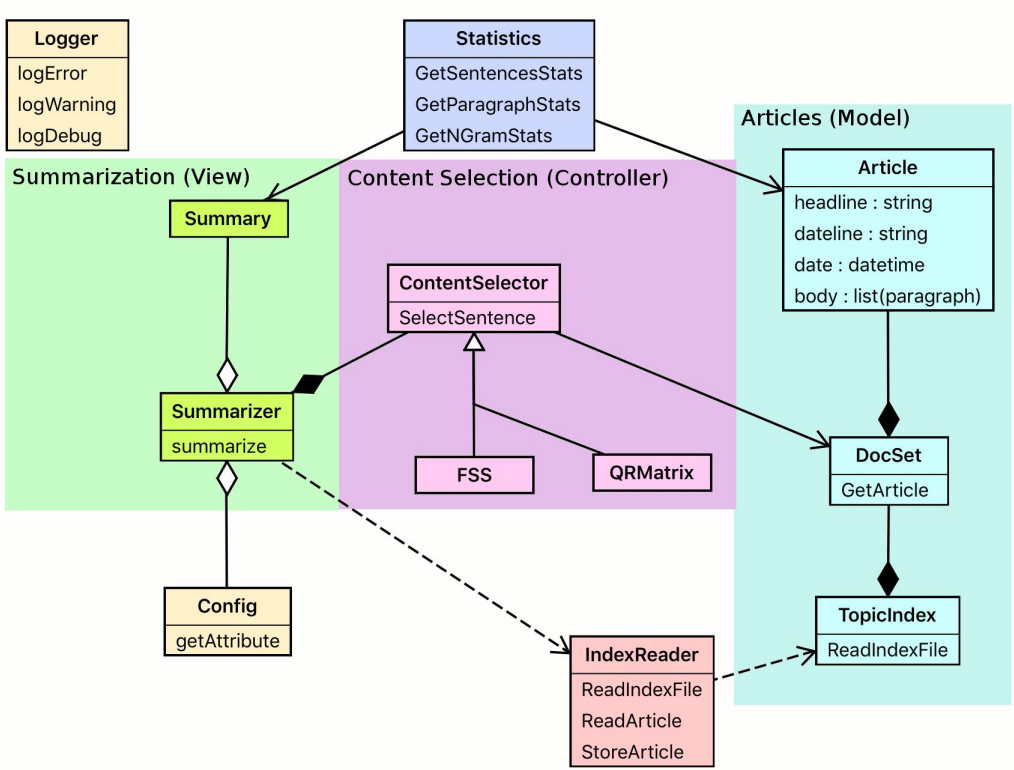
Summarization D3

Team E2JK++

Eric Lindberg
John Greve
Josh Mathias
Kekoa Riggin

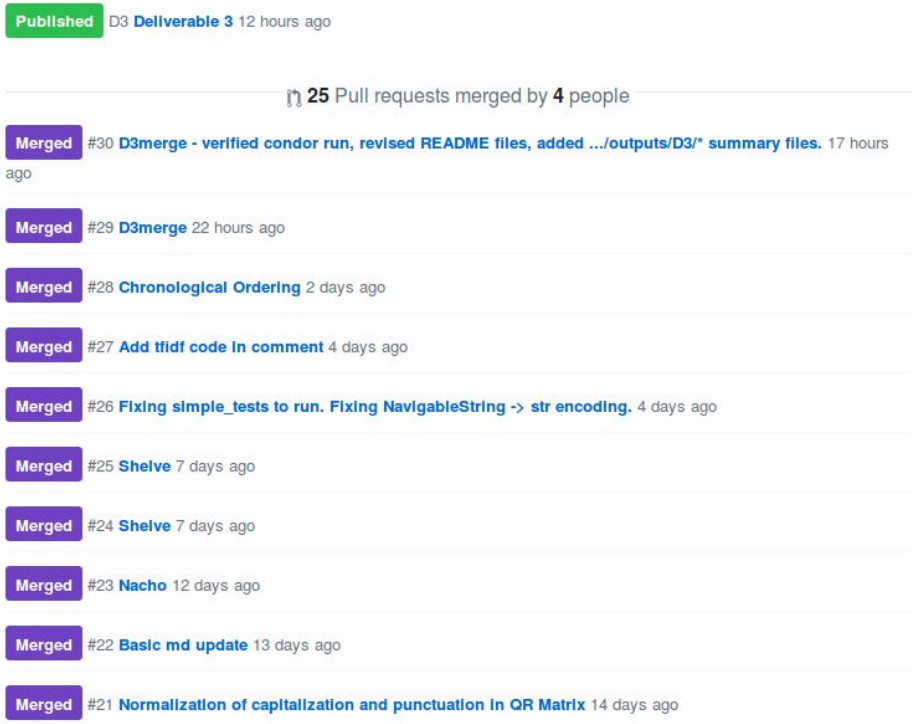
Infrastructure Enhancements

- **Article Serialization and Storage**
  Implemented shelve solution and sped up article reading in general
- **Logging**
  Logging levels controlled by configuration, making it easier to debug the system and track issues.
- **Configuration**
  Created a more sophisticated configuration reader, making it easier to test different features in the same code base.



Git Activity

- Branches for each feature
  - QR Matrix
  - MD Update
  - Shelve
  - TF*IDF
  - Chronological Ordering
- Pull Requests
  - Review Implementation
  - Merge to Master Branch
  - Delete Feature Branch



Content Selection

Complete sentences

- Solution to problems with `nltk.sent_tokenize()`

Token Normalization

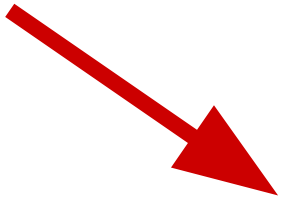
- Removal of tokens with no alphanumerics
- `token.lower()`

R-1 +0.02878

Content Selection - *Topic-focused Summarization*

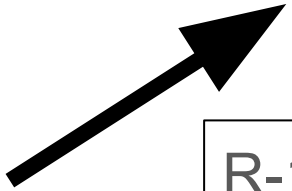
TF • IDF

ROUGE-1	0.15607
ROUGE-2	0.03121
ROUGE-4	0.00295



Count-based Weight

ROUGE-1	0.22262
ROUGE-2	0.05187
ROUGE-4	0.00501



R-1 +0.03548

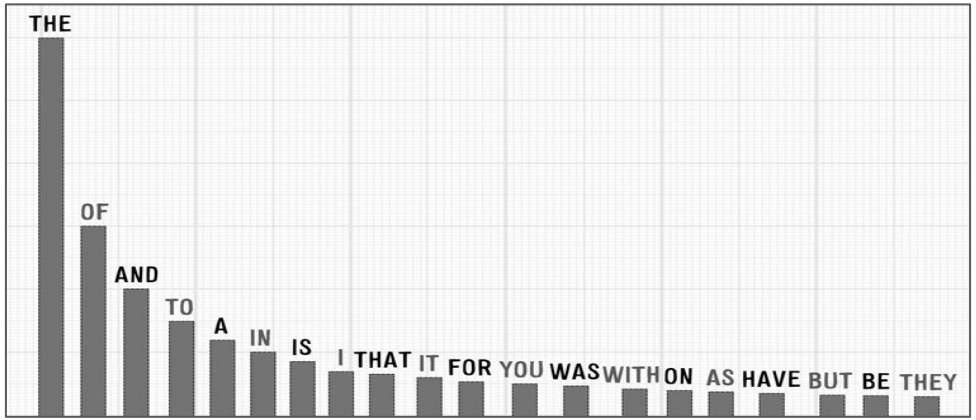
(Hong & Nenkova 2004; Luhn 1957)

Content Selection - *Topic-focused Summarization*

Stop Words

- Solution to problems with QR Matrix (Conroy & O'Leary, 2001)

R-1 +0.0067



Information Ordering

Chronological Expert
(Bollegala et al., 2004)

- First consideration: Date of article
- Second consideration: Position in article

Content Realization

Sentences are written as-is to file with no modification.

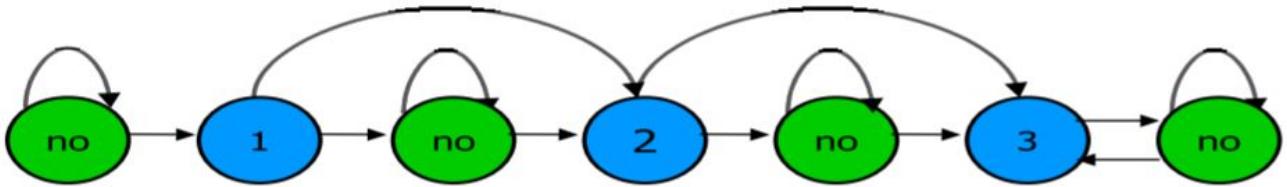
Issues and Successes

HMM content selection and ordering:

"Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition"

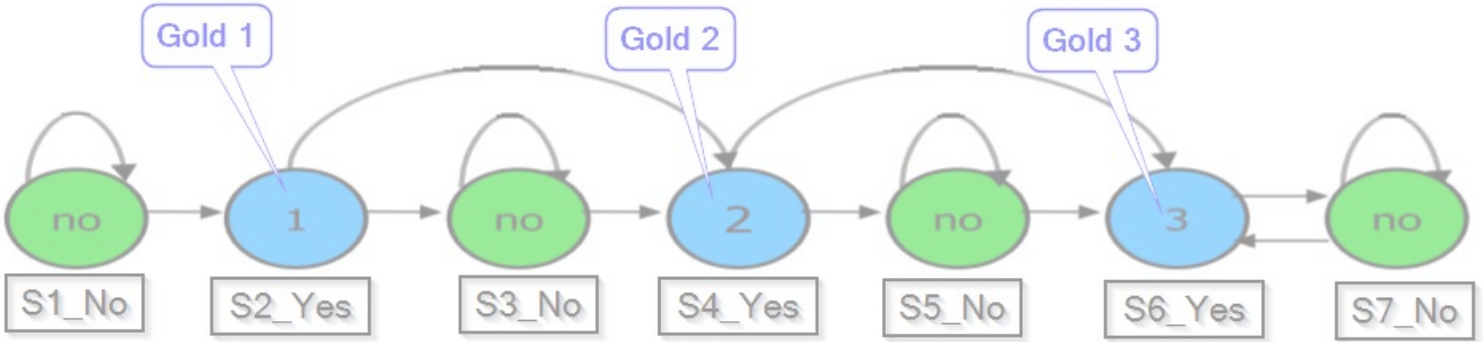
(Conroy & O'Leary, 2001)

> "These parameters are estimated based on training data: for example, **the probability of transition between summary state 2j and summary state 2j + 2 is the number of times summary sentence j + 1 directly followed summary sentence j in the training documents, divided by the number of documents**; and the probability of transition between summary state 2j and non-summary state 2j + 1 is defined to be one minus this probability."

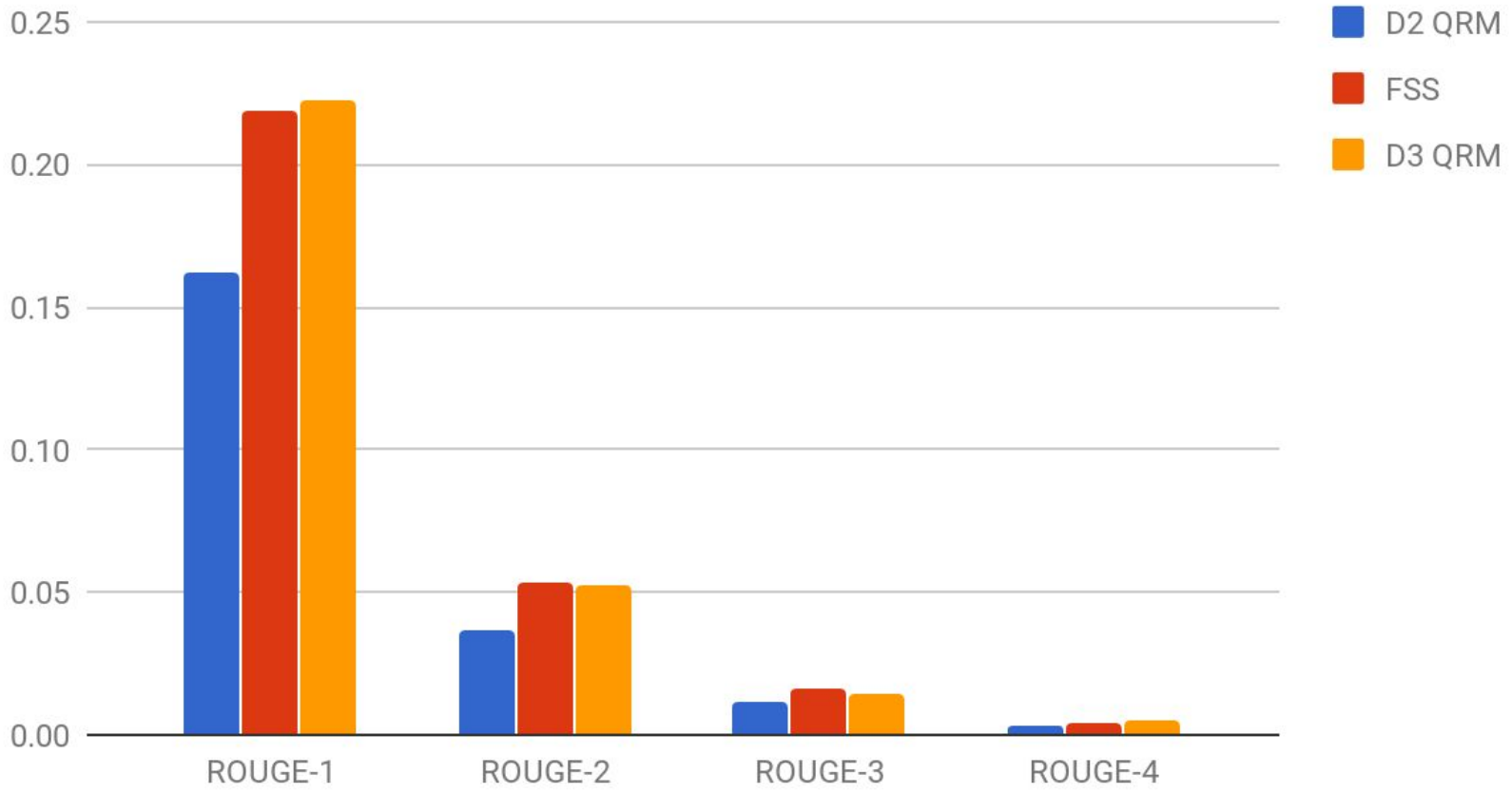


HMM content selection and ordering

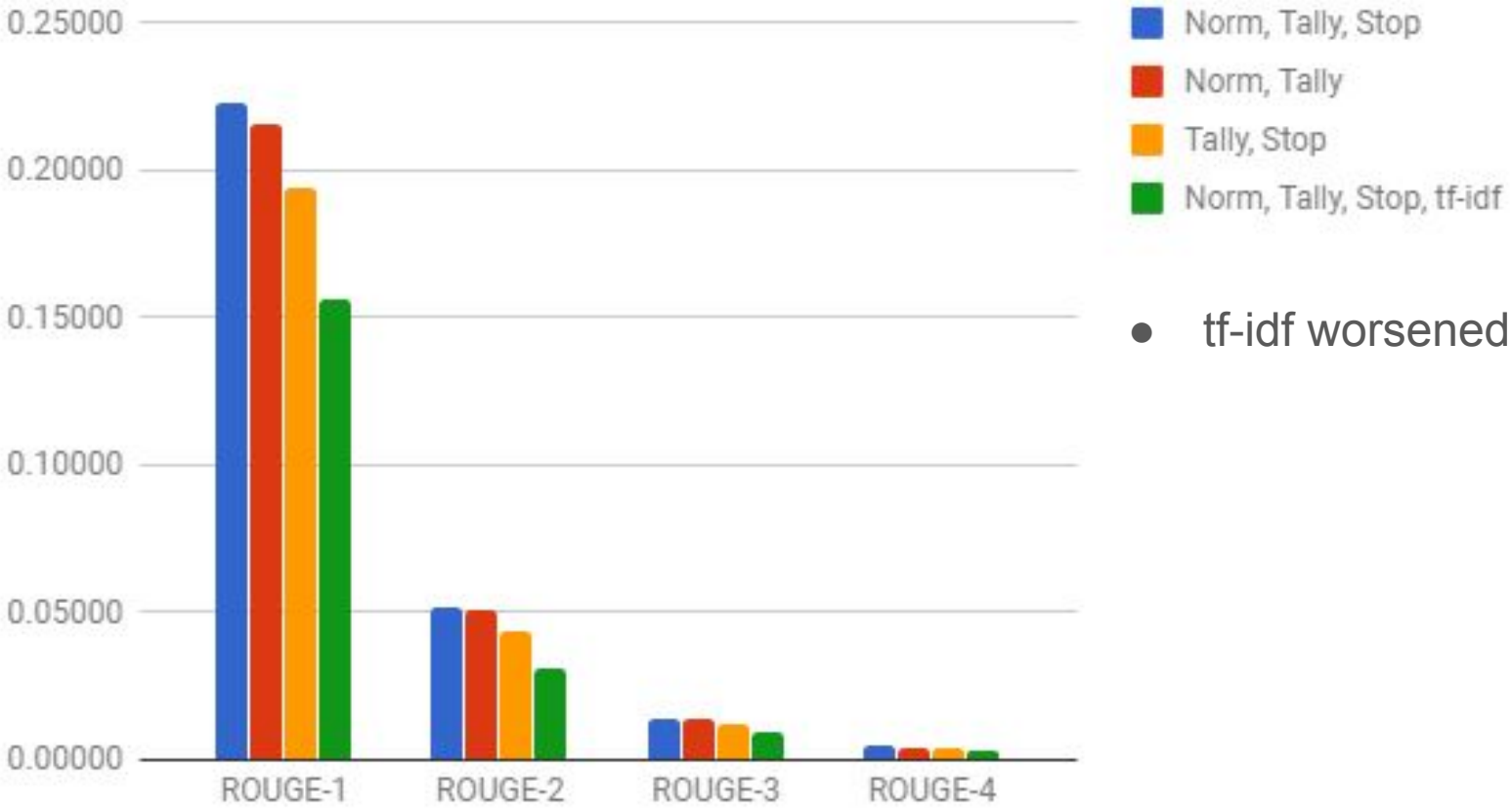
```
01 for RAW in all DOC_SETS:
02     GOLD_SUMMARIES = load_summaries_for_docset( RAW )
03     for g in range( 0, 3 ):
04         j = g+1 # Paper uses one-based indices
05         key = 'S{}_Yes --> S{}_Yes'.format( 2*j, 2*(j+1) )
06         # e.g. for j=1 : "S2_Yes --> S4_Yes"
07         for GOLD in GOLD_SUMMARIES:
08             TOTAL_OBSERVATIONS[ key ] += 1
09             r = find_index_of_sentence( RAW, GOLD[g] )
10             assert GOLD[g] == RAW[r] # safety check.
11             if GOLD[g+1] == RAW[r+1]:
12                 SUCCESSFUL_OBSERVATIONS[ key ] += 1
```



Results: Improvement from D2

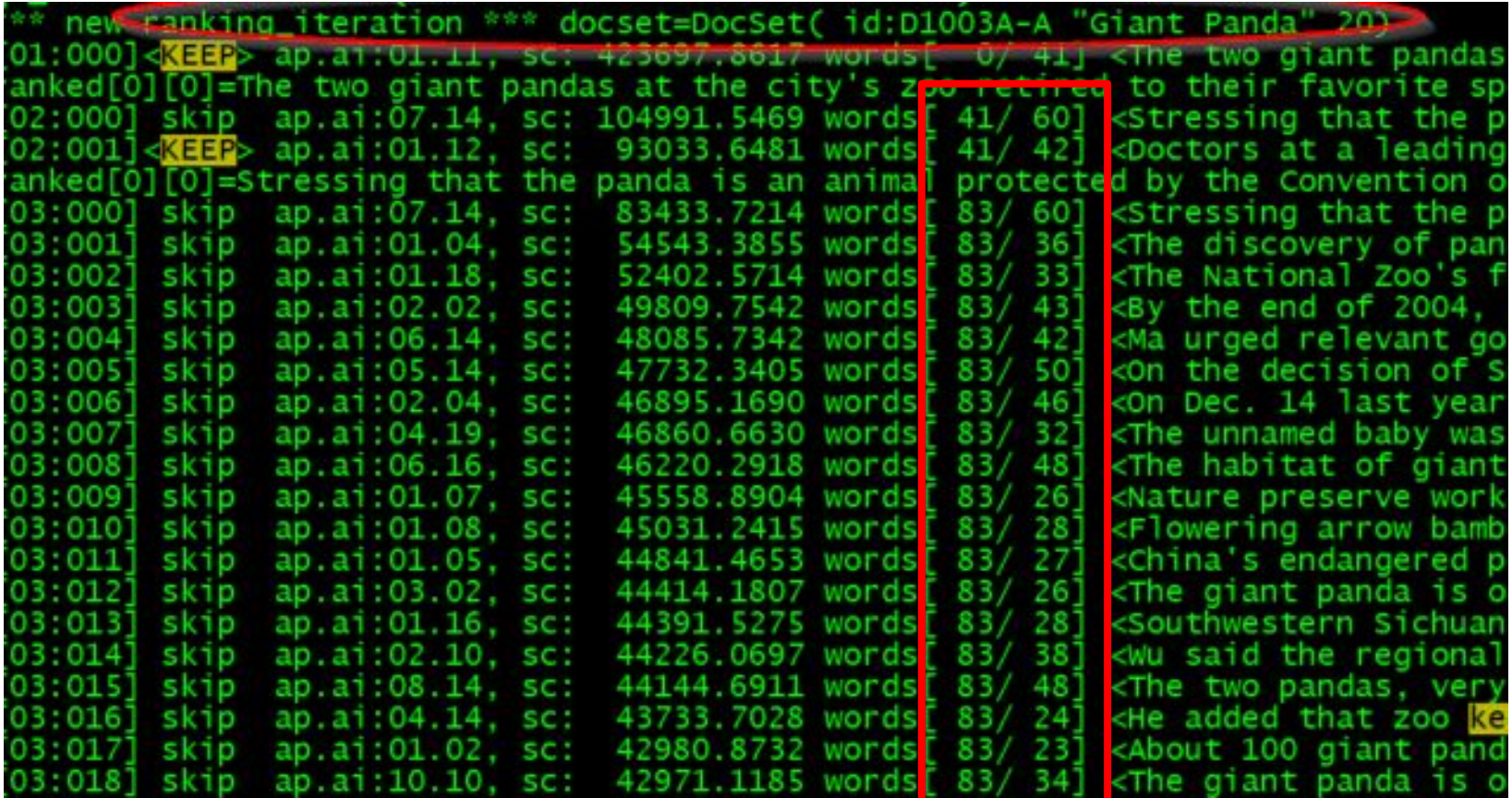


Incremental Improvements and Error Analysis



- tf-idf worsened scores

Error Analysis : Sentence Compression...



Error Analysis : Stop Word metrics...

STOP_WORDS_rev	freq	%ge	cum%	
the	29218	12.91%	12.91%	######
of	13069	5.77%	18.68%	###
to	11765	5.20%	23.88%	###
and	10955	4.84%	28.72%	##
a	10667	4.71%	33.44%	##
in	9528	4.21%	37.65%	##
that	5409	2.39%	40.04%	#
said	5297	2.34%	42.38%	#
's	4405	1.95%	44.32%	#
for	3931	1.74%	46.06%	#
on	3621	1.60%	47.66%	#
is	3443	1.52%	49.18%	#
was	3308	1.46%	50.64%	#
it	2853	1.26%	51.90%	#
with	2694	1.19%	53.09%	#
he	2524	1.12%	54.21%	#
by	2408	1.06%	55.27%	#
as	2384	1.05%	56.33%	#
have	2340	1.03%	57.36%	#
from	2327	1.03%	58.39%	#

Thank you!

Related Reading

A preference learning approach to sentence ordering for multi-document summarization D. Bollegala, N. Okazaki, and M. Ishizuka [2004]

Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition J. Conroy and D. P. O'Leary [2001]

Improving the Estimation of Word Importance for News Multi-Document Summarization K. Hong and N. Nenkova [2014]

*A Statistical Approach to Mechanized Encoding and Searching of Literary Information** H. P. Luhn [1957]

Beautiful Soup Documentation https://www.crummy.com/software/BeautifulSoup/bs4/doc/

NLTK http://www.nltk.org/