Christian Kohlschütter, *Peter Fankhauser, Wolfgang Nejdl*

Boilerplate Detection using Shallow Text Features

KNOWLEDGE • INFORMATION • LEARNING

Boilerplate Text

Login | Contact | Imprint

Leibniz Universität Hannover

Web Science – **Investigating the Future** of Information and Communication

Home / Profile | People | Research Areas | Jobs | News / Events | Publications

Language:

Deutsch English

- About L3S
- Contact
- Organigram
- Vision 2009-2013
- Mentoring Guidelines
- Facts and Figures

L3S Research Center

The L3S Research Center focuses on

The Advisory Board visiting L3S Research Center

fundamental and application-oriented research in all areas of *Web Science*. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center's total research volume is more than 6 million euros per year, with a large number of projects in the areas of

- **Intelligent Access to Information**
- **Next Generation Internet**
- **E-Science**

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by complementary background knowledge that L3S researchers themselves bring to their work, and the center's cooperations and projects with scholars and researchers not only from computer sciences, but also including library sciences, linguistics, psychology, law, economics, and business administration.

The experience L3S has gained over the years in participating in a variety of projects financed by the European Union has led to a large number of cooperations with research institutions and companies throughout all of Europe, and in many research results and products. Since 2008 alone, the L3S has been involved in 12 EU projects as part of the EU's Seventh Framework Programme, four of them (LivingKnowledge, Okkam, EUWB and EERQI) integrated projects, as well as the STELLAR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled "Future Internet – Internet, Information and I," L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.

News:

- Best Paper Nomination at WSDM 2010
- PHAROS is presented at ConventionCamp '09
- December 2009: L3S at International PhD. workshop in Beijing
- First Workshop on "Information, Internet, and I"
- Best Paper Prize for PhD proposal
- Making Web Diversity a true asset - Workshop Announcement
- ZDF: Leben in einer vernetzten Welt
- Why do we need a Content-Centric Future Internet?

Further News

©2010 L3S Research Center • Appelstrasse 9a • 30167 Hannover • Phone +49. 511. 762-17713 • Email: info@L3S.de

Boilerplate Text

Login | Contact | Imprint

Leibniz Universität Hannover

Web Science – **Investigating the Future** of Information and Communication

Home / Profile | People | Research Areas | Jobs | News / Events | Publications

Language:

Deutsch English

- About L3S
- Contact
- Organigram
- Vision 2009-2013
- Mentoring Guidelines
- Facts and Figures

L3S Research Center

The L3S Research Center focuses on

The Advisory Board visiting L3S Research Center

fundamental and application-oriented research in all areas of *Web Science*. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center's total research volume is more than 6 million euros per year, with a large number of projects in the areas of

- **Intelligent Access to Information**
- **Next Generation Internet**
- **E-Science**

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by complementary background knowledge that L3S researchers themselves bring to their work, and the center's cooperations and projects with scholars and researchers not only from computer sciences, but also including library sciences, linguistics, psychology, law, economics, and business administration.

The experience L3S has gained over the years in participating in a variety of projects financed by the European Union has led to a large number of cooperations with research institutions and companies throughout all of Europe, and in many research results and products. Since 2008 alone, the L3S has been involved in 12 EU projects as part of the EU's Seventh Framework Programme, four of them (LivingKnowledge, Okkam, EUWB and EERQI) integrated projects, as well as the STELLAR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled "Future Internet – Internet, Information and I," L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.

News:

- Best Paper Nomination at WSDM 2010
- PHAROS is presented at ConventionCamp '09
- December 2009: L3S at International PhD. workshop in Beijing
- First Workshop on "Information, Internet, and I"
- Best Paper Prize for PhD proposal
- Making Web Diversity a true asset - Workshop Announcement
- ZDF: Leben in einer vernetzten Welt
- Why do we need a Content-Centric Future Internet?

Further News

©2010 L3S Research Center • Appelstrasse 9a • 30167 Hannover • Phone +49. 511. 762-17713 • Email: info@L3S.de

Boilerplate Removal

L3S Research Center

The L3S Research Center focuses on

The Advisory Board visiting L3S Research Center

fundamental and application-oriented research in all areas of *Web Science*. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center's total research volume is more than 6 million euros per year, with a large number of projects in the areas of

- **Intelligent Access to Information**
- **Next Generation Internet**
- **E-Science**

The L3S is a research-driven institution that attracts outstanding students and researchers from all over the world with its open and invigorating research culture. For young researchers, the L3S is encouraging, innovative, international, independent, and supportive.

L3S activities primarily focus on research, but also include consulting and technology transfer. This is made possible by complementary background knowledge that L3S researchers themselves bring to their work, and the center's cooperations and projects with scholars and researchers not only from computer sciences, but also including library sciences, linguistics, psychology, law, economics, and business administration.

The experience L3S has gained over the years in participating in a variety of projects financed by the European Union has led to a large number of cooperations with research institutions and companies throughout all of Europe, and in many research results and products. Since 2008 alone, the L3S has been involved in 12 EU projects as part of the EU's Seventh Framework Programme, four of them (LivingKnowledge, Okkam, EUWB and EERQI) integrated projects, as well as the STELLAR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled "Future Internet – Internet, Information and I," L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.

Boilerplate Removal

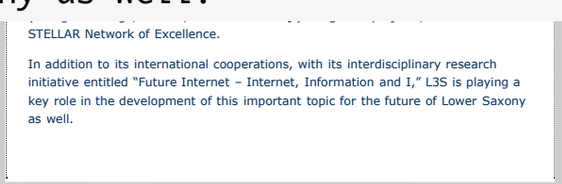
L3S Research Center

The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society, including academia and education; and connect the Internet to the real world.

In the context of a large number of projects, the L3S explores numerous issues covering the entire spectrum of challenges in Web Science as a field of research. Since its founding in 2001, the L3S has brought together numerous scholars and researchers who actively take on these challenges and perform interdisciplinary research in the fields of information retrieval, databases, the Semantic Web, performance modeling, service computing, and mobile networks. The center's total research volume is more than 6 million euros per year, with a large number of projects in the areas of

* Intelligent Access to Information
* Next Generation Internet
* E-Science

ThIn addition to its international cooperations, with its interdisciplinary research initiative entitled "Future Internet – Internet, Information and I," L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.


STELLAR Network of Excellence.

In addition to its international cooperations, with its interdisciplinary research initiative entitled "Future Internet – Internet, Information and I," L3S is playing a key role in the development of this important topic for the future of Lower Saxony as well.

```
</tr>
<tr>
    <td width="10"> </td>
    <td width="360"><br />
    <div align="center"><span class="bu">The Advisory Board visiting L3S Research Center<br />
    <br />
    </span></div>
    </td>
</tr>
ody>

3S Research Center focuses on fundamental and application-oriented research in all areas of <i>Web

ontext of a large number of projects, the L3S explores numerous issues covering the entire spectrum

<b>Intelligent Access to Information</b></li>
<b>Next Generation Internet</b></li>
<b>E-Science</b></li>
iv>
                <div class="content"><p>The L3S is a research-driven institution that attracts outs
ctivities primarily focus on research, but also include consulting and technology transfer. This is
xperience L3S has gained over the years in participating in a variety of projects financed by the E
dition to its international cooperations, with its interdisciplinary research initiative entitled &

                <!-- end iterate on instances -->
    </div>



                    </div>
                    <!-- IE column clearing -->
                    <div id="ie_clearing"> </div>
                </div>
                <!-- end: #col3 -->
        </div>
        <!-- end: #main -->
        <!-- begin: #footer -->
        <div id="footer">
                &copy;2010 L3S Research Center &bull; Appelstrasse 9a &bull; 30167 Hannover &bull;
        </div>
        <!-- end: #footer -->
</div>
```

Existing Approaches

- Machine Learning vs. Heuristics
- Site-specific Solutions
  (Rule-based Scraping, DOM, Text, Link Graph)
- Vision-based models
- Tokens, N-Grams
- Shallow Text Features
- Context

<h2>Hello World!</h2><p>This is a test.

Shallow Text Features

- Examine Document at Text Block Level
  - Numbers: Words, Tokens contained in block
  - Average Lengths: Tokens, Sentences
  - Ratios: Uppercased words, full stops
  - Classes: Block-level HTML tags <P>, <H*n*>, <DIV>
  - Densities: Link Density (Anchor Text Percentage), **Text Density**

Text Density

Kohlschütter/Nejdl [CIKM2008]
Kohlschütter [WWW2009]

Wrap text at a fixed line width (e.g. 80 chars)

> The L3S Research Center focuses on fundamental and application-oriented research in all areas of Web Science. L3S researchers develop new methods and technologies that enable intelligent, seamless access to information via the Web; link individuals and communities in all areas of the knowledge society,

$$\rho(b) = \frac{\#\text{ tokens in b}}{\#\text{ wrapped lines in b}}$$

> About L3S
> Contact
> Organigram
> Vision 2009-2013
> Mentoring Guidelines

Contextual Features

- Intra-Document:
  - Relative/Absolute Position of Block
  - Features of the previous/next block
- Inter-Document
  - Text Block Frequency

Experiments

1. Classification Accuracy?
   Decision Trees, SVM, 10-fold cross validation, F-Measure/ROC AuC, ...
2. Main Content Extraction
   Compare to BTE (Finn et al., 2001) and n-grams (Pasternack et al., 2009)
   In Paper also: Victor (Spousta et al., 2008), NCleaner (Evert, 2008)
3. Ranking Improvement?
   Precision@10, NDCG@10
   50 top-k TREC-Queries for BLOGS06 (3M docs)

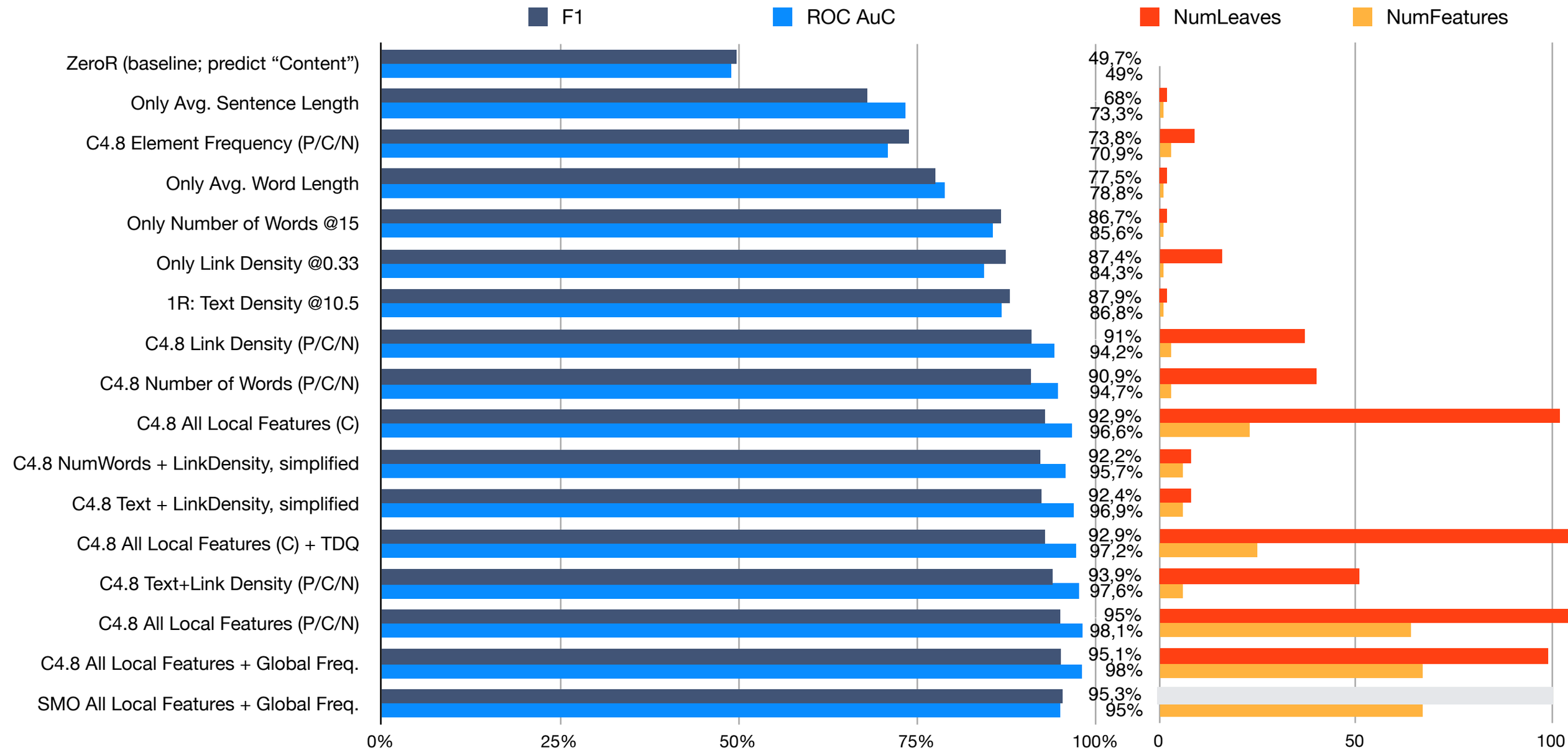
GoogleNews Dataset

- L3S-GN1
  621 news articles from 408 web sites, randomly sampled from a 254,000 pages crawl of English Google News over 4 months, manually assessed by L3S colleagues

Class	# Blocks	# Words	# Tokens
Total	72662	520483	644021
Boilerplate	79%	35%	46%
Any Content	21%	65%	54%
Headline	1%	1%	1%
Article Full-text	12%	51%	42%
Supplemental	3%	3%	2%
User Comments	1%	1%	1%
Related Content	4%	9%	8%

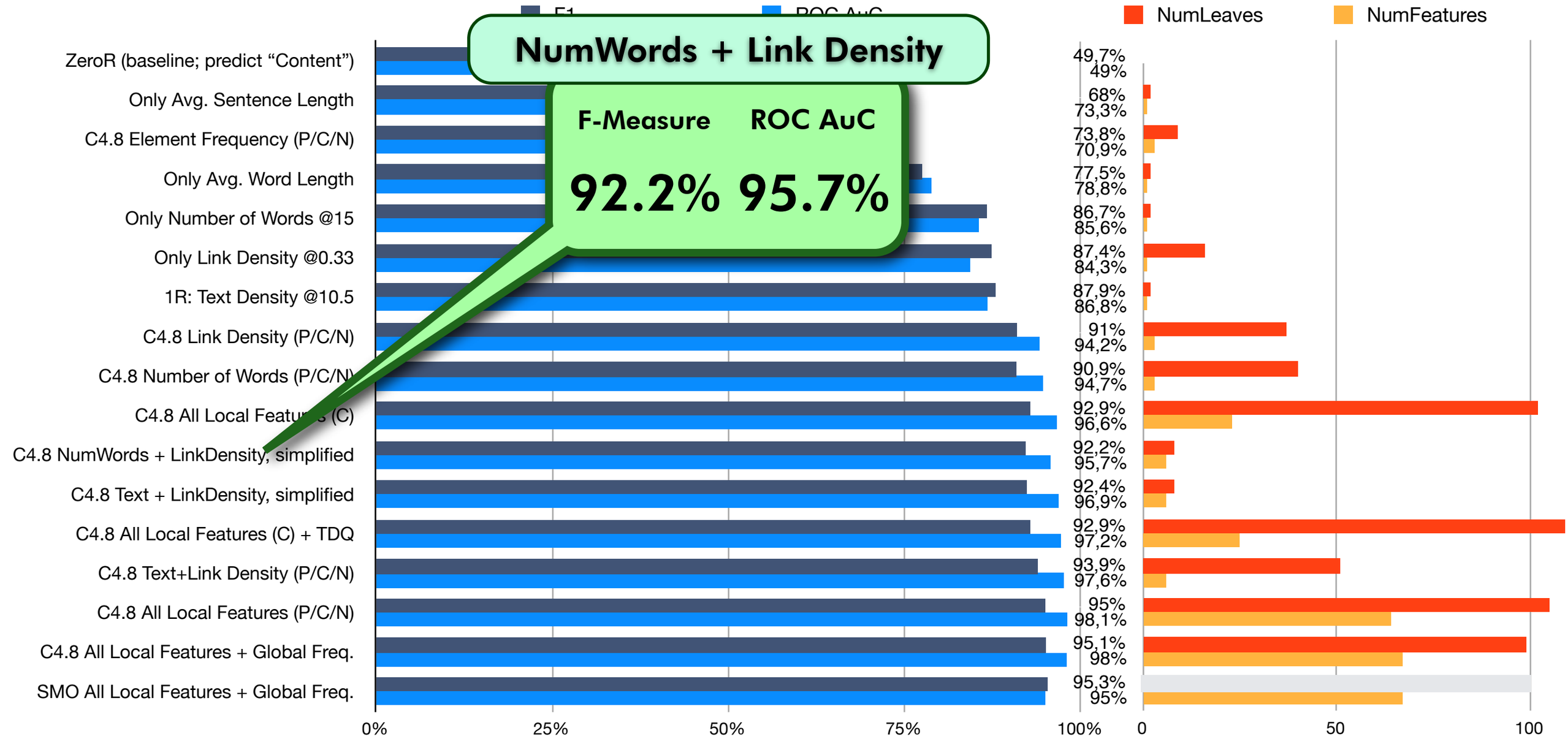
Classification Accuracy

Block-Level (weighted by number of words)



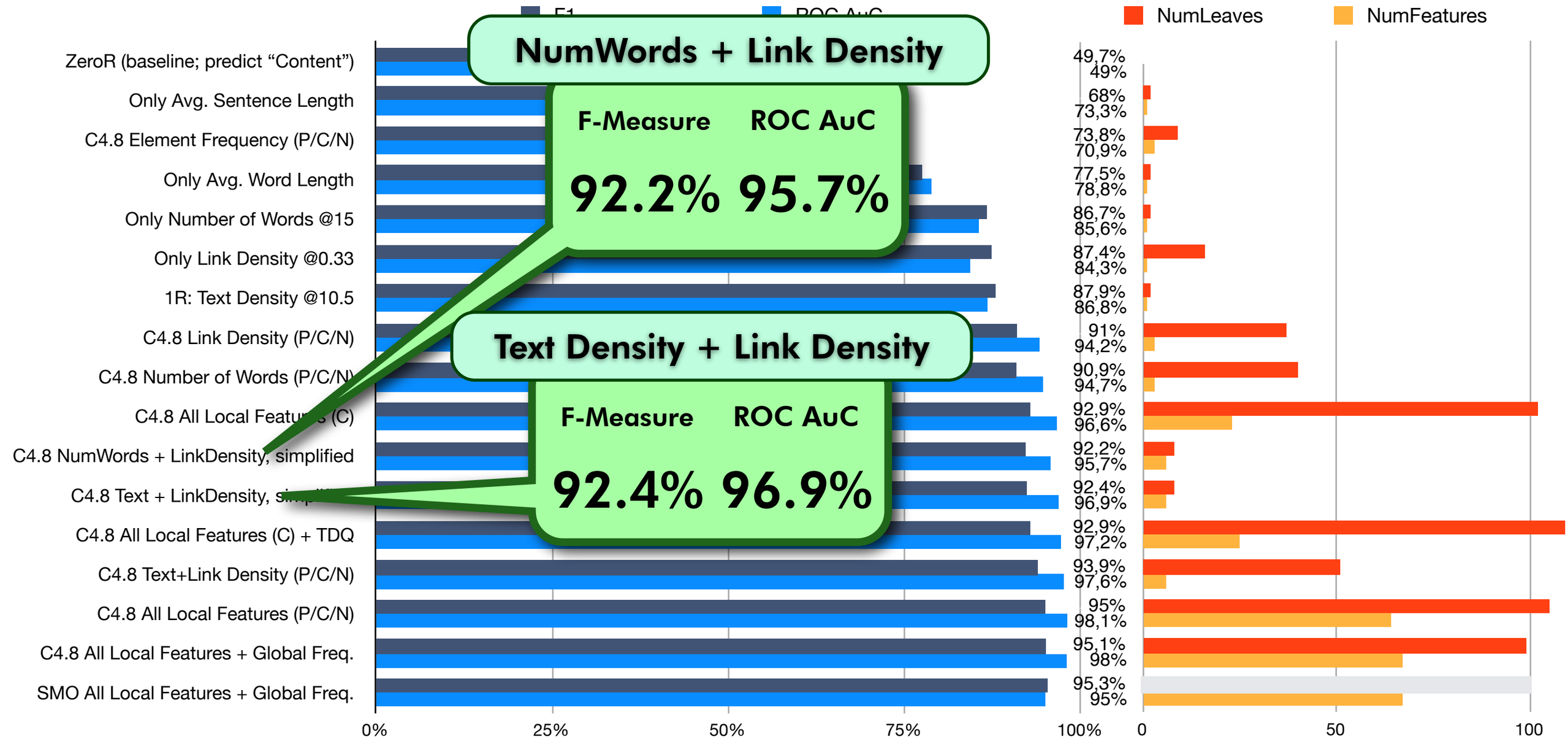
Classification Accuracy

Block-Level (weighted by number of words)



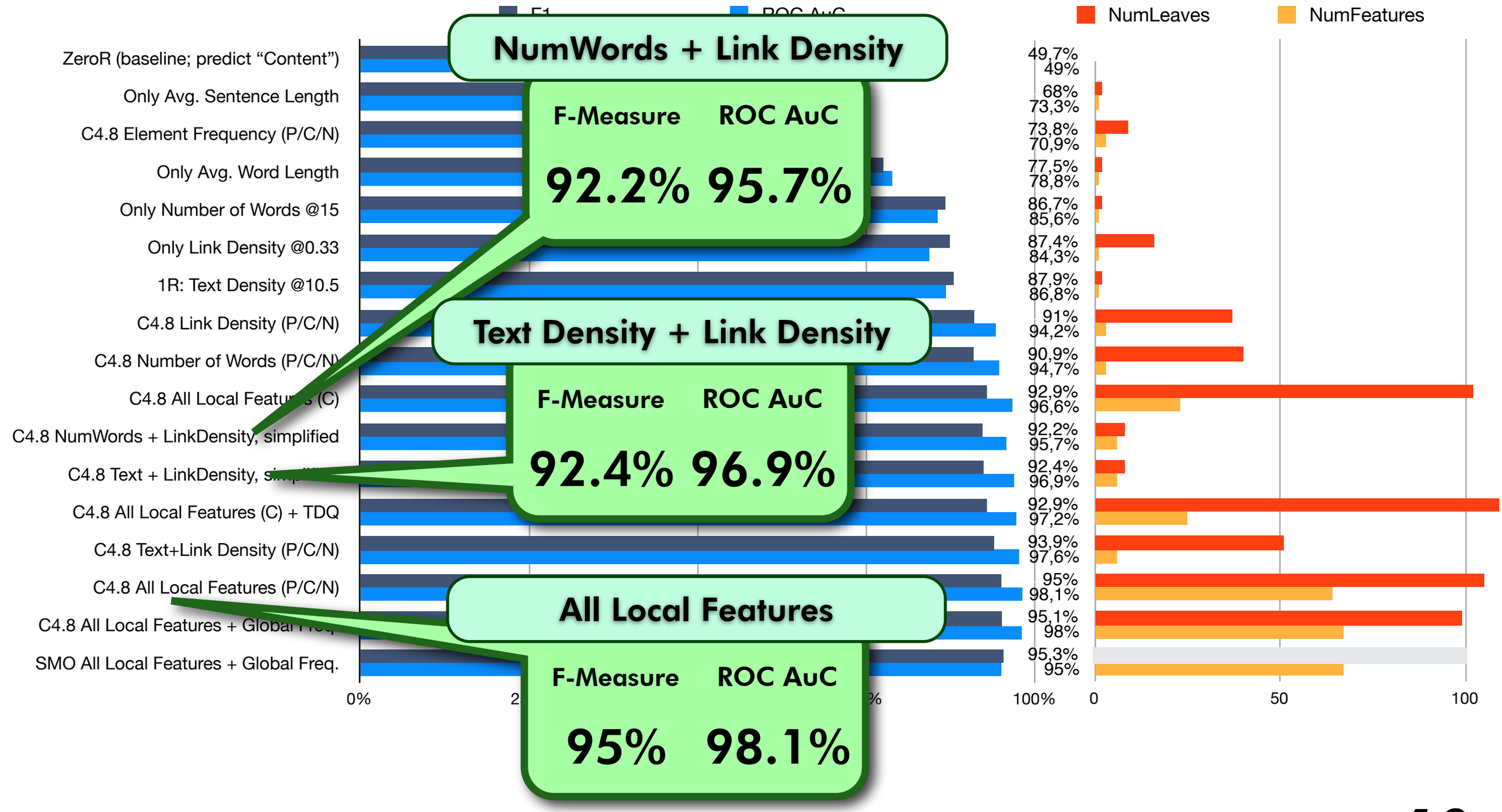
Classification Accuracy

Block-Level (weighted by number of words)



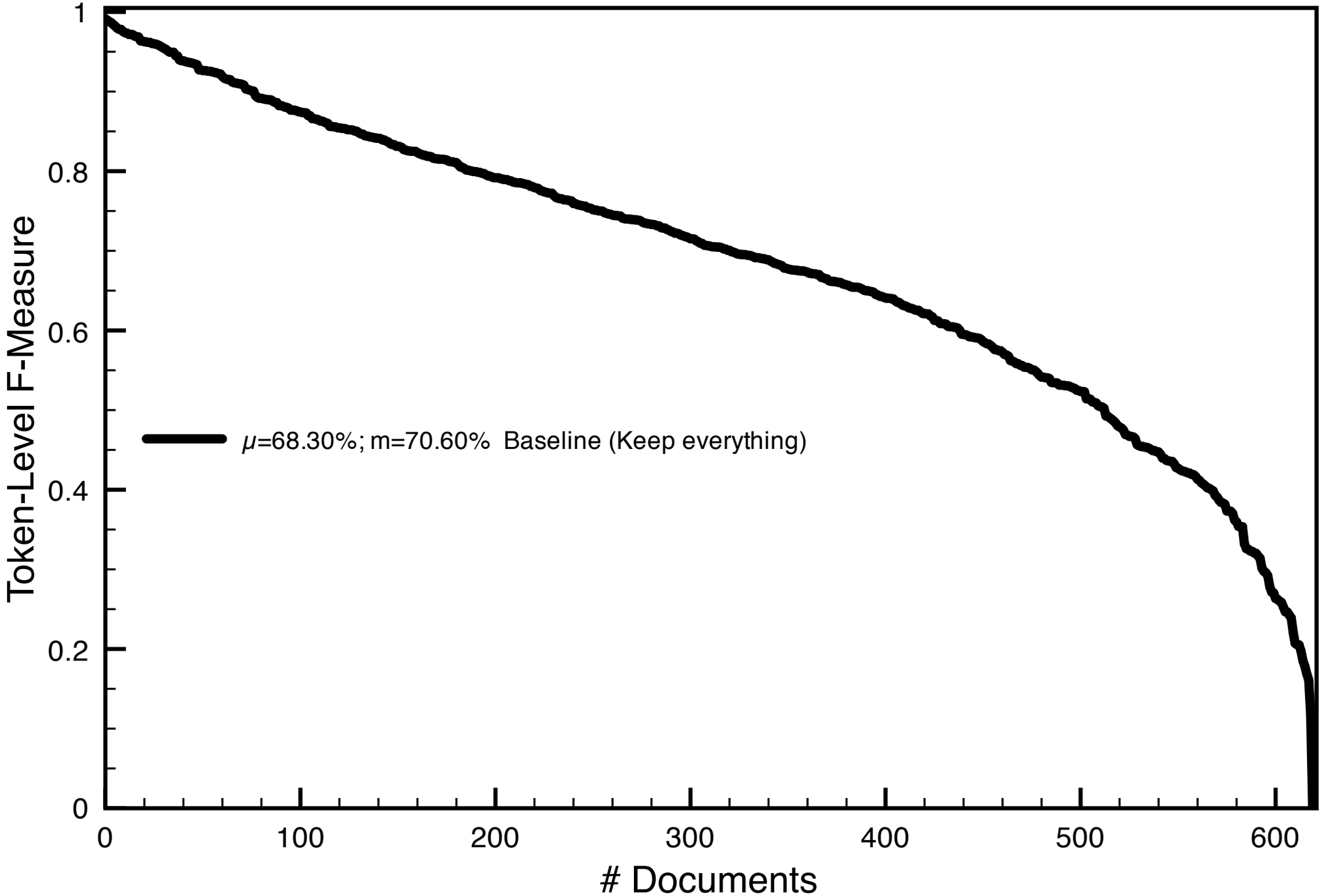
Classification Accuracy

Block-Level (weighted by number of words)

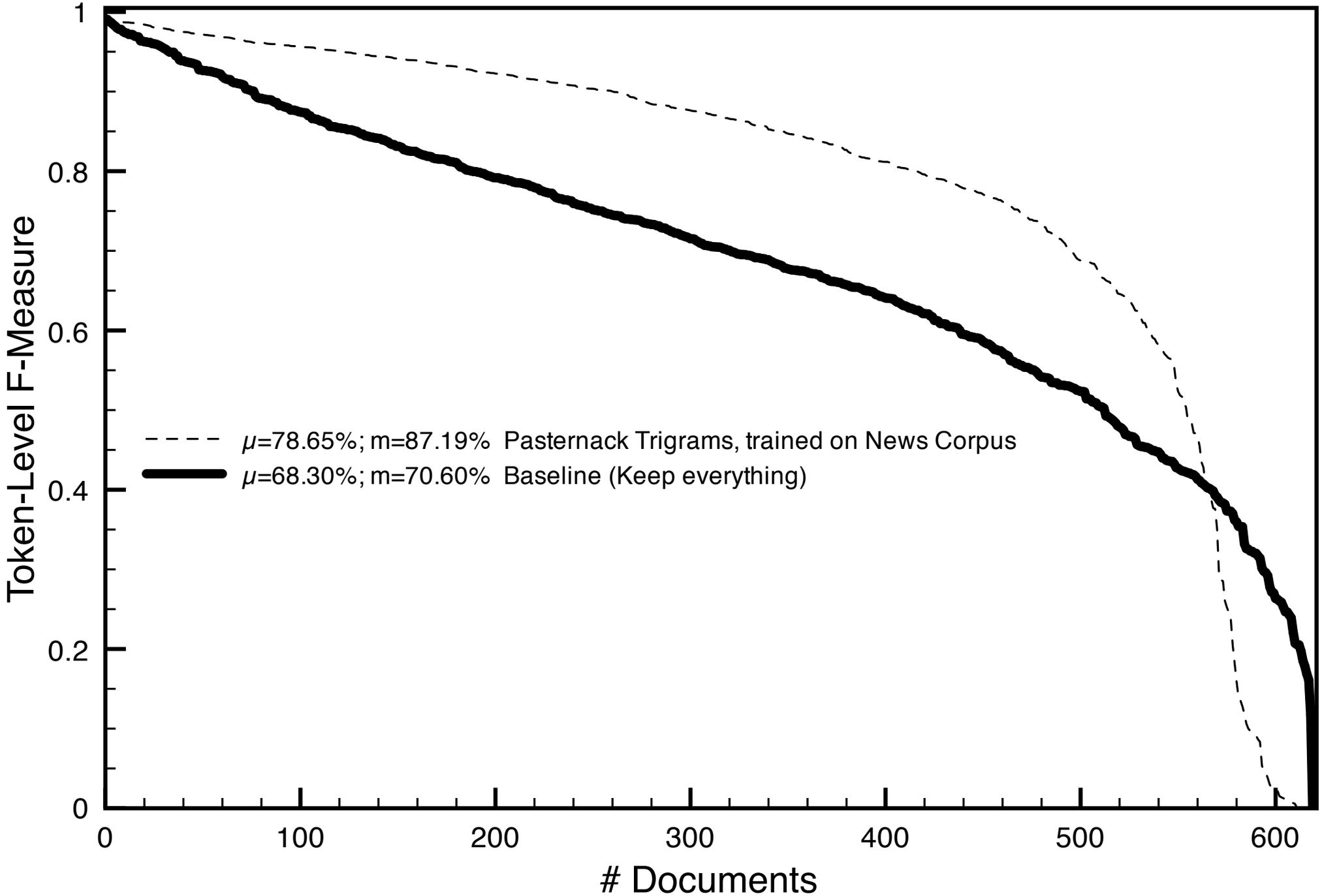


"*Main Content*" *Extraction*

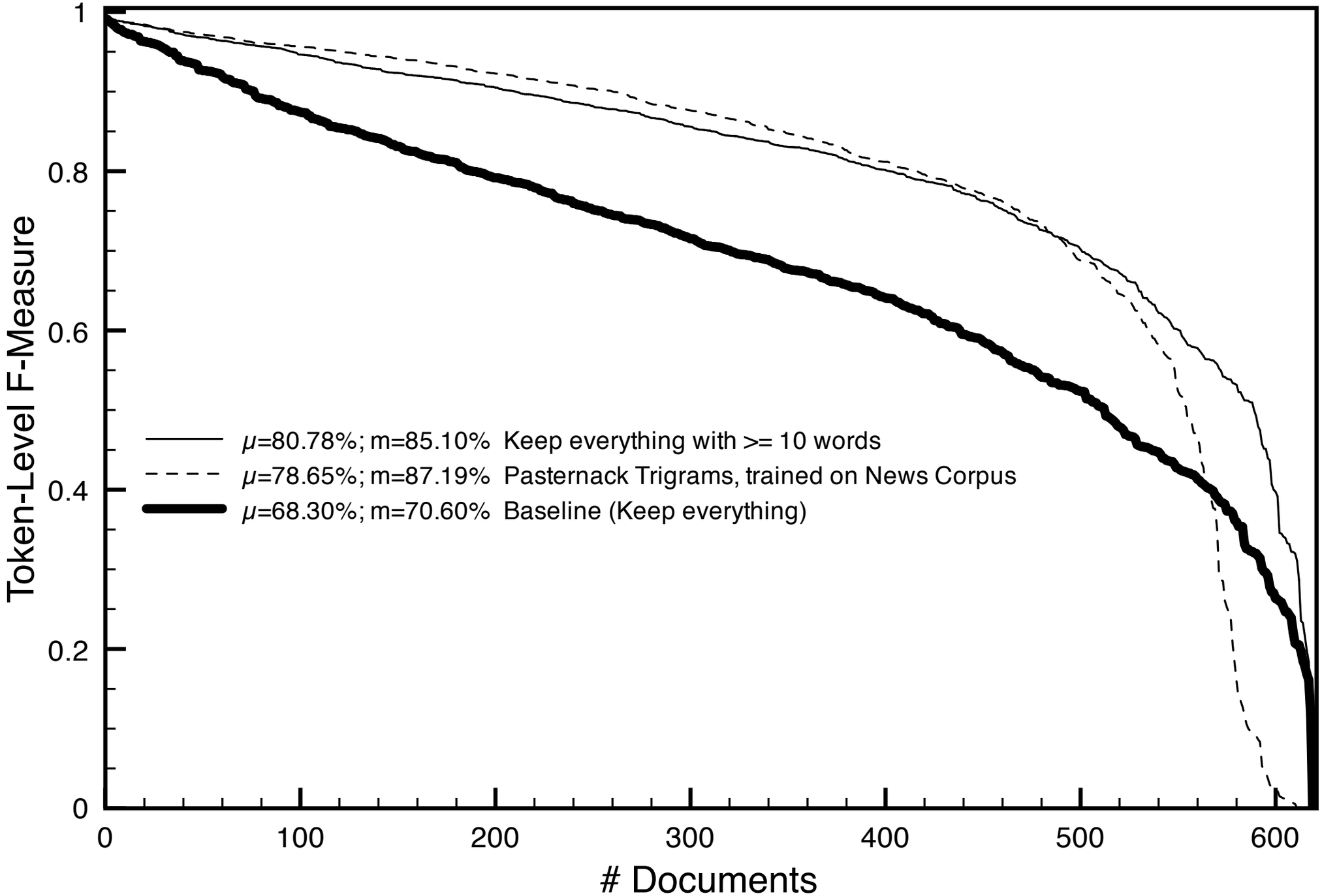
"*Main Content" Extraction*



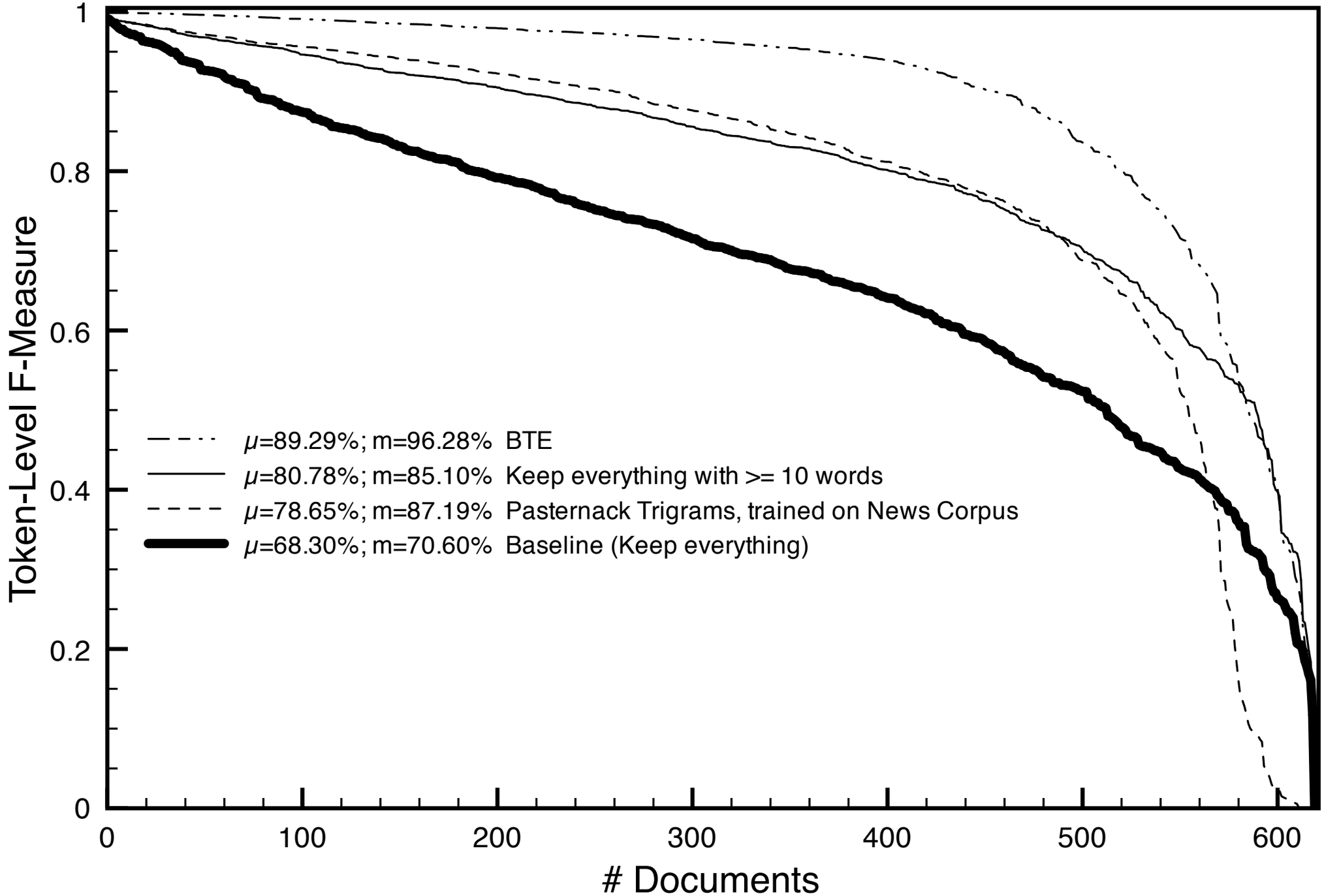
"*Main Content" Extraction*



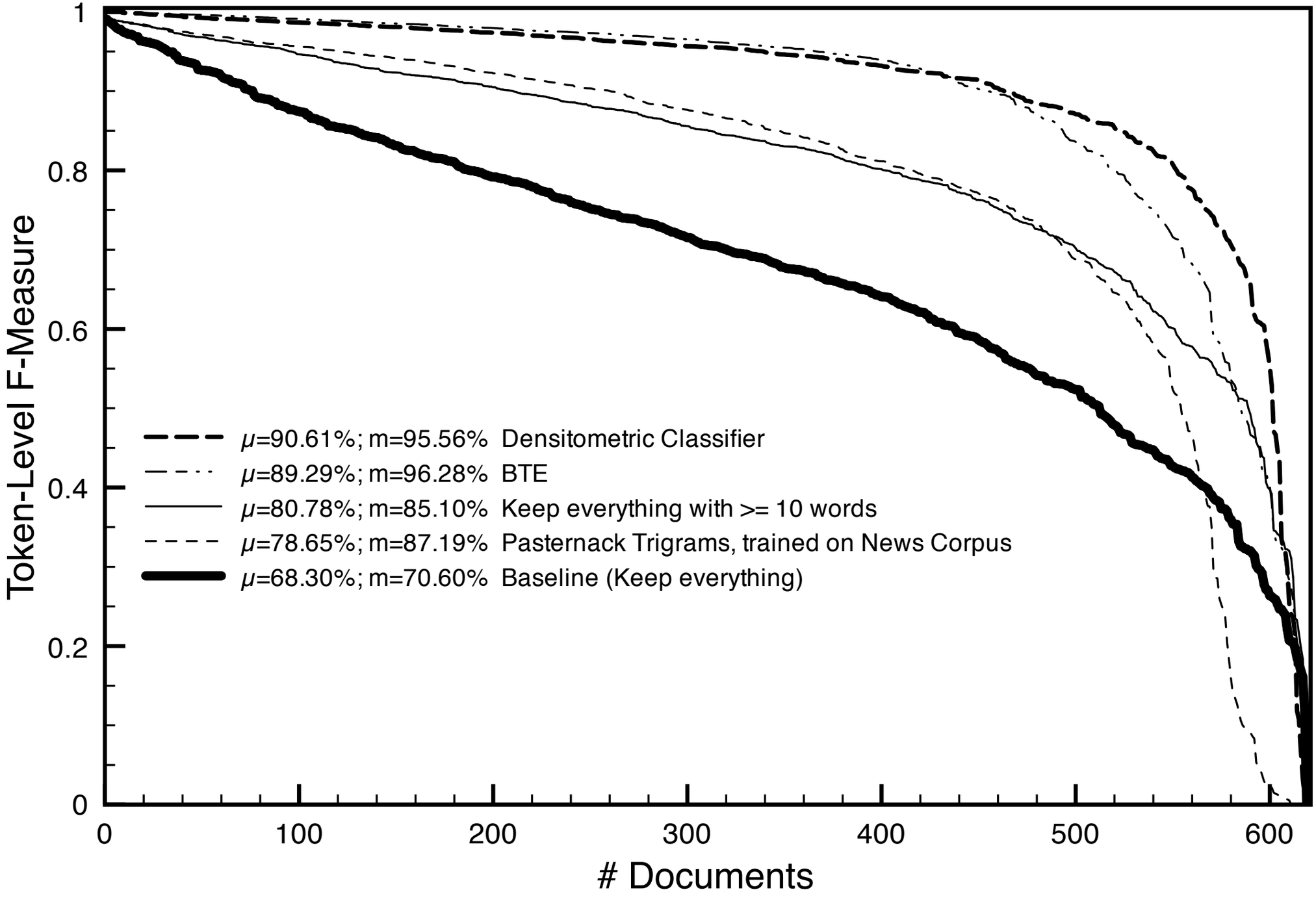
"*Main Content" Extraction*



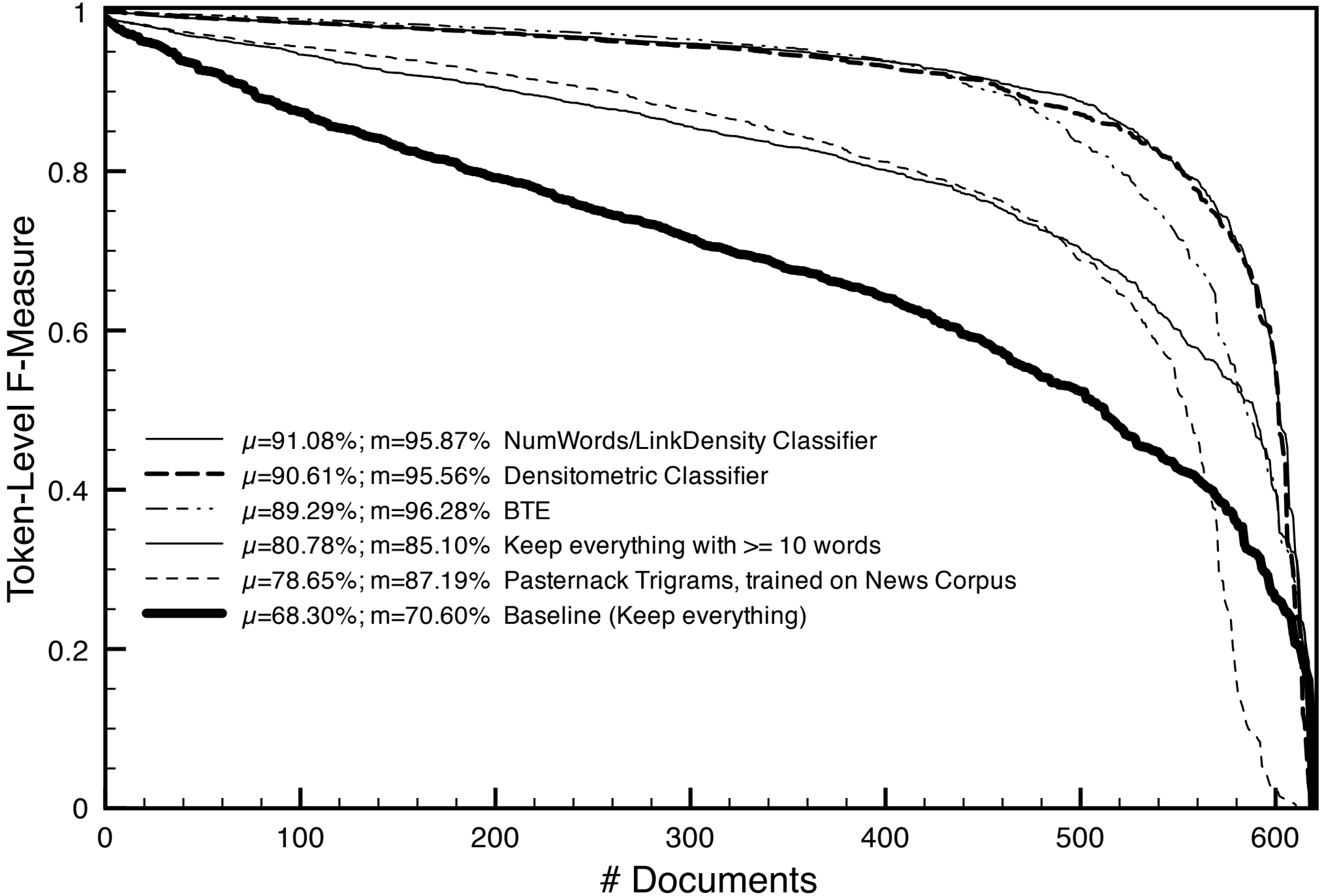
"*Main Content" Extraction*



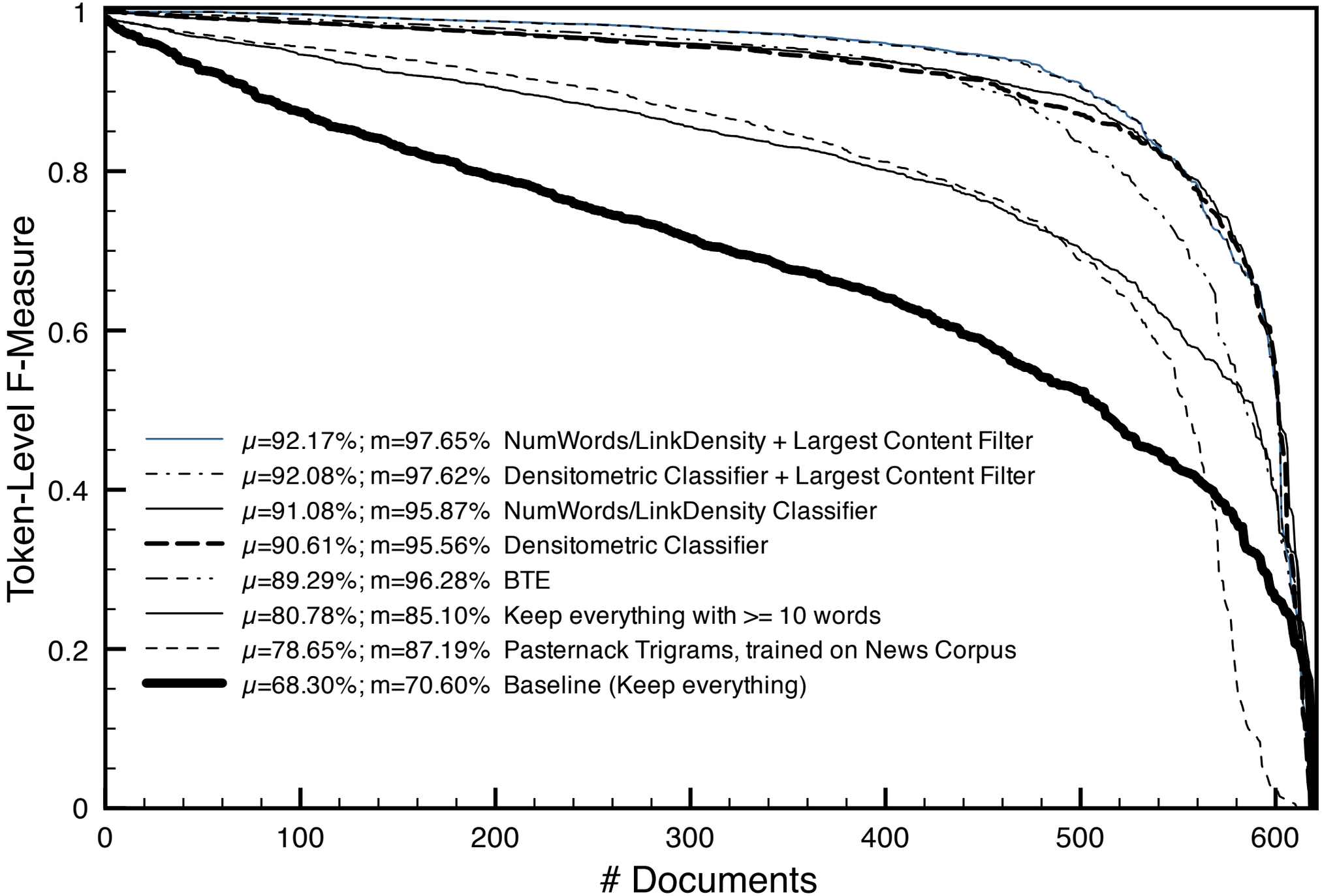
"*Main Content" Extraction*



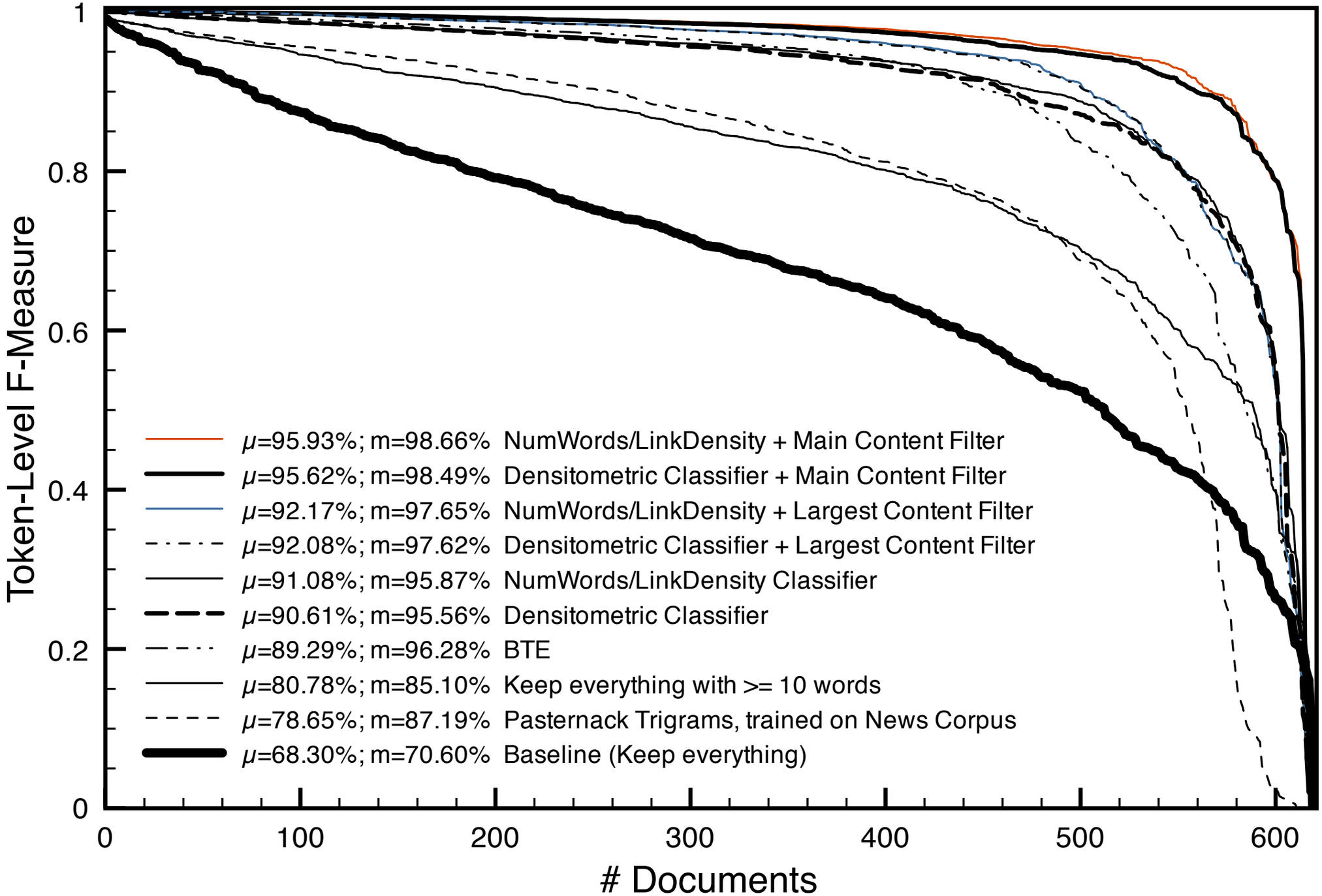
"*Main Content" Extraction*



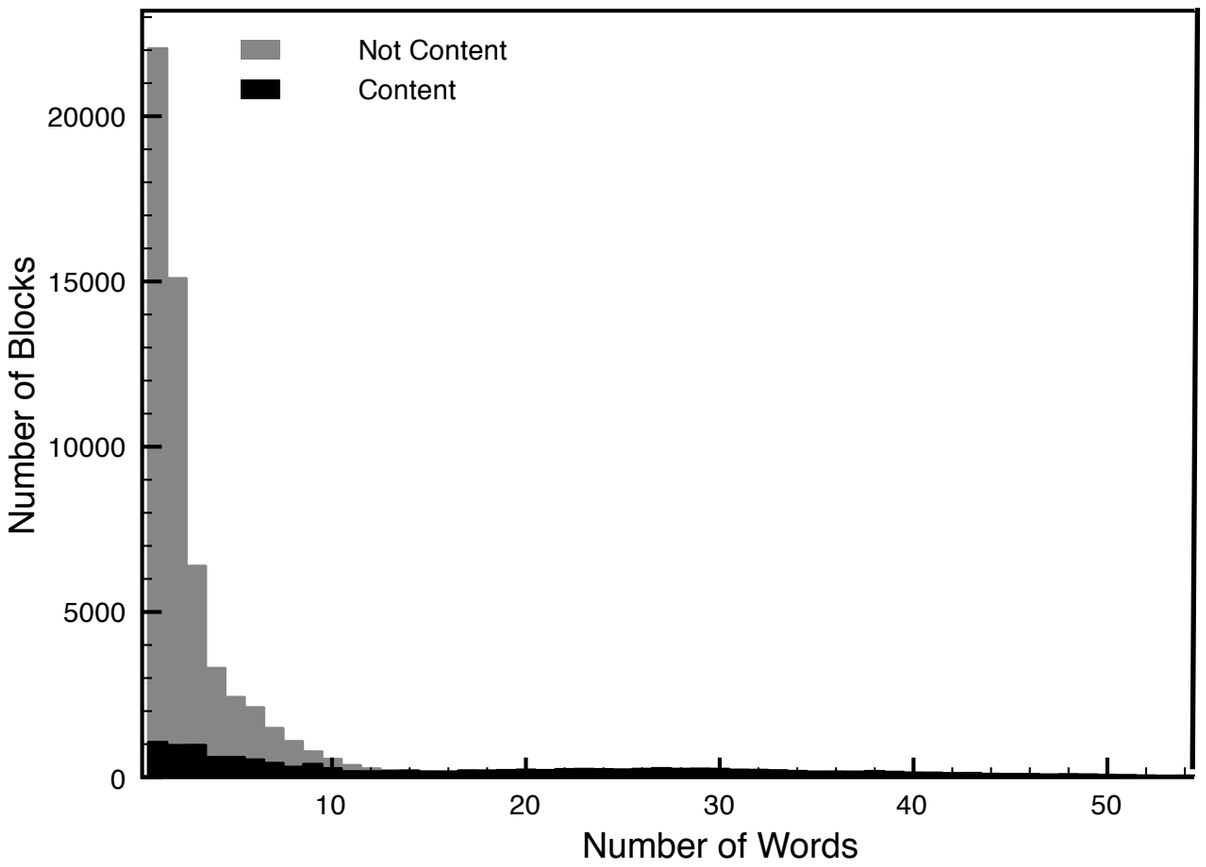
"*Main Content" Extraction*



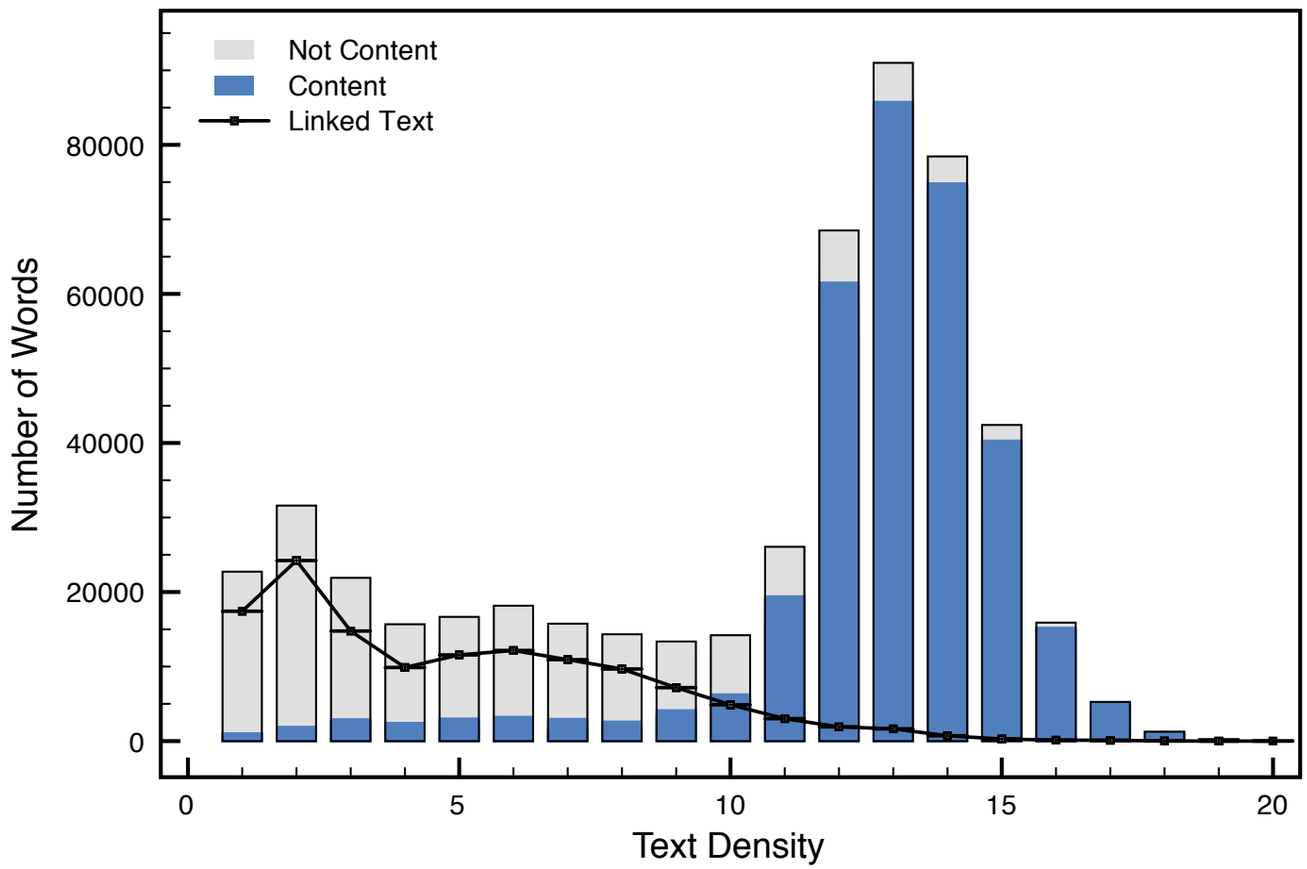
"Main Content" Extraction



Number of Words



Text Density



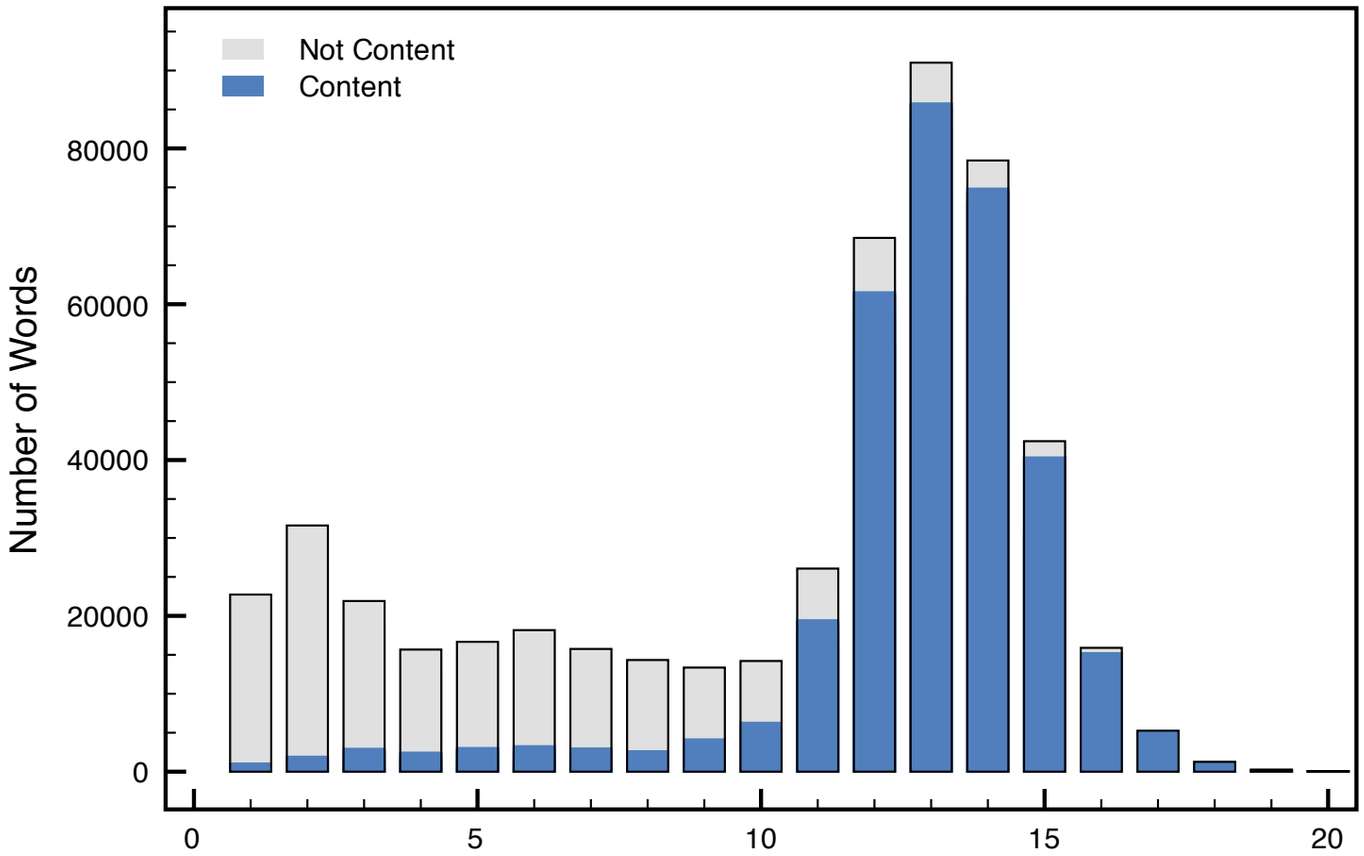
NumWords + Link Density

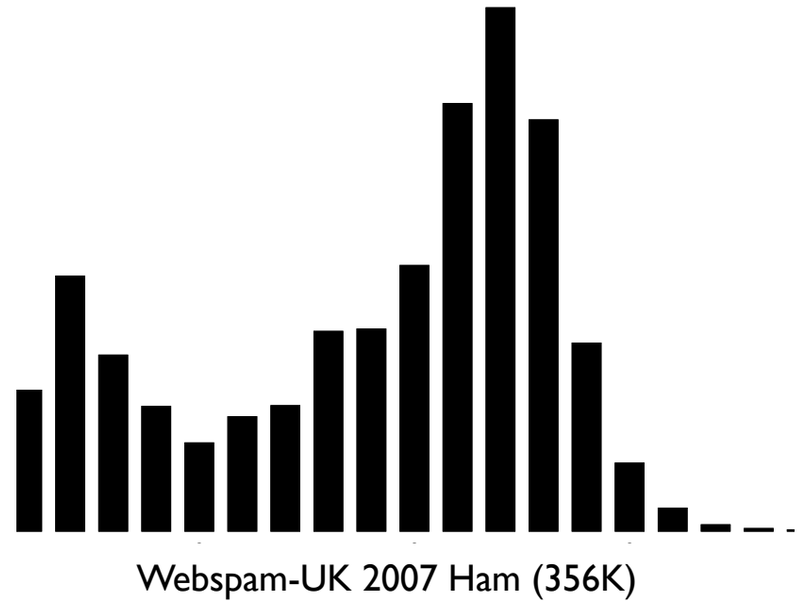
```
curr_linkDensity <= 0.333333
| prev_linkDensity <= 0.555556
| | curr_numWords <= 16
| | | next_numWords <= 15
| | | | prev_numWords <= 4: BOILERPLATE
| | | | prev_numWords > 4: CONTENT
| | | next_numWords > 15: CONTENT
| | curr_numWords > 16: CONTENT
| prev_linkDensity > 0.555556
| | curr_numWords <= 40
| | | next_numWords <= 17: BOILERPLATE
| | | next_numWords > 17: CONTENT
| | curr_numWords > 40: CONTENT
curr_linkDensity > 0.333333: BOILERPLATE
```

Text Density + Link Density

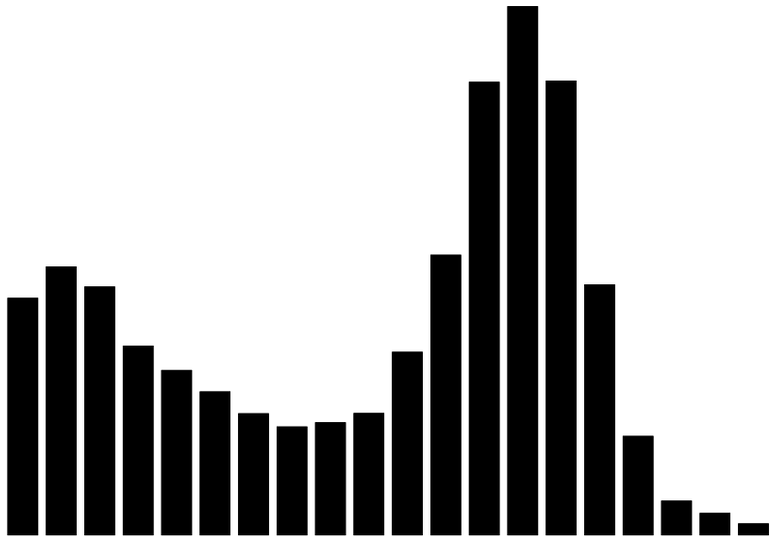
```
curr_linkDensity <= 0.333333
| prev_linkDensity <= 0.555556
| | curr_textDensity <= 9
| | | next_textDensity <= 10
| | | | prev_textDensity <= 4: BOILERPLATE
| | | | prev_textDensity > 4: CONTENT
| | | next_textDensity > 10: CONTENT
| | curr_textDensity > 9
| | | next_textDensity = 0: BOILERPLATE
| | | next_textDensity > 0: CONTENT
| prev_linkDensity > 0.555556
| | next_textDensity <= 11: BOILERPLATE
| | next_textDensity > 11: CONTENT
curr_linkDensity > 0.333333: BOILERPLATE
```

GoogleNews L3S-GN1

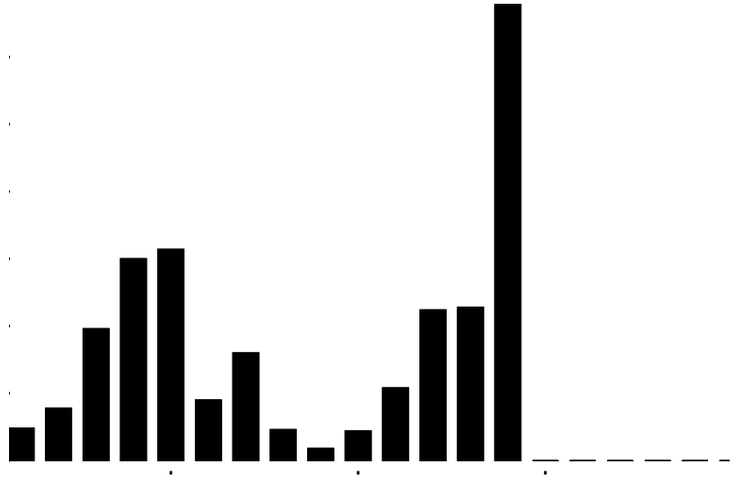




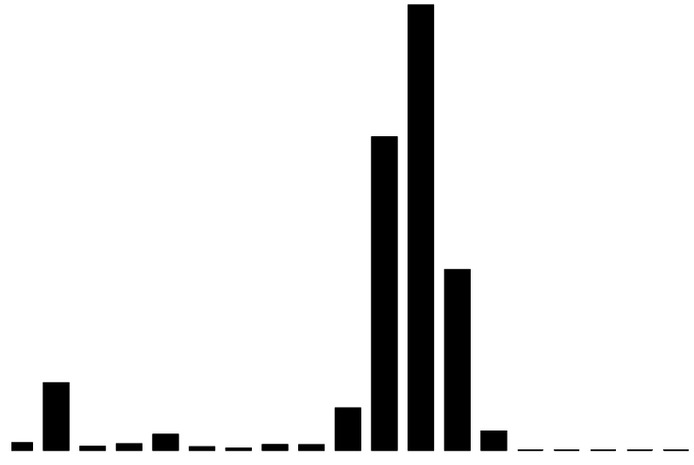
Webspam-UK 2007 Ham (356K)



BLOGS06 (3M)



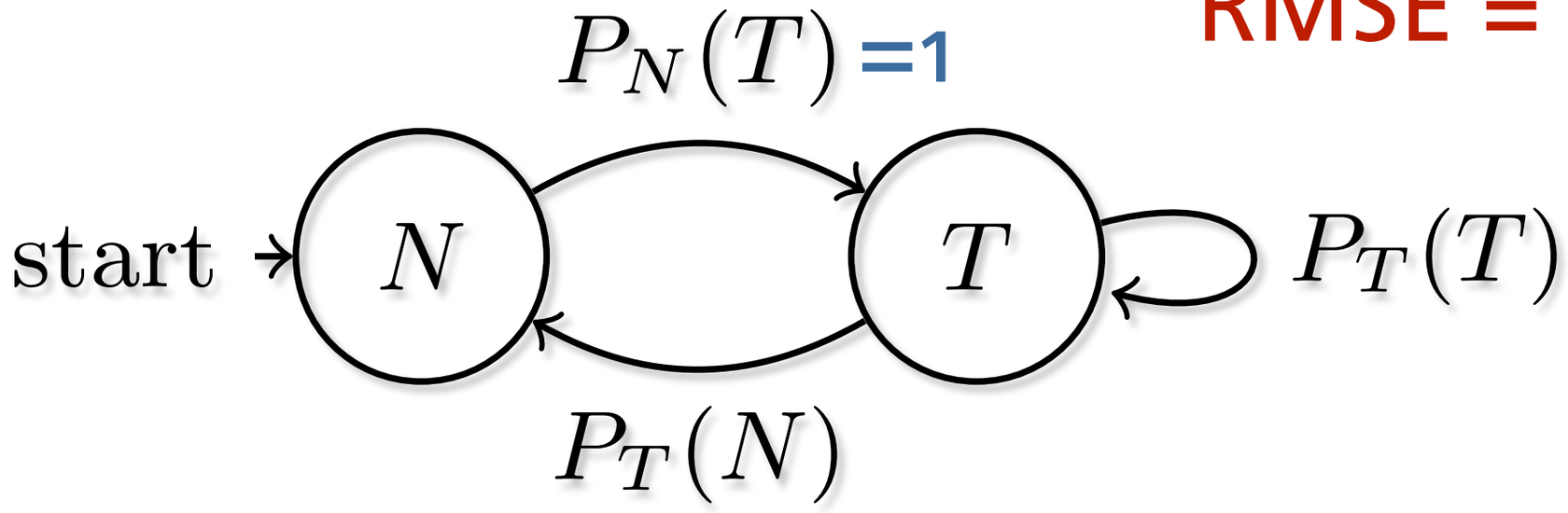
Invidiual web page
About.com: New York City Travel



WSDM Paper
Kohlschütter, Fankhauser, Nejdl

Shannon Random Writer

$R^2_{adj} = 96.7\%$
RMSE = 0.0046

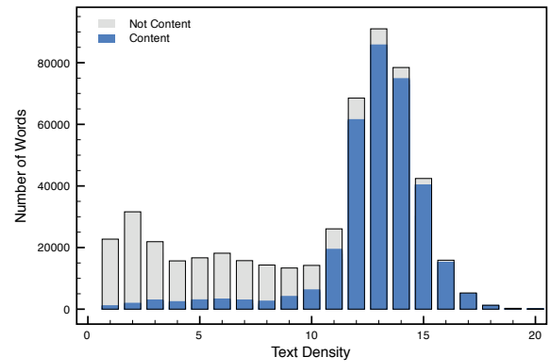


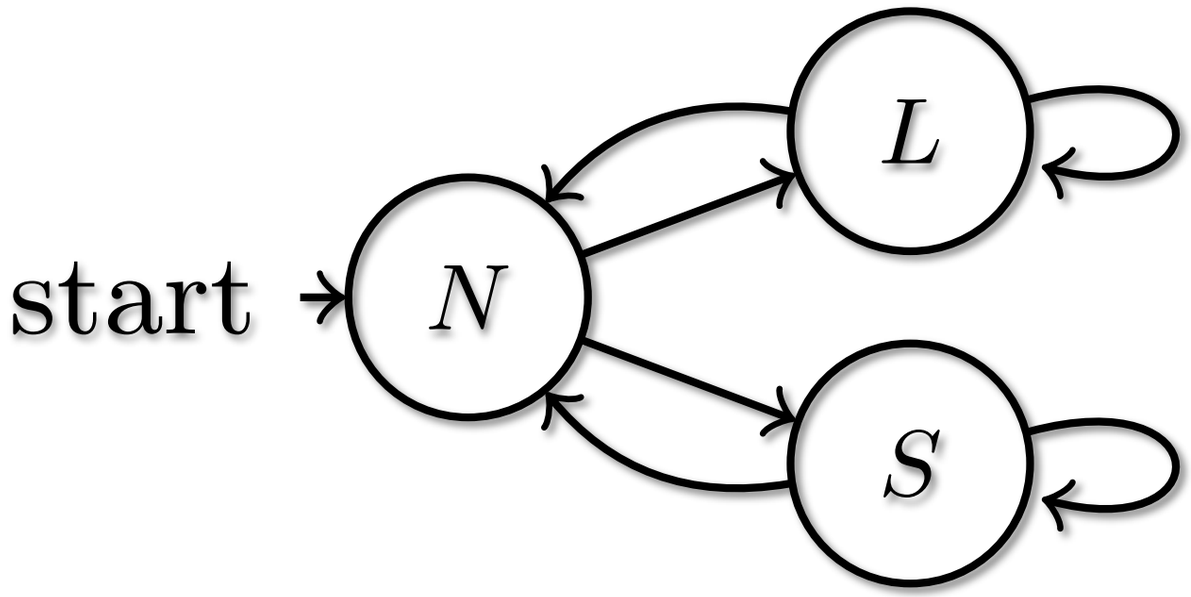
Bernoulli trial: Transition to next block is success p
emission of another word is failure *1-p*

$$Pr(Y = k) = (1 - p)^k p$$

$$Pr(Y = x) = (1 - p)^{x-1} \cdot p = P_T(T)^{x-1} \cdot P_T(N)$$

Stratified Model



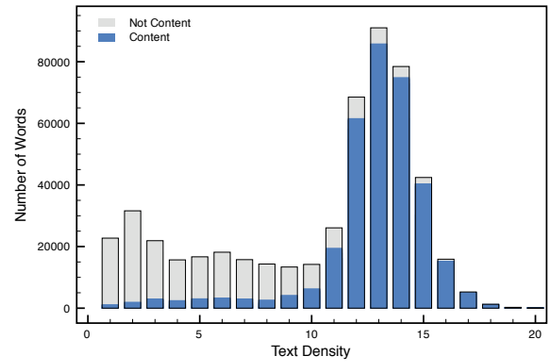


L = "Long Text"
S = "Short Text"

$$P_S(N) \gg P_L(N)$$
$$P_N(L) = 1 - P_N(S)$$

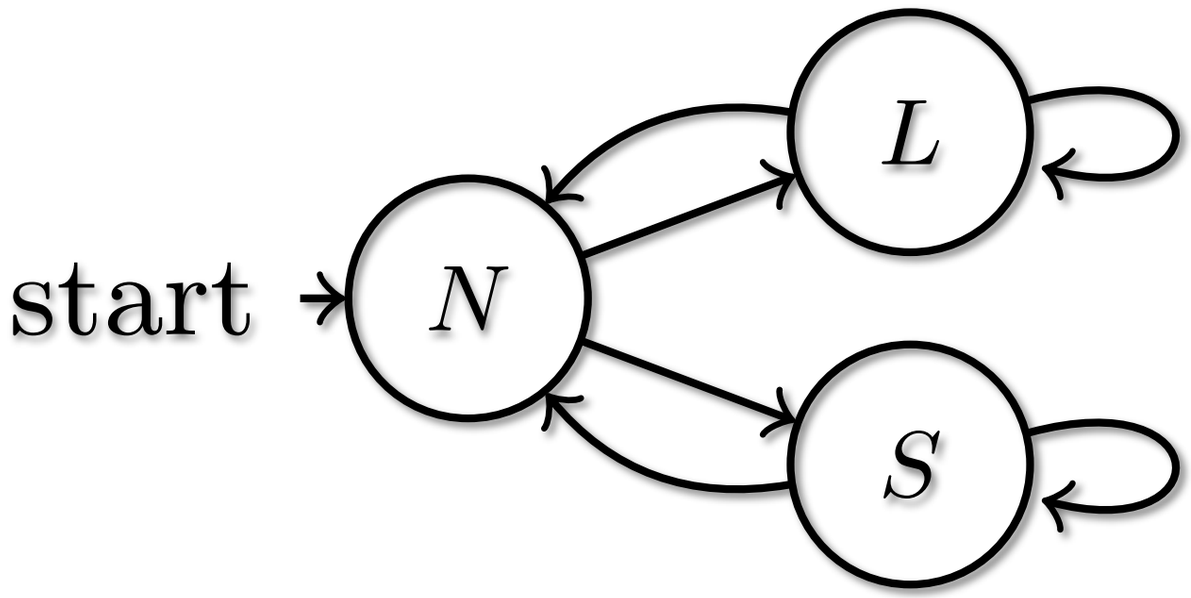
$$Pr(Y = x) = P_N(S) \cdot \left[P_S(S)^{x-1} \cdot P_S(N)\right] +$$
$$+ P_N(L) \cdot \left[P_L(L)^{x-1} \cdot P_L(N)\right]$$

Stratified Model



$R^2_{adj} = 98.8\%$

$RMSE = 0.0027$



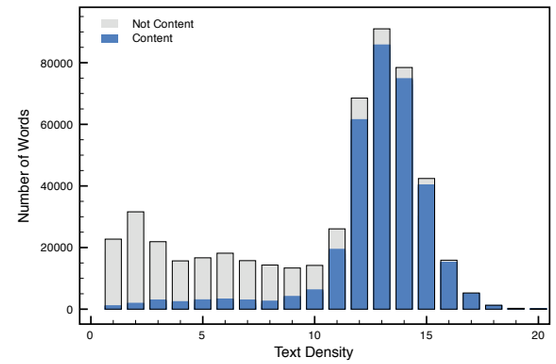
L = "Long Text"

S = "Short Text"

$$P_S(N) \gg P_L(N)$$
$$P_N(L) = 1 - P_N(S)$$

$$Pr(Y = x) = P_N(S) \cdot \left[P_S(S)^{x-1} \cdot P_S(N)\right] +$$
$$+ P_N(L) \cdot \left[P_L(L)^{x-1} \cdot P_L(N)\right]$$

Stratified Model

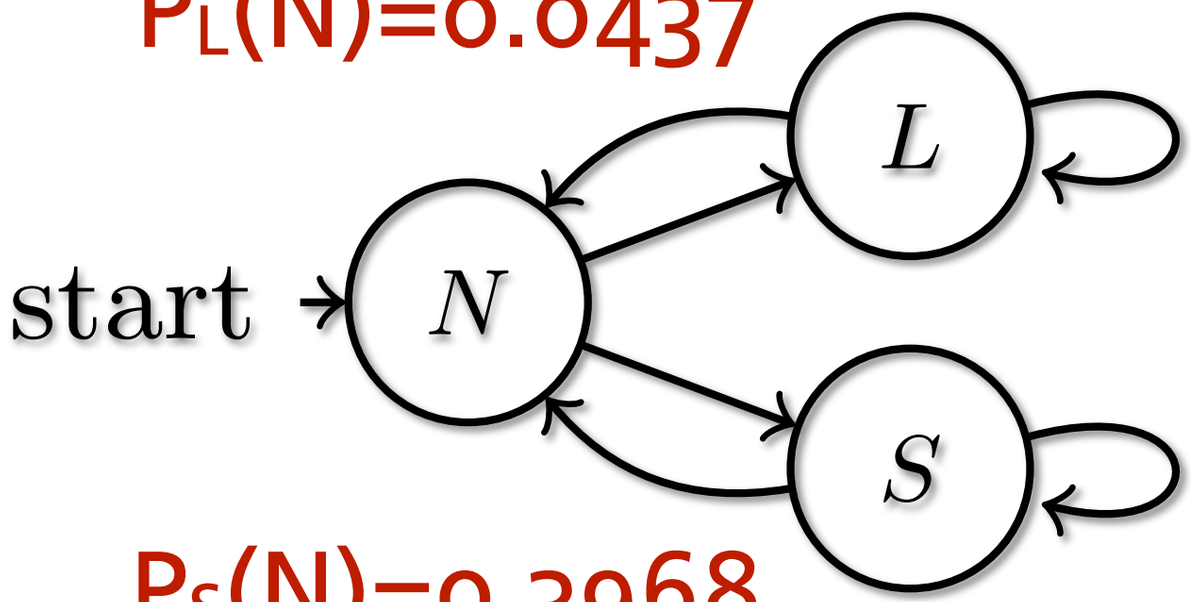


$1 + E = 1 + 1/p = 23.8$

$P_L(N)=0.0437$

$R^2_{adj} = 98.8\%$

$RMSE = 0.0027$



$P_S(N)=0.3968$

$1 + E = 1 + 1/p = 3.52$

L = "Long Text"

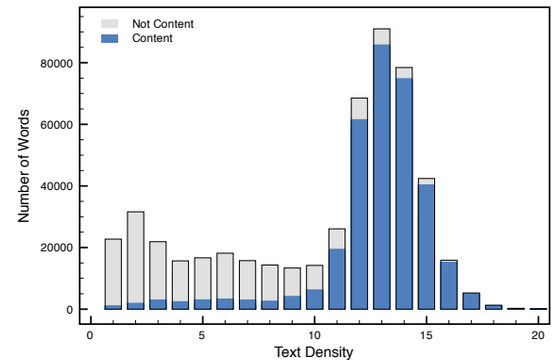
S = "Short Text"

$$P_S(N) \gg P_L(N)$$

$$P_N(L) = 1 - P_N(S)$$

$$Pr(Y = x) = P_N(S) \cdot \left[P_S(S)^{x-1} \cdot P_S(N)\right] + \\ + P_N(L) \cdot \left[P_L(L)^{x-1} \cdot P_L(N)\right]$$

Stratified Model

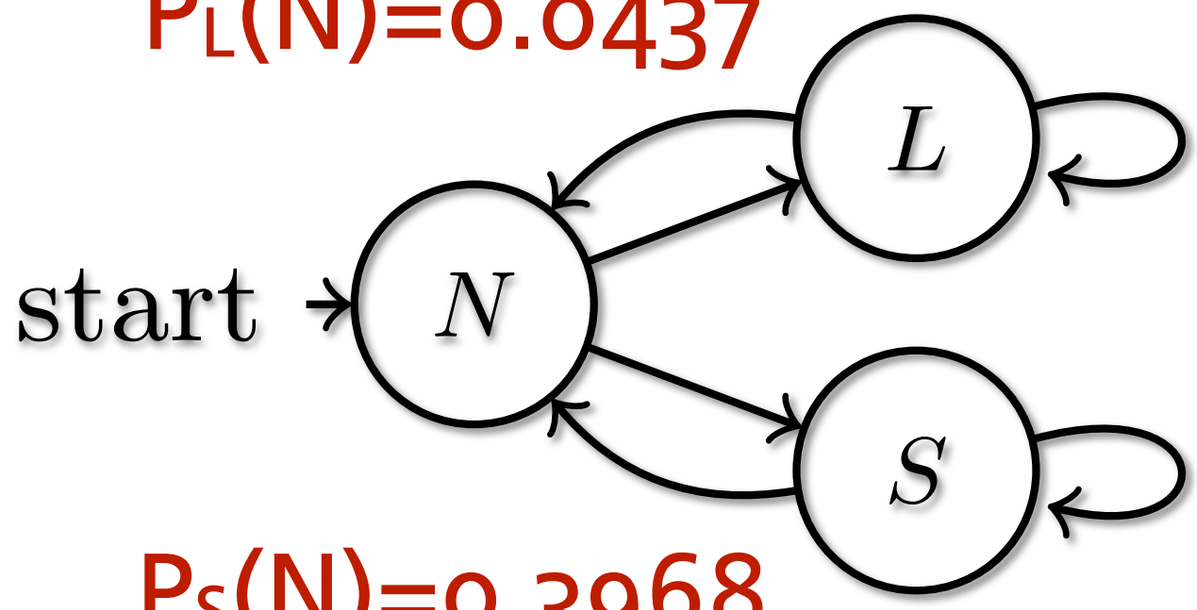


$1 + E = 1 + 1/p = 23.8$

$P_L(N)=0.0437$

$R^2_{adj} = 98.8\%$

$RMSE = 0.0027$



L = "Long Text"

S = "Short Text"

$$P_S(N) \gg P_L(N)$$

$$P_N(L) = 1 - P_N(S)$$

$P_S(N)=0.3968$

$1 + E = 1 + 1/p = 3.52$

$P_N(S)=76\%$

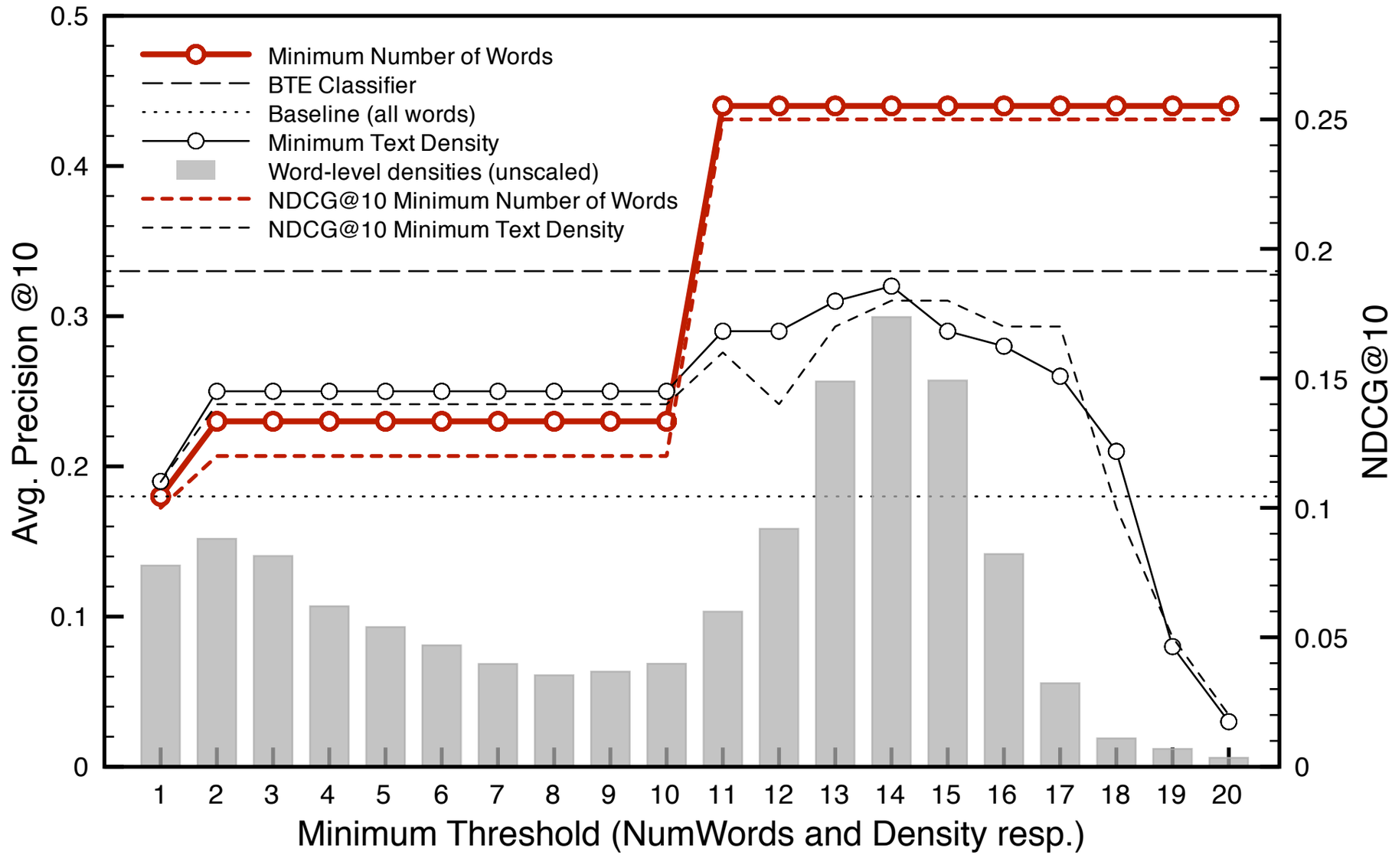
GoogleNews assessment:
79% of blocks were boilerplate

$$Pr(Y = x) = P_N(S) \cdot \left[P_S(S)^{x-1} \cdot P_S(N)\right] + \\ + P_N(L) \cdot \left[P_L(L)^{x-1} \cdot P_L(N)\right]$$

Retrieval Experiment

Baseline: P@10=0.18; NDCG@10=0.0985

BTE: P@10=0.33; NDCG@10=0.1627

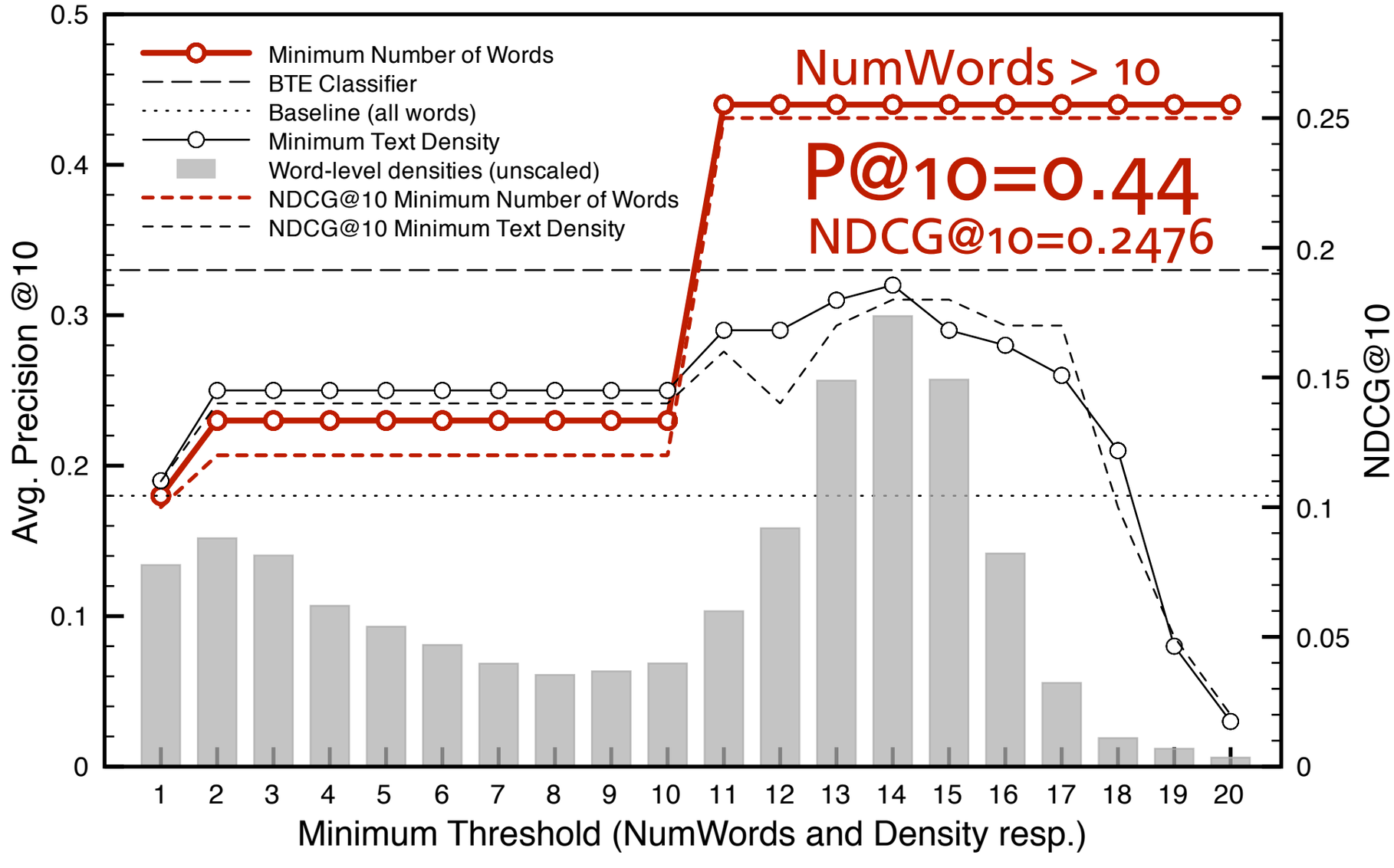


50 top-k TREC queries on BLOGS06 dataset (~3M docs)

Retrieval Experiment

Baseline: P@10=0.18; NDCG@10=0.0985

BTE: P@10=0.33; NDCG@10=0.1627

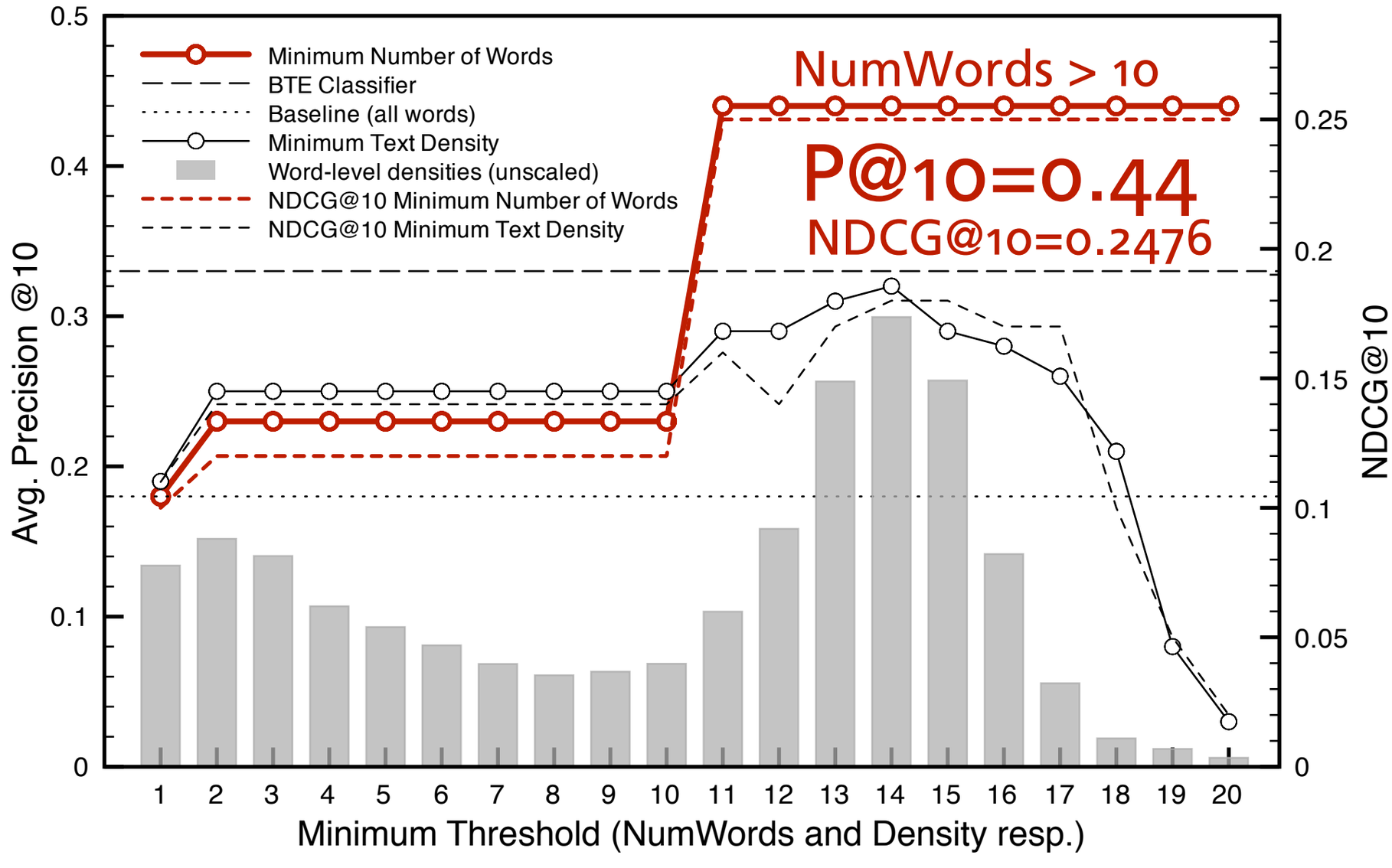


50 top-k TREC queries on BLOGS06 dataset (~3M docs)

Retrieval Experiment

Improvement over Baseline: 144%/151% P@10=0.18; NDCG@10=0.0985

Improvement over BTE: 33%/ 52% P@10=0.33; NDCG@10=0.1627



50 top-k TREC queries on BLOGS06 dataset (~3M docs)

Conclusions

Conclusions

- Text Creation can be modeled as a Stratified Stochastic Process

Conclusions

- Text Creation can be modeled as a Stratified Stochastic Process
- Very high Classification/Extraction Accuracy (92-98%) at almost no cost

Conclusions

- Text Creation can be modeled as a Stratified Stochastic Process
- Very high Classification/Extraction Accuracy (92-98%) at almost no cost
- Increase of Retrieval Precision (33%-151%) at almost no cost

Next Steps

- Multi-Lingual, Multi-Domain Corpora
- Further explore the relationship to Quantitative Linguistics
- Model Linking Behavior
- Use it, for free (Apache 2.0 License)
  http://boilerpipe.googlecode.com

KOHLSCHUETTER@L3S.DE