A framework to aggregate news, show relationships and eliminate editorial bias

http://wordsinmedia.com

Keerat Sharma
ksharm2@pride.hofstra.edu
keerat@gmail.com

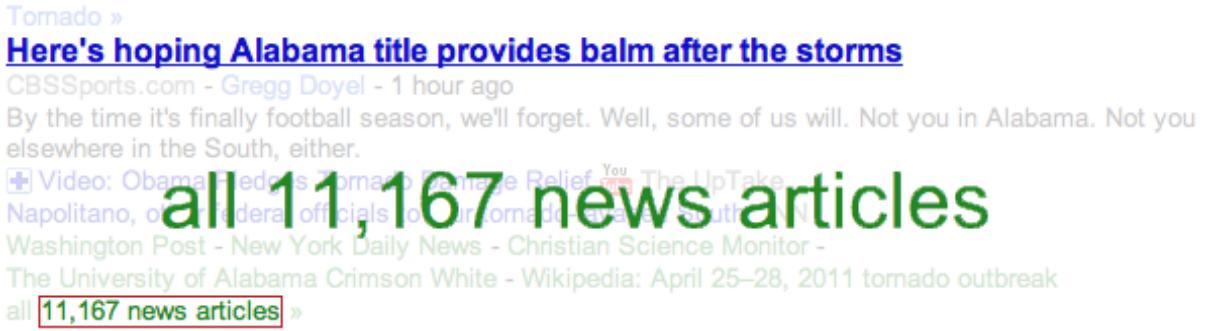
We need news

"The basis of our government being the opinion of the people, the very first object should be to keep that right; and were it left to me to decide whether we should have *a government without newspapers, or newspapers without a government, I should not hesitate a moment to prefer the latter.*"

Thomas Jefferson,16 January 1787


Los Angeles Times
THE WALL STREET JOURNAL
TUESDAY, JULY 31, 2007 - VOL. CCL NO. 25
Do Bats Kill?
The New Hunt For Business—
COMMODITY KING
Snowman Video
DR. EVIL'S PLAN TO STOP GLOBAL WARMING
Rolling Stone
HIP-HOP REPORT
JAY-Z
NAS
DIDDY
YOUNG JEEZY
TUPAC
Jon Stewart & Stephen Colbert
AMERICA'S ANCHORS
By Maureen Dowd
THE WHO RETURN!
BORAT COMEDY OF THE YEAR
SPECIAL COMMENT
LIVE
The Washington Post
Justices To Rule On D.C. Gun Ban
2nd Amendment Case Could Affect Laws Nationwide
Advance May End Stem Cell Debate
Huckabee Gaining Ground in Iowa
FOX NEWS channel
AT LARGE
The New York Times
Nation Sets Its Sights on Building Same Economy
IRAQ WAR ENDS
NEW YORK POST

Problems



- Lots of content 24/7
- Too much to process
- Editorialized variations

Existing solutions

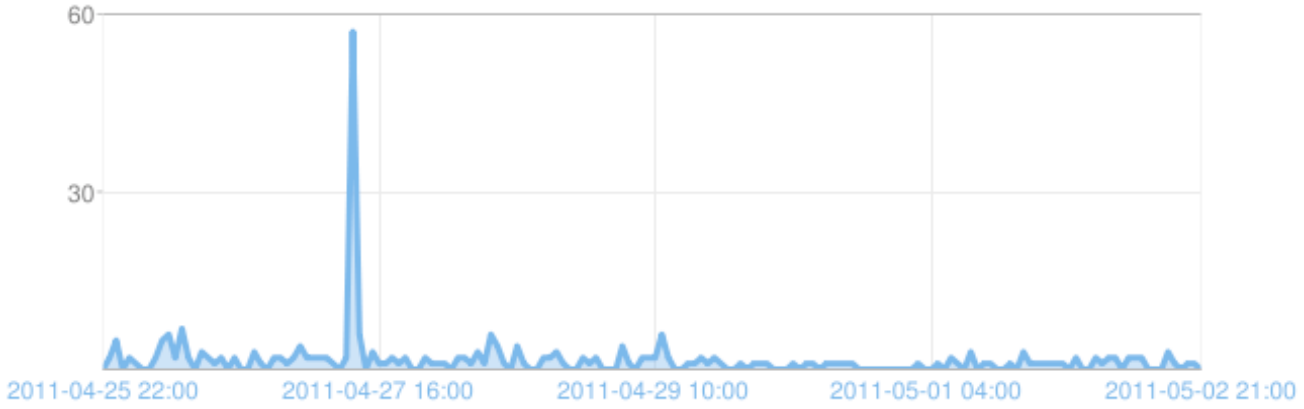
- Read everything
- You pick a few “trusted” sources
- Google News et al
- wikipedia
- why can’t we aggregate news like logs/ stocks/bank balances etc?

A framework to extract facts news etc..

- Acquiring News
- Analyzing Text
- Database
- Reporting
- Lessons and Future Research

Structured Data

www.wordsinmedia.com:80 220.181.108.179 - - [02/May/2011:02:23:48 +0000] "GET / HTTP/1.1" 200 4890 "-" "Baiduspider+(+http://www.baidu.com/search/spider.htm)"
www.wordsinmedia.com:80 123.125.71.96 - - [02/May/2011:03:29:01 +0000] "GET / HTTP/1.1" 200 4890 "-" "Baiduspider+(+http://www.baidu.com/search/spider.htm)"
www.wordsinmedia.com:80 220.181.108.177 - - [02/May/2011:03:29:08 +0000] "GET / HTTP/1.1" 200 4890 "-" "Baiduspider+(+http://www.baidu.com/search/spider.htm)"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:40:58 +0000] "POST /w/arpc HTTP/1.1" 200 1320 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:40:58 +0000] "POST /w/arpc HTTP/1.1" 200 4568 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:40:59 +0000] "POST /w/arpc HTTP/1.1" 200 784 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:41:07 +0000] "POST /w/arpc HTTP/1.1" 200 678 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:41:08 +0000] "POST /w/arpc HTTP/1.1" 200 702 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:41:08 +0000] "POST /w/arpc HTTP/1.1" 200 627 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:41:08 +0000] "POST /w/arpc HTTP/1.1" 200 499 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:41:09 +0000] "POST /w/arpc HTTP/1.1" 200 459 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"
www.wordsinmedia.com:80 98.247.205.231 - - [02/May/2011:03:41:11 +0000] "POST /w/arpc HTTP/1.1" 200 458 "http://www.wordsinmedia.com/w/F81EA09428A27265D79446B37BDEF8BD.cache.html" "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; en-US) AppleWebKit/534.16 (KHTML, like Gecko) Chrome/10.0.648.205 Safari/534.16"



Data Structure for News

- Acquire Content
- Sentences
- Link Grammar
- Article Histogram

RSS

- Created for humans
- Advertisements
- Navigational elements
- Images

Raw RSS

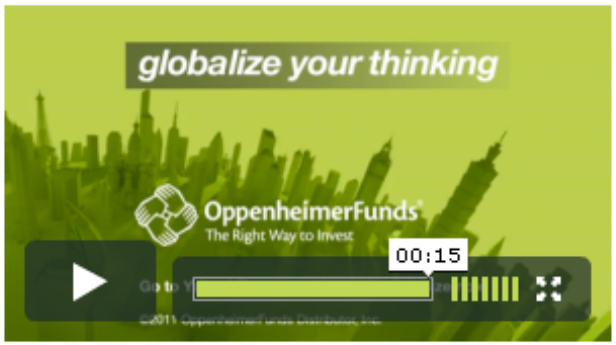

```
<title>U.S. Nuclear Industry Faces New Uncertainty</title>
<description>A fragile bipartisan consensus on nuclear
power&#x2019;s promise for the United States may have
dissolved.&lt;br clear=&quot;both&quot; style=&quot;clear:
both;&quot;/&gt;&lt;br clear=&quot;both&quot; style=&quot;clear:
both;&quot;/&gt;&lt;a href=&quot;http://ads.pheedo.com/
click.phdo?s=0159e1d628b59e2ae52f62785ae17898&amp;p=1&quot;
&gt;&lt;img alt=&quot;&quot; style=&quot;border: 0;&quot;
border=&quot;0&quot; src=&quot;http://ads.pheedo.com/img.phdo?
s=0159e1d628b59e2ae52f698 &amp;p=1&quot;/&gt;&lt;/a&gt; </
description>
```

HTML Client

U.S. Nuclear Industry Faces New Uncertainty

(Mon, 14 Mar 2011 03:24:16 GMT)

A fragile bipartisan consensus on nuclear power's promise for the United States may have dissolved.



Unescape HTML

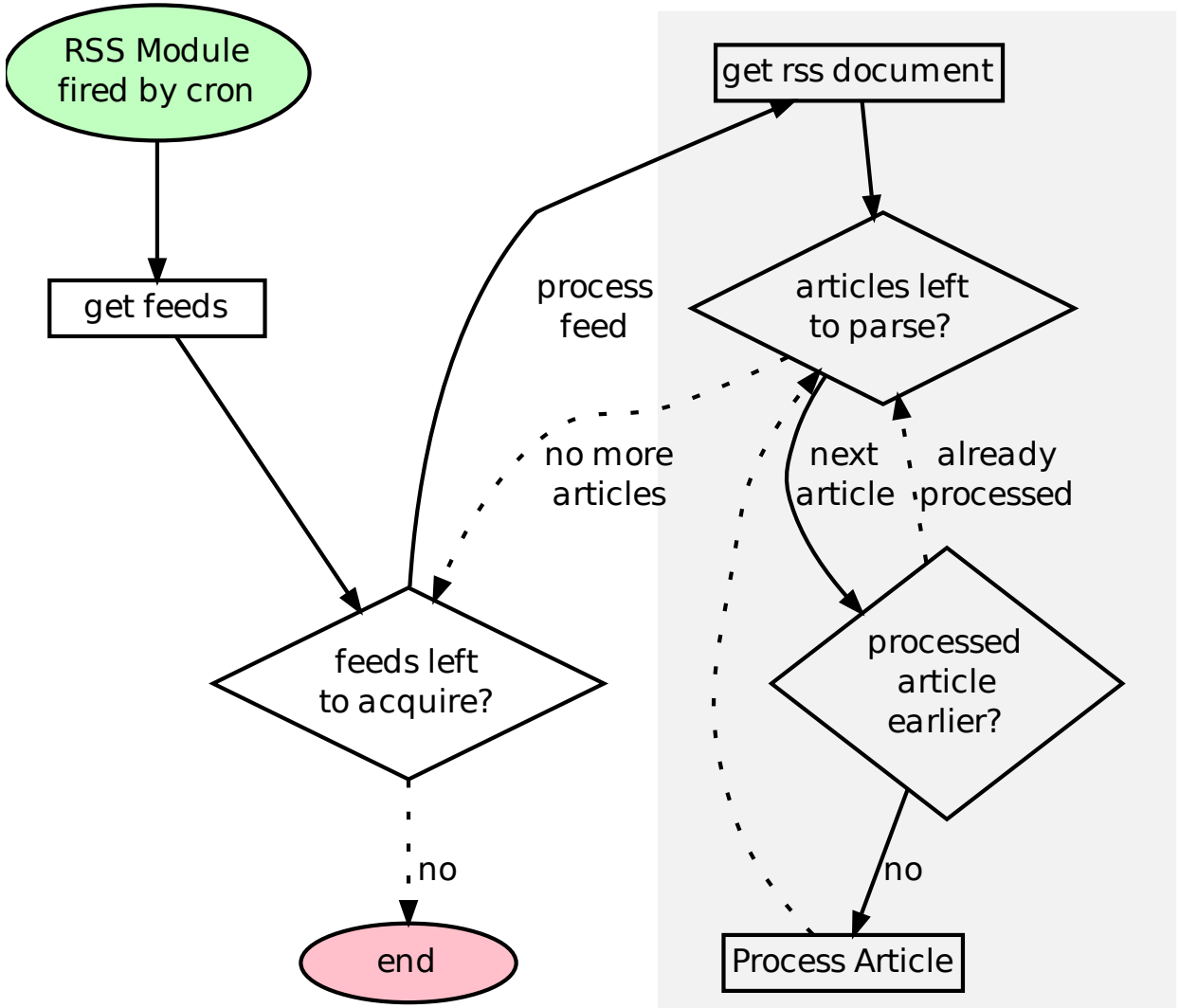
```
A fragile bipartisan consensus on nuclear power’s promise for
the United States may have dissolved.<br clear="both"
style="clear: both;"/><br clear="both" style="clear: both;"/><a
href="http://ads.pheedo.com/click.phdo?
s=0159e1d628b59e2ae52f62785ae17898&p=1" ><img alt=""
style="border: 0;" border="0" src="http://ads.pheedo.com/
img.phdo?s=0159e1d628b598 &p=1"/></a>
```

Artifact Removal

“A fragile bipartisan consensus on nuclear power’s promise for the United States may have dissolved. ”


```
<title>U.S. Nuclear Industry Faces New Uncertainty</title>
<description>A fragile bipartisan consensus on nuclear
power&#x2019;s promise for the United States may have
dissolved.&lt;br clear=&quot;both&quot; style=&quot;clear:
both;&quot;/&gt;&lt;br clear=&quot;both&quot; style=&quot;clear:
both;&quot;/&gt;&lt;a href=&quot;http://ads.pheedo.com/
click.phdo?s=0159e1d628b59e2ae52f62785ae17898&amp;p=1&quot;
&gt;&lt;img alt=&quot;&quot; style=&quot;border: 0;&quot;
border=&quot;0&quot; src=&quot;http://ads.pheedo.com/img.phdo?
s=0159e1d628b59e2ae52f698 &amp;p=1&quot;/&gt;&lt;/a&gt; </
description>
```

Systematize



Processing Articles

- Ingest text (article)
- Do something complex (parse)
- Emit something we can aggregate

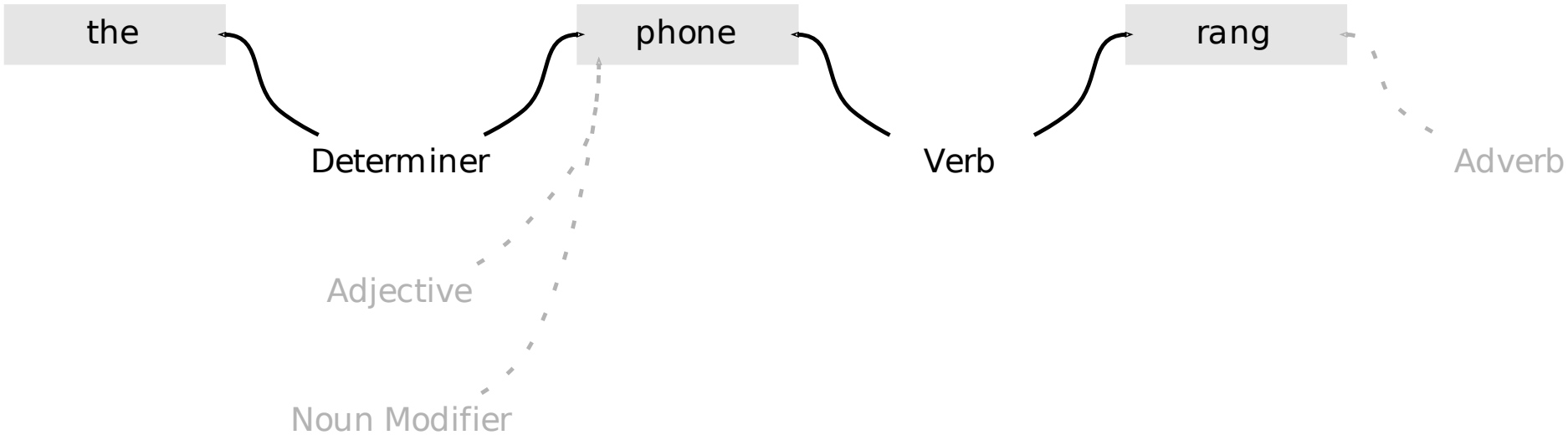
Link Grammar

- Sentences are graphs
- Each word is a node
- Words have a set of in/out edge types
- w_1(out) == w_2(in)
- Dictionary for words and edge types

Example Dictionary

Word	Inbound	Outbound
loudly	Adverb	
phone	Adjective, Determiner, Noun Modifier	Verb
rang	Verb	Adverb
yellow	Noun	Adjective
the		Determiner

Legal Sentence

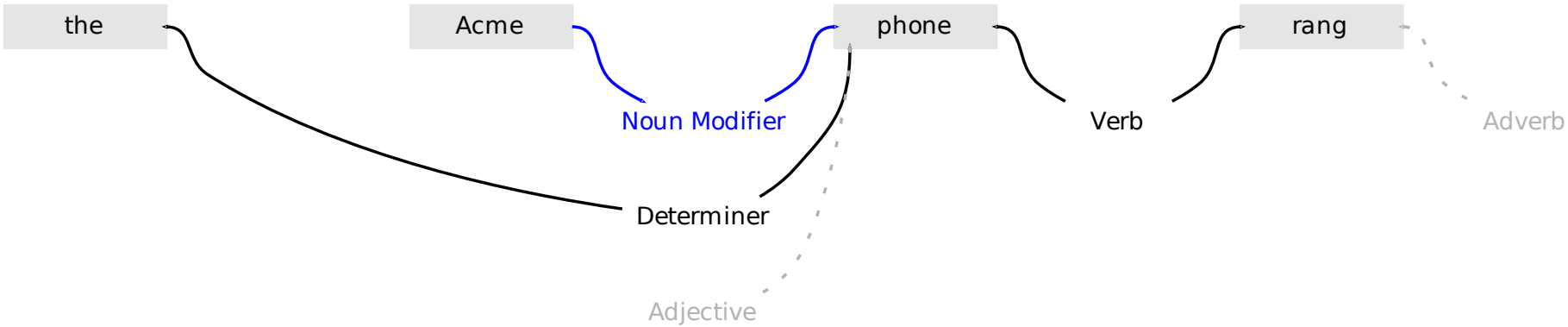


Word	Inbound	Outbound
loudly	Adverb	
phone	Adjective, Determiner, Noun Modifier	Verb
rang	Verb	Adverb
yellow	Noun	Adjective
the		Determiner

Illegal Sentence

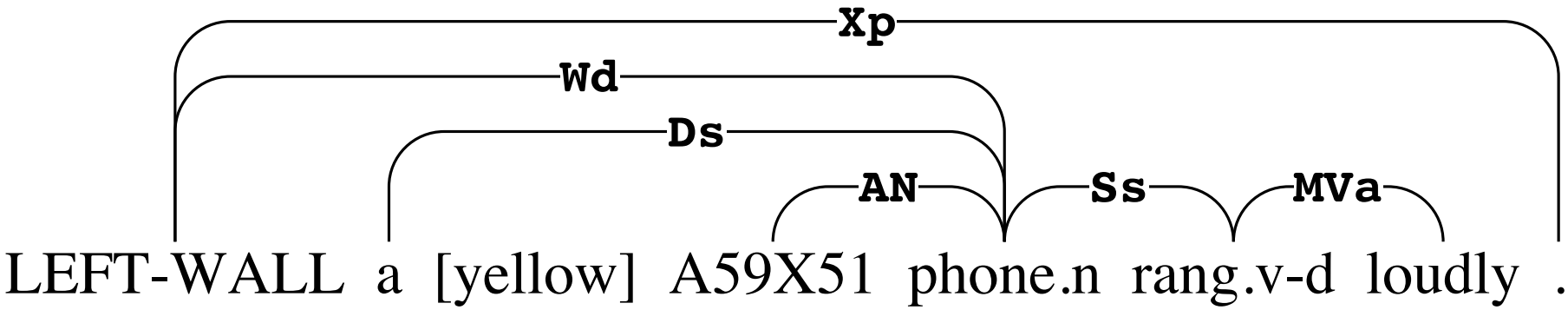
Word	Inbound	Outbound
loudly	Adverb	
phone	Adjective, Determiner, Noun Modifier	Verb
rang	Verb	Adverb
yellow	Noun	Adjective
the		Determiner

Intelligent Edges



Word	Inbound	Outbound
loudly	Adverb	
phone	Adjective, Determiner, Noun Modifier	Verb
rang	Verb	Adverb
yellow	Noun	Adjective
the		Determiner

Word contextualization



foreach sentence

- Link parser
- Extract words of interest
- Stem words

Stemming

Original Word	Stemmed Word
dismiss	dismiss
dismissed	dismiss
dismissal	dismiss
dismisses	dismiss
dismissing	dismiss

Porter Stemming

```
# Step 1c
if ($w =~ /y$/) { $stem = $`; if ($stem =~ /$_v/o) { $w = $stem."i"; } }

# Step 2
if ($w =~ /(ational|tional|enci|anci|izer|bli|alli|entli|eli|ousli|ization
       |ation|ator|alism|iveness|fulness|ousness|aliti|iviti|biliti|logi)$/)
{ $stem = $`; $suffix = $1;
  if ($stem =~ /$mgr0/o) { $w = $stem . $step2list{$suffix}; }
}

# Step 3
if ($w =~ /(icate|ative|alize|iciti|ical|ful|ness)$/)
{ $stem = $`; $suffix = $1;
  if ($stem =~ /$mgr0/o) { $w = $stem . $step3list{$suffix}; }
}
```

Example Article

The negotiations on weapon management are failing. This is concerning, as failure to keep the parties at the negotiating table will result in illicit weapons on the black market.

Link Parse and Extract

The negotiations on weapon management are failing:

```
[negotiations:negotiations.n, weapon:weapon.n,
management:management.n-u, failing:failing.g]
```

This is concerning, as failure to keep the parties at the negotiating table will result in illicit weapons on the black market:

```
[concerning:concerning.g, failure:failure.n,
parties:parties.n, negotiating:negotiating.g,
table:table.n, illicit:illicit.a, weapons:weapons.n,
black:black.a]
```

Stem and Aggregate

Word as seen in sentence	Stemmed word
negotiations	negoti
weapon	weapon
management	manag
failing	fail
concerning	concern
failure	failur
parties	parti
negotiating	negoti
table	tabl
illicit	illicit
weapons	weapon
black	black

Article Histogram

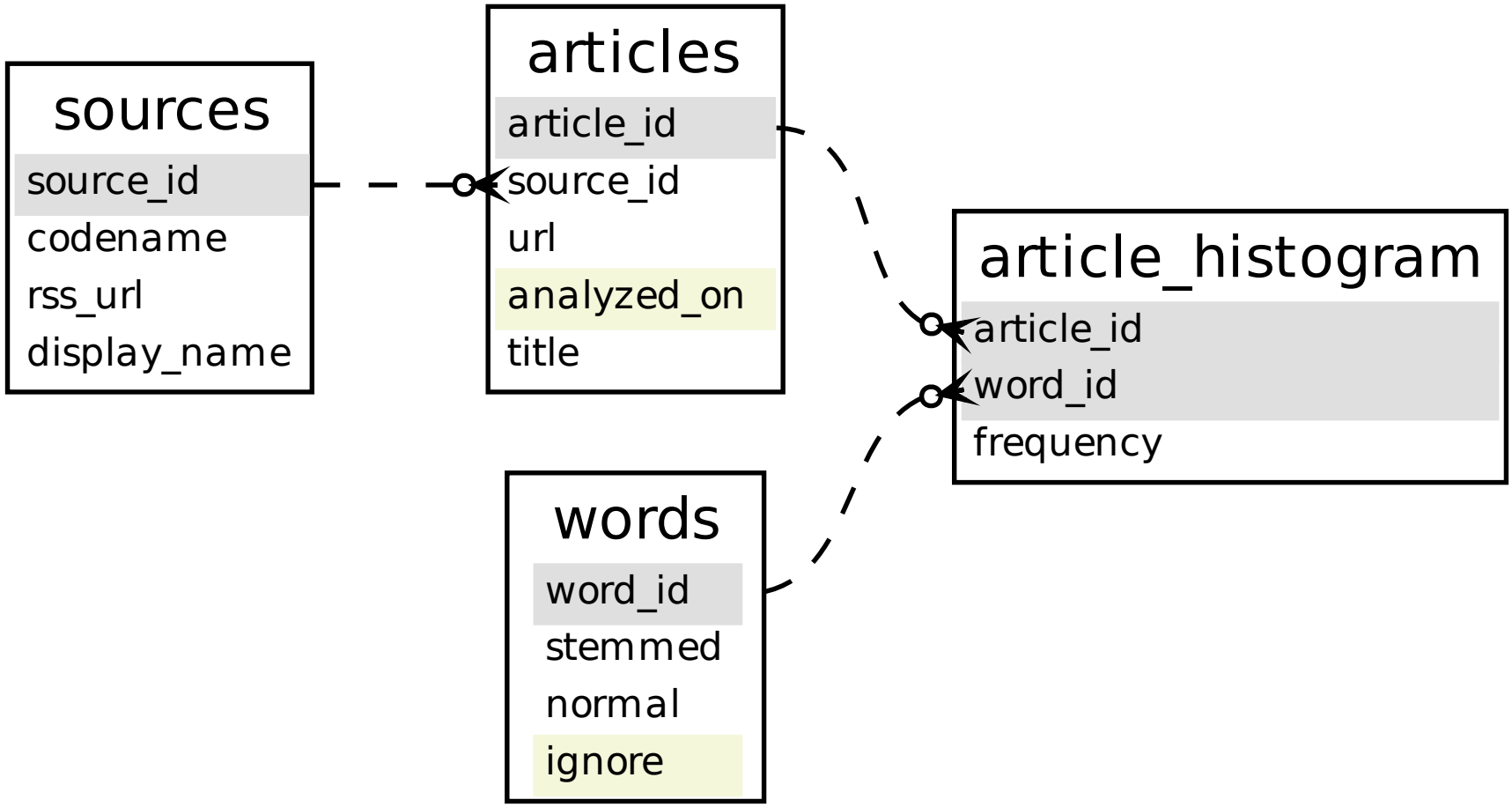
The negotiations on weapon management are failing. This is concerning, as failure to keep the parties at the negotiating table will result in illicit weapons on the black market.

Stemmed Word	Frequency	Occurred in Article as
negoti	2	negotiating negotiations
weapon	2	weapon weapons
black	1	black
concern	1	concerning
fail	1	failing
failur	1	failure
illicit	1	illicit
manag	1	management
parti	1	parties
tabl	1	table

Article Histogram

- Timestamp
- Stemmed words and their frequencies
- Stemmed words and “normal” versions

Database


sources
source_id
codename
rss_url
display_name
articles
article_id
source_id
url
analyzed_on
title
article_histogram
article_id
word_id
frequency
words
word_id
stemmed
normal
ignore

What's prominent?

```
select w.word_id, w.normal, sum(wh.frequency) as count
  from words w, articles a, word_histogram wh
  where w.ignore = 0 and
  a.analyzed_on > '2011-03-01' and
  a.analyzed_on < '2011-03-31' and
  wh.article_id = a.article_id and
  wh.word_id = w.word_id
  group by wh.word_id order by count desc limit 10;
```

word id	normal	count
1156	Japan	1955
4625	Libya	1179
2117	nuclear	980
267	power	620
688	plant	602
4426	earthquake	509
103	forces	456
1204	Japanese	419
3883	reactor	412
686	protest	360

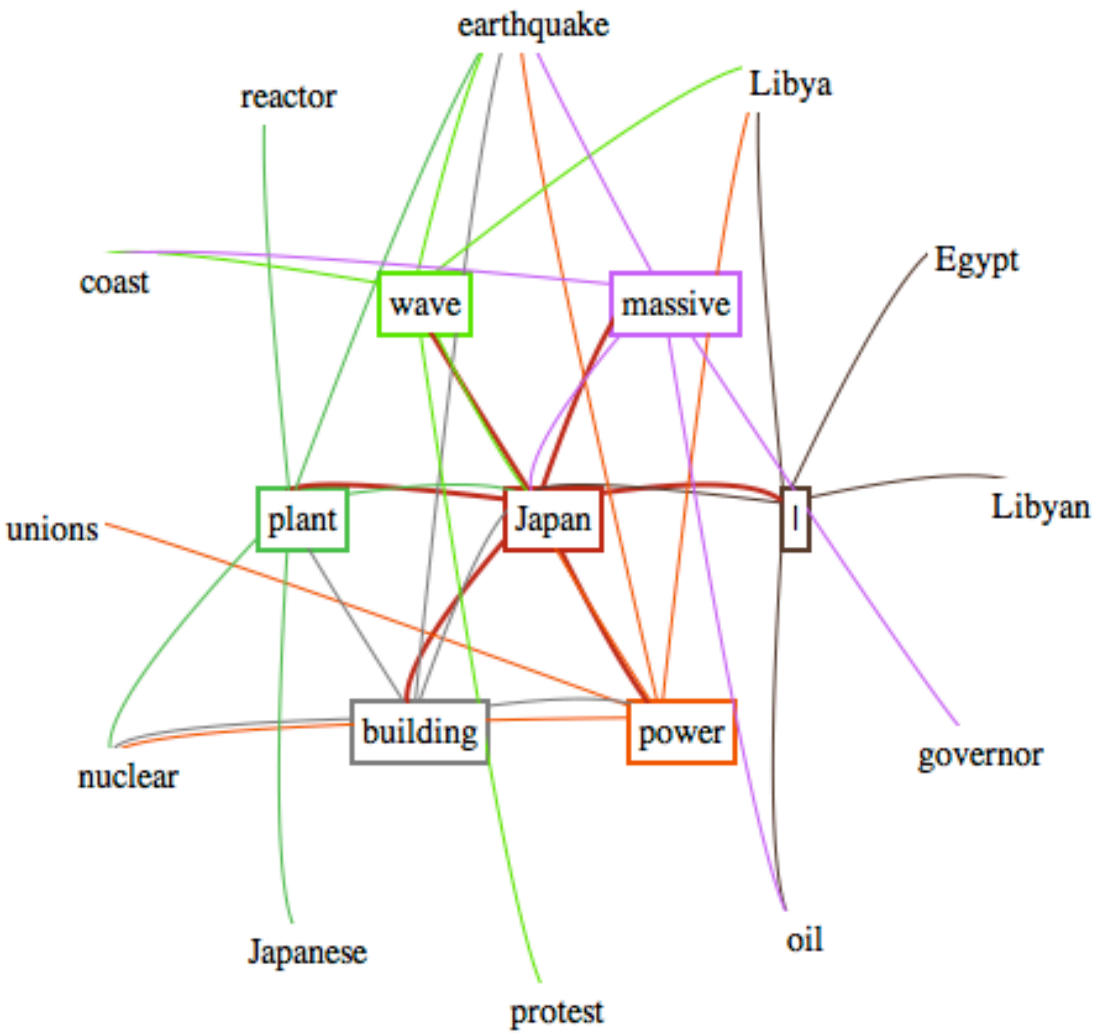
Relationships

```
select w.word_id, w.normal, sum(wh2.frequency) wcount
  from words w, word_histogram wh, word_histogram wh2, articles a
  where wh.word_id = 1156 and
  wh.article_id = a.article_id
  and wh2.article_id = wh.article_id
  and w.word_id = wh2.word_id
  and w.ignore = 0 and wh2.word_id != wh.word_id
  and a.analyzed_on > '2011-03-01'
  and a.analyzed_on < '2011-03-31'
  group by w.word_id order by 3 desc limit 10;
```

word id	normal	count
2117	nuclear	753
688	plant	426
4426	earthquake	415
3883	reactor	335
267	power	266
1204	Japanese	257
905	radiation	232
455	crisis	181
463	disaster	181
70	safety	113

“Japan” mid-March

Analyzing: **Japan**



Related trends:

From 2011-03-11 to 2011-03-11:
northern

From 2011-03-11 to 2011-03-12:
nuclear plant massive quake devastating power | 8.9-Magnitude reactor

From 2011-02-27 to 2011-03-12:
wave

From 2011-03-09 to 2011-03-12:
earthquake coast magnitude damaging building northeastern warning

From 2011-02-14 to 2011-03-12:
Japanese

By Source

From “The Wall Street Journal”		
word id	normal	count
2117	nuclear	110
688	plant	65
4426	earthquake	63
455	crisis	31
649	relief	28
3883	reactor	26
3328	damaging	25
6426	humanitarian	25
1204	Japanese	25
905	radiation	23
From “Associated Press”		
word id	normal	count
2117	nuclear	20
455	crisis	12
4426	earthquake	12
905	radiation	11
688	plant	6
463	disaster	5
1320	amount	5
1675	agency	4
275	system	4
551	threat	4

```
select w.word_id, w.normal, sum(wh2.frequency) wcount
  from words w, word_histogram wh, word_histogram wh2, articles a, sources s
  where wh.word_id = 1156 and
  wh.article_id = a.article_id
  and wh2.article_id = wh.article_id
  and w.word_id = wh2.word_id
  and w.ignore = 0 and wh2.word_id != wh.word_id
  and a.analyzed_on > '2011-03-01'
  and a.analyzed_on < '2011-03-31'
  and a.source_id = s.source_id
  and s.source_id = ?
  group by w.word_id order by 3 desc limit 10;
```

Takeaways

- Acquisition is getting harder (rss, paywalls)
- Sentence Fracturing is complex
- Scalability
- Schema: linkages, sources
- Synonyms

30 er.. 230 years later

"To your request of my opinion of the manner in which a newspaper should be conducted, so as to be most useful, I should answer, "by restraining it to true facts & sound principles only." Yet I fear such a paper would find few subscribers. It is a melancholy truth, that a suppression of the press could not more completely deprive the nation of its benefits, than is done by its abandoned prostitution to falsehood. Nothing can now be believed which is seen in a newspaper."

Thomas Jefferson, 11 June 1807

Conclusion

- Aggregation IS possible
- Can be reused beyond news
- Extend stored content to measure bias

Thanks

- Dr Krishnan Pillaipakkamnatt [Advisor]
- Dr Habib Ammari
- Dr Xiang Fu

?