

Summarization D3

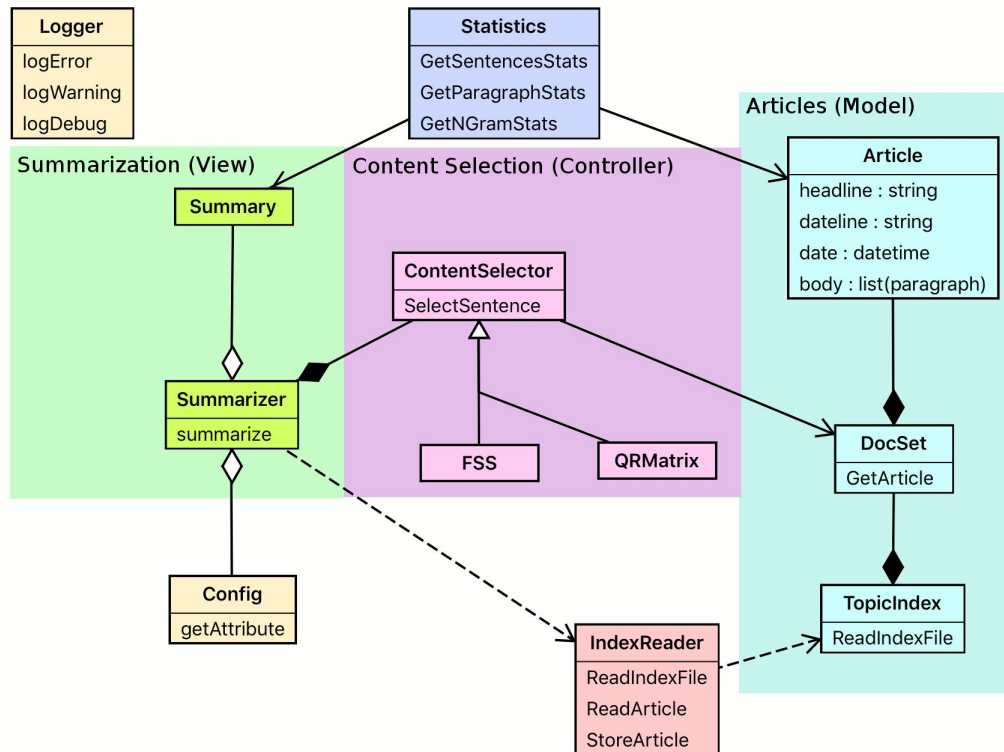
Team E2JK++

Eric Lindberg
John Greve
Josh Mathias
Kekoa Riggins



Infrastructure Enhancements

- **Article Serialization and Storage**
Implemented shelf 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

Published	D3 Deliverable 3	12 hours ago
🔗 25 Pull requests merged by 4 people		
Merged	#30 D3merge - verified condor run, revised README files, added .../outputs/D3/* summary files.	17 hours ago
Merged	#29 D3merge	22 hours ago
Merged	#28 Chronological Ordering	2 days ago
Merged	#27 Add tfidf code in comment	4 days ago
Merged	#26 Fixing simple_tests to run. Fixing NavigableString -> str encoding.	4 days ago
Merged	#25 Shelve	7 days ago
Merged	#24 Shelve	7 days ago
Merged	#23 Nacho	12 days ago
Merged	#22 Basic md update	13 days ago
Merged	#21 Normalization of capitalization and punctuation in QR Matrix	14 days ago

Content Selection

Complete sentences

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

Token Normalization

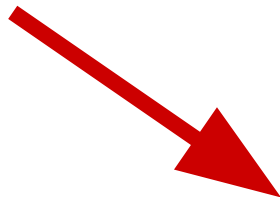
- Removal of tokens with no alphanumerics
- `token.lower()`

R-1 +0.02878

Content Selection - *Topic-focused Summarization*

TF • IDF

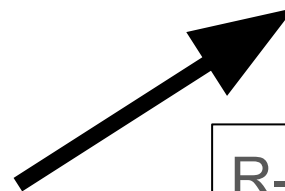
ROUGE-1	0.15607
ROUGE-2	0.03121
ROUGE-4	0.00295



(Hong & Nenkova 2004; Luhn 1957)

Count-based Weight

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



R-1 +0.03548

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.



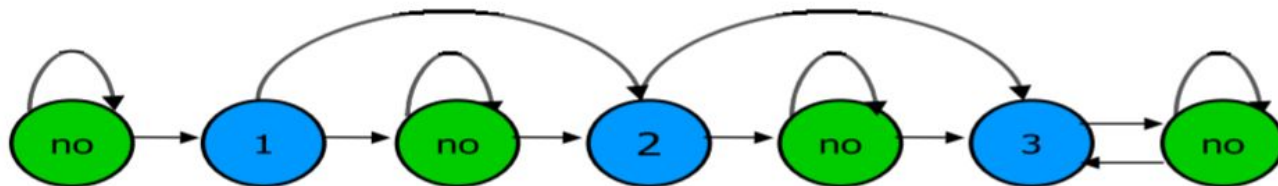
Issues and Successes

HMM content selection and ordering:

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

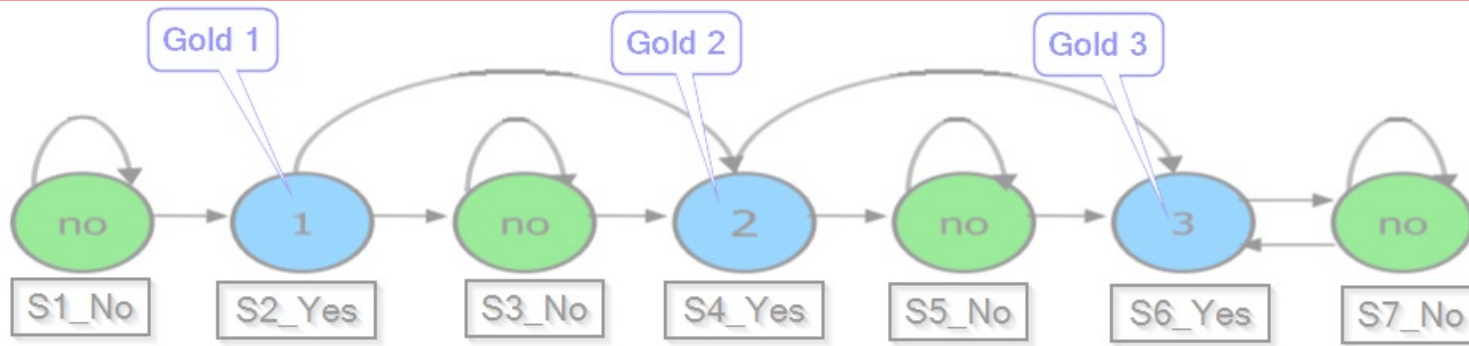
(Conroy & O'Leary, 2001)

"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."

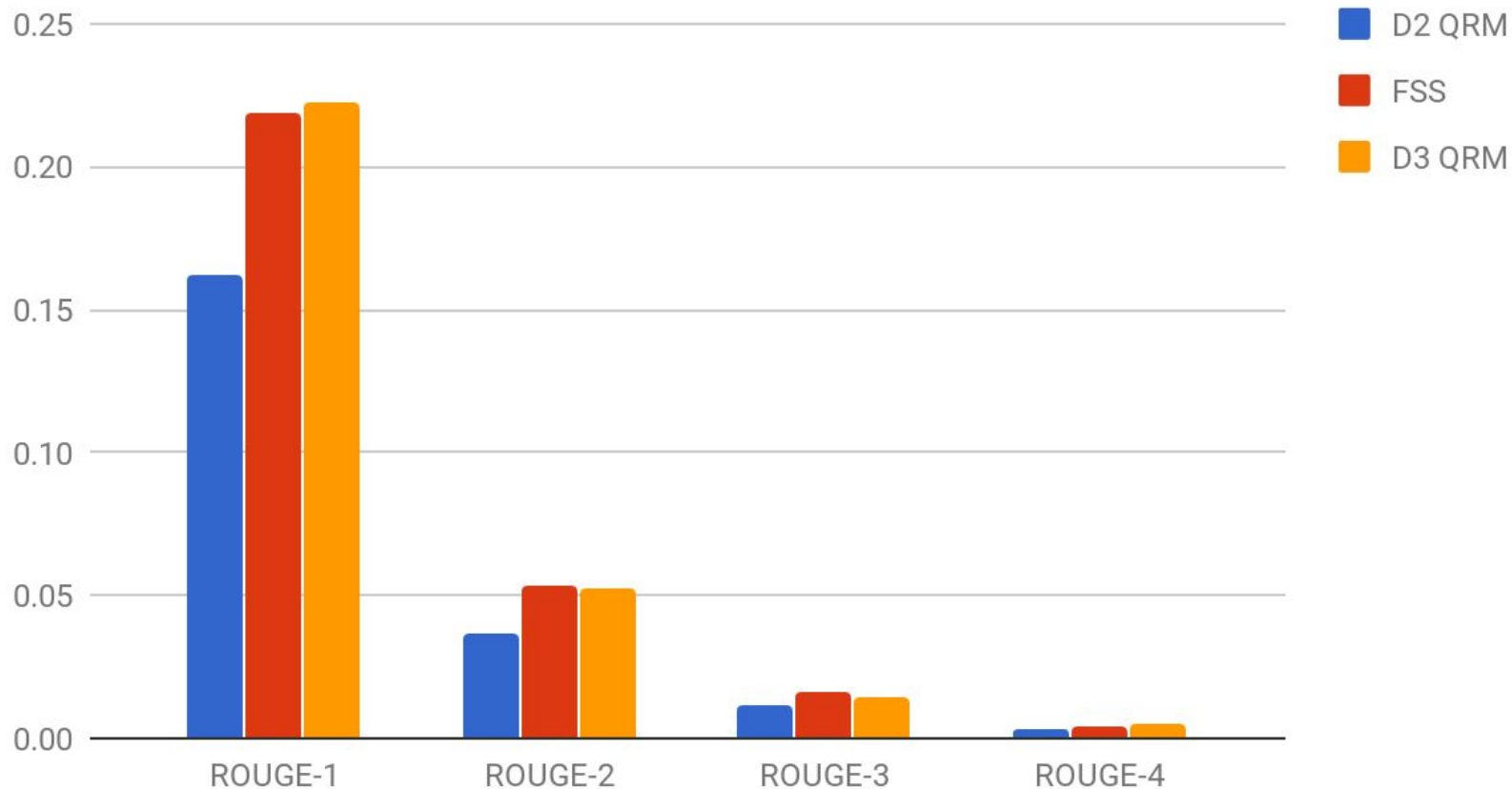


HMM content selection and ordering

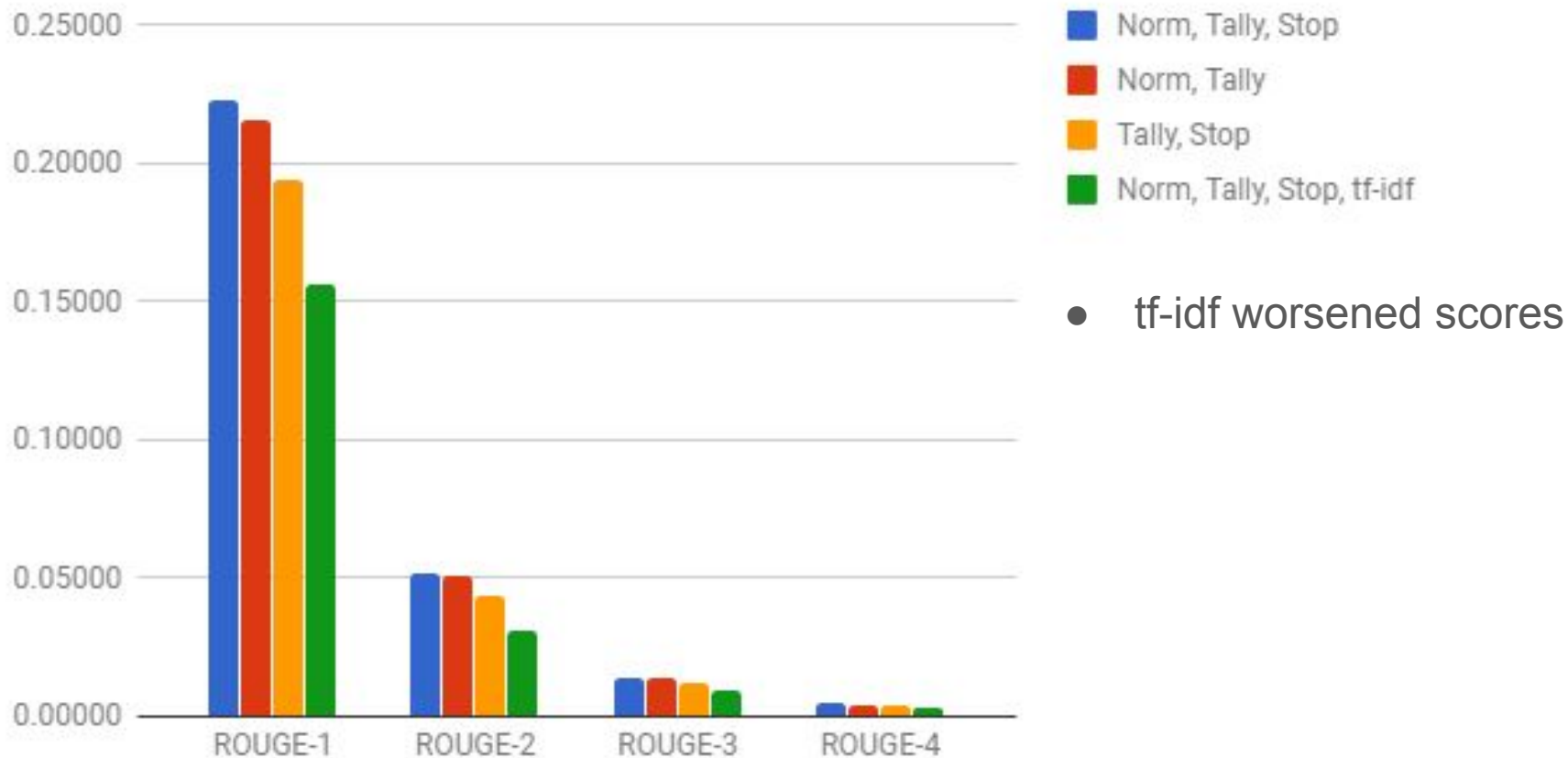
```
01 for RAW in all DOC_SETS:
02     GOLD_SUMMARIES = load_summaries_for_docset( RAW )
03     for g in range( 0, 3 ):
04         j = g+1 # Paper uses one-based indices
05         key = 'S{ }_Yes --> S{ }_Yes'.format( 2*j, 2*(j+1) )
06         # e.g. for j=1 : "S2_Yes --> S4_Yes"
07         for GOLD 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



Error Analysis : Sentence Compression...

```
** new ranking iteration *** docset=DocSet( id:D1003A-A "Giant Panda" 20)
01:000] <KEEP> ap.ai:01.11, sc: 423697.8617 words[ 0/ 41] <The two giant pandas
anked[0][0]=The two giant pandas at the city's zoo retired to their favorite sp
02:000] skip ap.ai:07.14, sc: 104991.5469 words[ 41/ 60] <Stressing that the p
02:001] <KEEP> ap.ai:01.12, sc: 93033.6481 words[ 41/ 42] <Doctors at a leading
anked[0][0]=Stressing that the panda is an animal protected by the Convention o
03:000] skip ap.ai:07.14, sc: 83433.7214 words[ 83/ 60] <Stressing that the p
03:001] skip ap.ai:01.04, sc: 54543.3855 words[ 83/ 36] <The discovery of pan
03:002] skip ap.ai:01.18, sc: 52402.5714 words[ 83/ 33] <The National Zoo's f
03:003] skip ap.ai:02.02, sc: 49809.7542 words[ 83/ 43] <By the end of 2004,
03:004] skip ap.ai:06.14, sc: 48085.7342 words[ 83/ 42] <Ma urged relevant go
03:005] skip ap.ai:05.14, sc: 47732.3405 words[ 83/ 50] <On the decision of S
03:006] skip ap.ai:02.04, sc: 46895.1690 words[ 83/ 46] <On Dec. 14 last year
03:007] skip ap.ai:04.19, sc: 46860.6630 words[ 83/ 32] <The unnamed baby was
03:008] skip ap.ai:06.16, sc: 46220.2918 words[ 83/ 48] <The habitat of giant
03:009] skip ap.ai:01.07, sc: 45558.8904 words[ 83/ 26] <Nature preserve work
03:010] skip ap.ai:01.08, sc: 45031.2415 words[ 83/ 28] <Flowering arrow bamb
03:011] skip ap.ai:01.05, sc: 44841.4653 words[ 83/ 27] <China's endangered p
03:012] skip ap.ai:03.02, sc: 44414.1807 words[ 83/ 26] <The giant panda is o
03:013] skip ap.ai:01.16, sc: 44391.5275 words[ 83/ 28] <Southwestern Sichuan
03:014] skip ap.ai:02.10, sc: 44226.0697 words[ 83/ 38] <Wu said the regional
03:015] skip ap.ai:08.14, sc: 44144.6911 words[ 83/ 48] <The two pandas, very
03:016] skip ap.ai:04.14, sc: 43733.7028 words[ 83/ 24] <He added that zoo ke
03:017] skip ap.ai:01.02, sc: 42980.8732 words[ 83/ 23] <About 100 giant pand
03:018] skip ap.ai:10.10, sc: 42971.1185 words[ 83/ 34] <The giant panda is o
```

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/>