchinson 70°F AccuWeather.com Weather Forecast
ADVANCE Termite & Pest Control, Inc. Capturing Termites Since 1995 Sentricon Colony Elimination System
absolute fresh absolute style absolute wow absolutely flowers exceeding expectations 620-728-0266
Cottonwood Hills A Master Planned Community Great Golf. Great Living.
Subscribe to the Print Edition 1-800-766-5730
FEE Insurance Group We have great term life insurance rates! Click here for more information. You'll like the way we do business!
Fort Hays State University Virtual College affordable success
Christmas in July!
HUTCHINSON CREDIT UNION
Save $180/year* click for details COX

Thursday, August 14, 2008 10 : 35 AM

today's top stories

Published online 8/13/2008

A home away from school

Day care has after-school duties as some clients start academic year

By Kristen Roderick - The Hutchinson News - kroderick@hutchnews.com

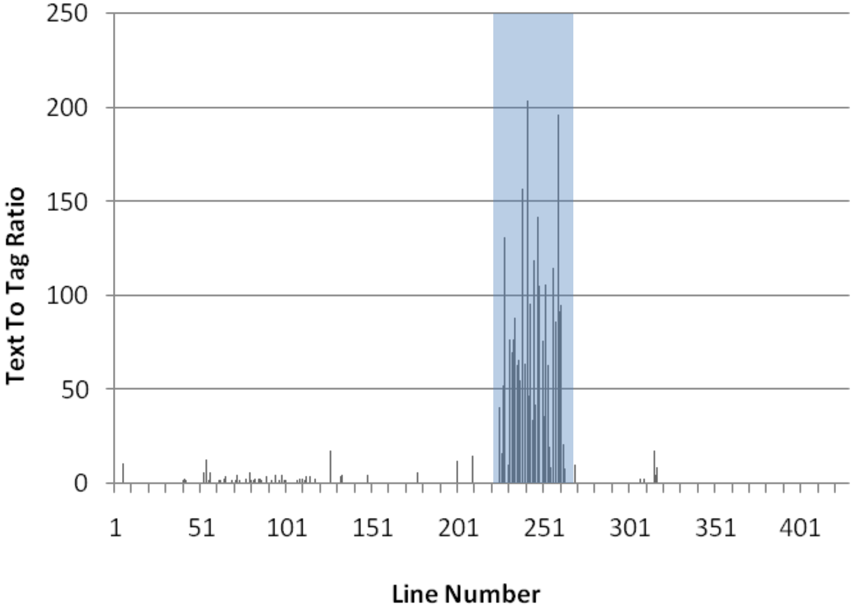
(Travis Morisse/The Hutchinson News) Mary Waln, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

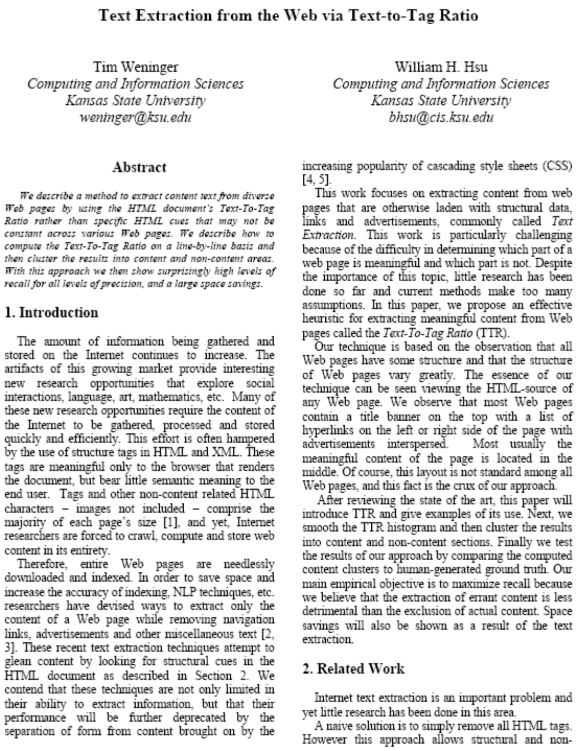
"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.



Text-to-Tag Ratio [3]

- Worst Case [1]
  - Non-HTML or all content pages



Text Extraction from the Web via Text-to-Tag Ratio

Tim Weninger
Computing and Information Sciences
Kansas State University
weninger@ksu.edu

William H. Hsu
Computing and Information Sciences
Kansas State University
bhsu@cis.ksu.edu

Abstract

We describe a method to extract content text from diverse Web pages by using the HTML document's Text-To-Tag Ratio rather than specific HTML cues that may not be constant across various Web pages. We describe how to compute the Text-To-Tag Ratio on a line-by-line basis and then cluster the results into content and non-content areas. With this approach we then show surprisingly high levels of recall for all levels of precision, and a large space savings.

1. Introduction

The amount of information being gathered and stored on the Internet continues to increase. The artifacts of this growing market provide interesting new research opportunities that explore social interactions, language, art, mathematics, etc. Many of these new research opportunities require the content of the Internet to be gathered, processed and stored quickly and efficiently. This effort is often hampered by the use of structure tags in HTML and XML. These tags are meaningful only to the browser that renders the document, but bear little semantic meaning to the end user. Tags and other non-content related HTML characters – images not included – comprise the majority of each page's size [1], and yet, Internet researchers are forced to crawl, compute and store web content in its entirety.

Therefore, entire Web pages are needlessly downloaded and indexed. In order to save space and increase the accuracy of indexing, NLP techniques, etc. researchers have devised ways to extract only the content of a Web page while removing navigation links, advertisements and other miscellaneous text [2, 3]. These recent text extraction techniques attempt to glean content by looking for structural cues in the HTML document as described in Section 2. We contend that these techniques are not only limited in their ability to extract information, but that their performance will be further deprecated by the separation of form from content brought on by the increasing popularity of cascading style sheets (CSS) [4, 5].

This work focuses on extracting content from web pages that are otherwise laden with structural data, links and advertisements, commonly called *Text Extraction*. This work is particularly challenging because of the difficulty in determining which part of a web page is meaningful and which part is not. Despite the importance of this topic, little research has been done so far and current methods make too many assumptions. In this paper, we propose an effective heuristic for extracting meaningful content from Web pages called the *Text-To-Tag Ratio* (TTR).

Our technique is based on the observation that all Web pages have some structure and that the structure of Web pages vary greatly. The essence of our technique can be seen viewing the HTML-source of any Web page. We observe that most Web pages contain a title banner on the top with a list of hyperlinks on the left or right side of the page with advertisements interspersed. Most usually the meaningful content of the page is located in the middle. Of course, this layout is not standard among all Web pages, and this fact is the crux of our approach.

After reviewing the state of the art, this paper will introduce TTR and give examples of its use. Next, we smooth the TTR histogram and then cluster the results into content and non-content sections. Finally we test the results of our approach by comparing the computed content clusters to human-generated ground truth. Our main empirical objective is to maximize recall because we believe that the extraction of errant content is less detrimental than the exclusion of actual content. Space savings will also be shown as a result of the text extraction.

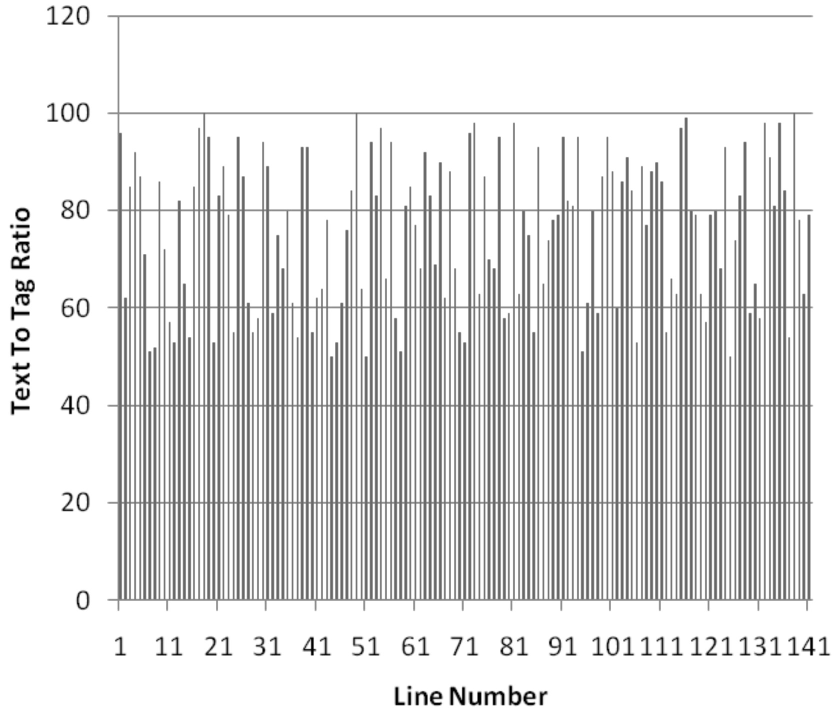
2. Related Work

Internet text extraction is an important problem and yet little research has been done in this area.

A naive solution is to simply remove all HTML tags. However this approach allows structural and non-

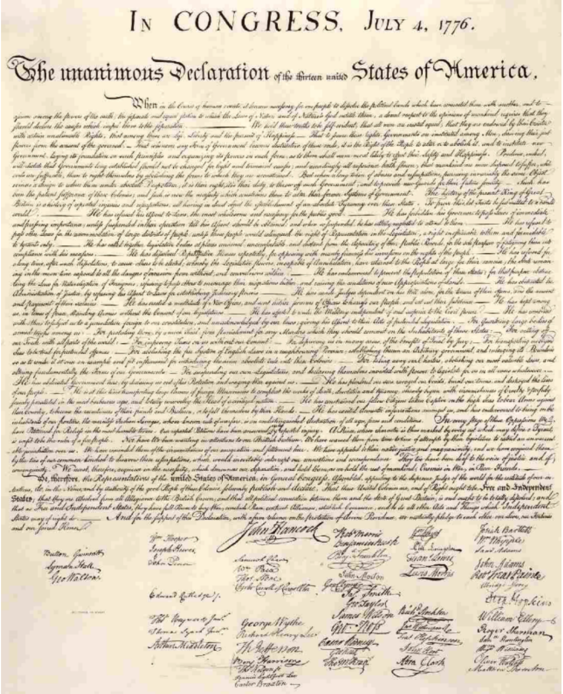
TIR'08 Paper

approximation



Text-to-Tag Ratio [4]

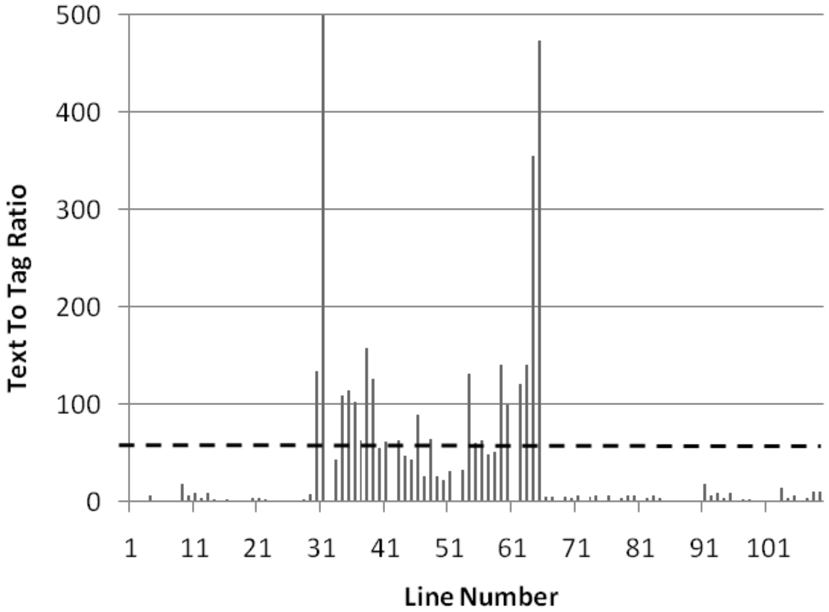
- Worst Cases [2]
  - › American Declaration of Independence Web page



IN CONGRESS, JULY 4, 1776.

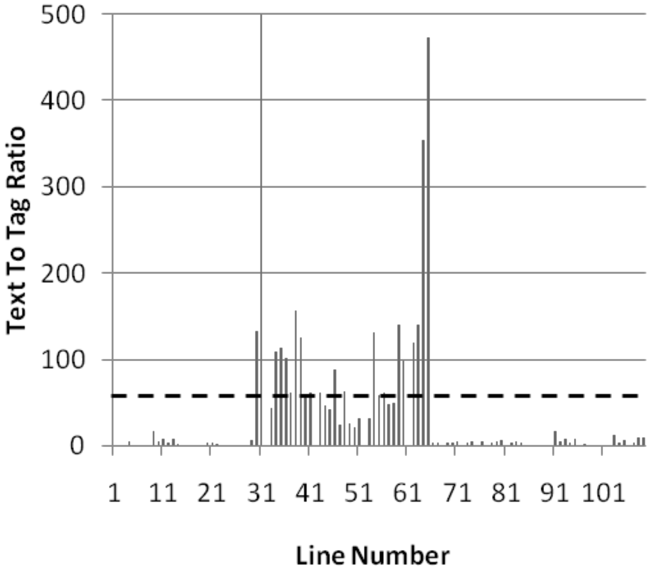
The unanimous Declaration of the thirteen united States of America,

American Declaration of Independence
TTR computed from digital copy at
http://www.ushistory.org/declaration/document/index.htm

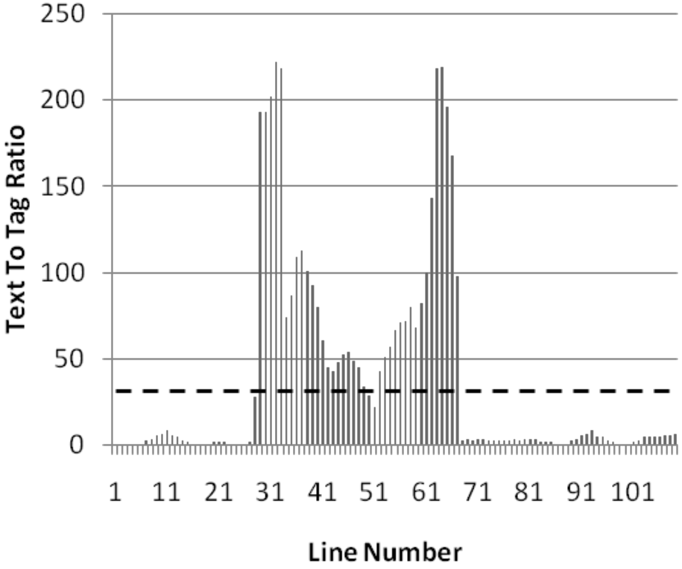


Methodology [1]

- Preprocessing
  - Content Blurring

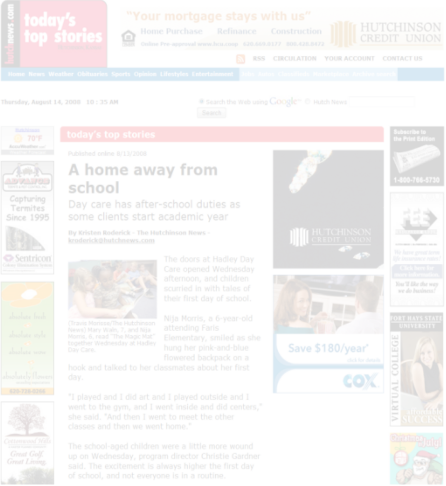


$$e_k = \frac{\sum_{i=k-r}^{k+r} TTRArray_i}{2r+1}$$



Methodology [2]

- Clustering [1]
  - K-Means, Farthest First, Expectation


hutchnews.com
today's top stories
"Your mortgage stays with us"
Home Purchase Refinance Construction
Online Pre-approval www.hcu.coop 620.669.0177 800.428.0472
HUTCHINSON CREDIT UNION
RSS CIRCULATION YOUR ACCOUNT CONTACT US
Home News Weather Obituaries Sports Opinion Lifestyles Entertainment
Search the Web using Google Hutch News
Search
70°F
Capturing Termites Since 1995
Sentricon
absolute fresh absolute style absolute wow absolutely flowers
620-728-0266
Cottonwood Hills
Great Golf. Great Living.
Subscribe to the Print Edition
1-800-766-5730
We have great term life insurance rates!
You'll like the way we do business!
Save $180/year
cox
FORT HAYS STATE UNIVERSITY
VIRTUAL COLLEGE
affordable SUCCESS

Thursday, August 14, 2008 10 : 35 AM

today's top stories

Published online 8/13/2008

A home away from school

Day care has after-school duties as some clients start academic year

By Kristen Roderick - The Hutchinson News - kroderick@hutchnews.com

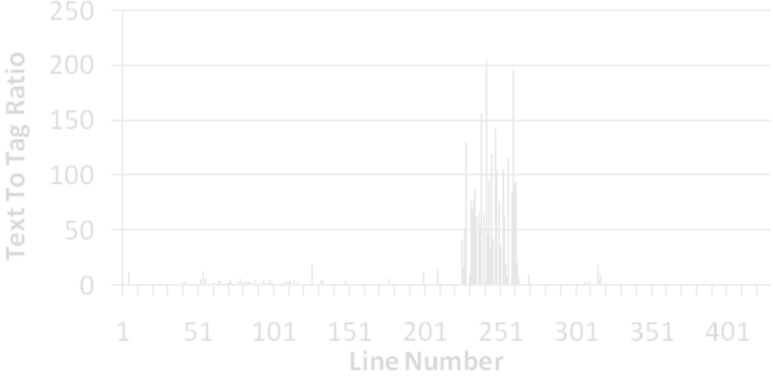
(Travis Morisse/The Hutchinson News) Mary Wain, 7, and Nija Morris, 6, read "The Magic Mat" together Wednesday at Hadley Day Care.

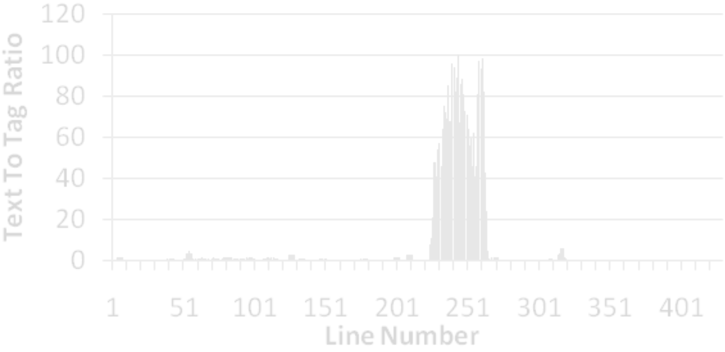
The doors at Hadley Day Care opened Wednesday afternoon, and children scurried in with tales of their first day of school.

Nija Morris, a 6-year-old attending Faris Elementary, smiled as she hung her pink-and-blue flowered backpack on a hook and talked to her classmates about her first day.

"I played and I did art and I played outside and I went to the gym, and I went inside and did centers," she said. "And then I went to meet the other classes and then we went home."

The school-aged children were a little more wound up on Wednesday, program director Christie Gardner said. The excitement is always higher the first day of school, and not everyone is in a routine.



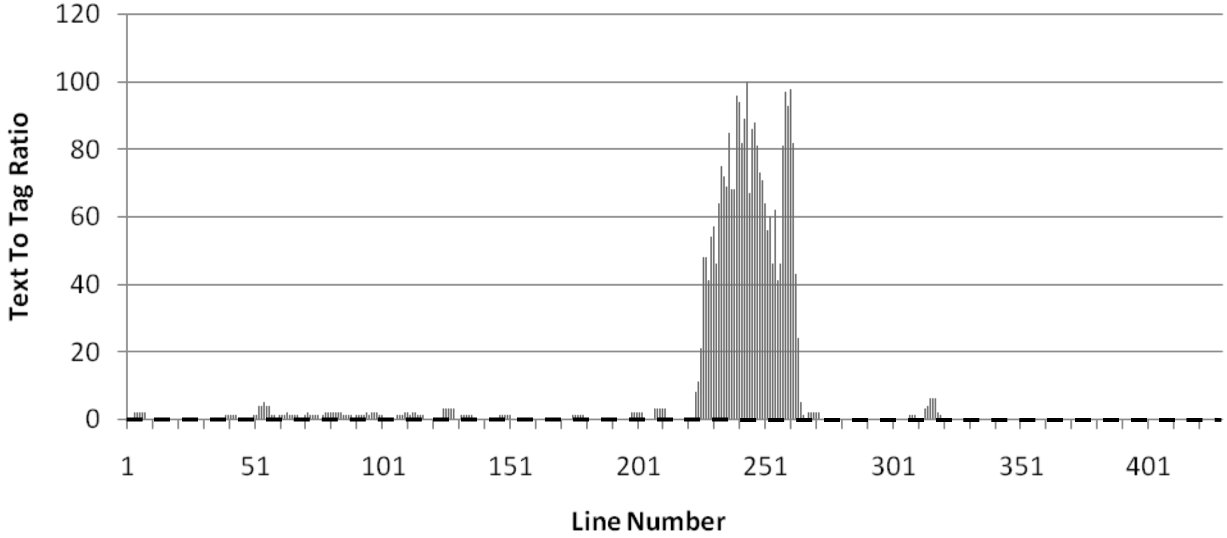


Cluster	1 cluster	2 clusters	3 clusters
1	6.85	0.56	10.12
2	-	**53.40**	**70.42**
3	-	-	0.59

K-Means clustering

Methodology [3]

- Clustering [2]
  - › Threshold clustering based on standard deviation



Std. Dev. Is 20.3TTR for
Hutchinson News document

Methodology [4]

- Clustering [3]
  - › Prediction clustering
    - Looks for jumps in the moving average of the TTRArray
    - Not formalized in this paper
    - Very good extension in ANNIE'08 paper.

Methodology [5]

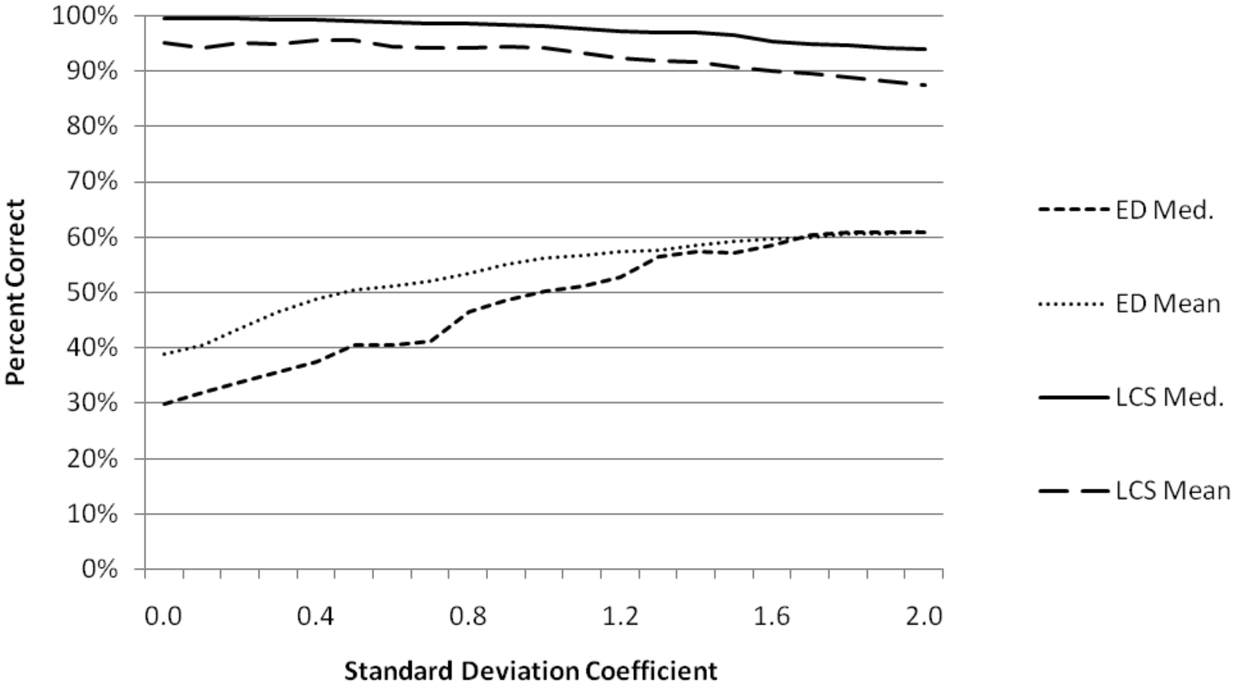
- Evaluation Metrics
  - Longest Common Subsequence (LCS)
    - Very Draconian
    - Treated as recall
  - Edit Distance Ratio (EDR)
    - Inverse Levenstein distance over longest sequence
    - Treated as precision

$$EDR = \frac{1 - \mathrm{edtDist}(o, m)}{\max\left(\mathrm{len}(o), \mathrm{len}(m)\right)}$$

- Evaluation method
  - 176 Pages selected by querying Yahoo search for “the”
  - Gold standard for each page created by a CS undergraduate.
  - Metrics computed against gold standard and averaged

Results [1]

- Threshold Only



Results [2]

- Longest Common Subsequence

	Threshold	EM	K-Means	Farthest First	Prediction
Mean (%)	**94.19**	92.62	92.47	85.88	81.14
Median (%)	98.65	**99.34**	98.68	94.18	94.42
Std Dev.	**14.03**	17.60	16.57	21.32	24.85
Matches	34	**43**	35	25	22

- Edit Distance Ratio

	Threshold	EM	K-Means	Farthest First	Prediction
Mean (%)	56.21	48.77	57.44	**62.53**	52.40
Median (%)	61.63	48.98	61.17%	**77.03**	55.30
Std Dev.	31.89	30.66	32.96	33.75	**30.01**

Results [3]

- Space savings
  - Mean file sizes

	HTML	Extracted Text	GZip HTML	GZip Text
File Size (Kb)	9,630.34	497.70	2,234.77	275.53

Conclusions and Future Work

- Text-To-Tag Ratio Approach
  - A valid content extraction technique
  - But has Limitations
- Need for better evaluation metrics
- Prediction clustering
  - Extended for ANNIE'08 in St. Louis, MO, USA
  - General histogram clustering
    - Uses Gaussian Blurring
    - Analysis of the slope of the tangent line
    - Extracting dimensions and re-clustering
  - Much better results exist, but were not available by the TIR deadline.

Questions?