NEAR DUPLICATE NEWS STORY DETECTION REVISITED

OMAR ALONSO[1], DENNIS FETTERLY[2], AND MARK MANASSE[2]

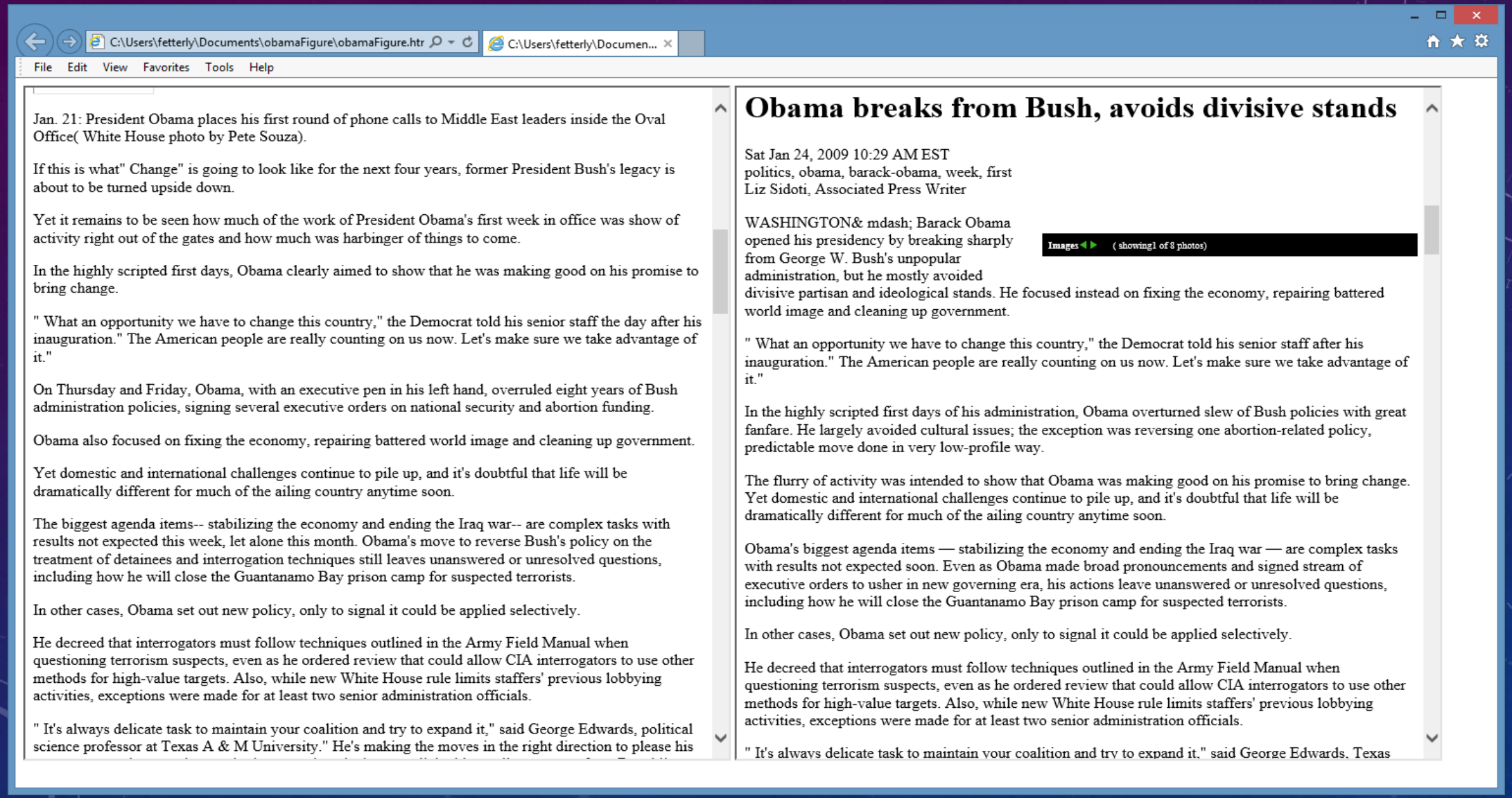
[1]BING SOCIAL

[2]MICROSOFT RESEARCH, SILICON VALLEY LAB

APPLICATIONS OF NEAR-DUPLICATE DETECTION

- Applications
  - Web crawling
  - Removing duplicate search results
  - Plagiarism detection
  - Near-duplicate files in local or remote filesystems
- Our application is near duplicate news document detection

EXAMPLE


C:\Users\fetterly\Documents\obamaFigure\obamaFigure.htr
C:\Users\fetterly\Documen...
File Edit View Favorites Tools Help
Jan. 21: President Obama places his first round of phone calls to Middle East leaders inside the Oval Office(White House photo by Pete Souza).
If this is what" Change" is going to look like for the next four years, former President Bush's legacy is about to be turned upside down.
Yet it remains to be seen how much of the work of President Obama's first week in office was show of activity right out of the gates and how much was harbinger of things to come.
In the highly scripted first days, Obama clearly aimed to show that he was making good on his promise to bring change.
" What an opportunity we have to change this country," the Democrat told his senior staff the day after his inauguration." The American people are really counting on us now. Let's make sure we take advantage of it."
On Thursday and Friday, Obama, with an executive pen in his left hand, overruled eight years of Bush administration policies, signing several executive orders on national security and abortion funding.
Obama also focused on fixing the economy, repairing battered world image and cleaning up government.
Yet domestic and international challenges continue to pile up, and it's doubtful that life will be dramatically different for much of the ailing country anytime soon.
The biggest agenda items-- stabilizing the economy and ending the Iraq war-- are complex tasks with results not expected this week, let alone this month. Obama's move to reverse Bush's policy on the treatment of detainees and interrogation techniques still leaves unanswered or unresolved questions, including how he will close the Guantanamo Bay prison camp for suspected terrorists.
In other cases, Obama set out new policy, only to signal it could be applied selectively.
He decreed that interrogators must follow techniques outlined in the Army Field Manual when questioning terrorism suspects, even as he ordered review that could allow CIA interrogators to use other methods for high-value targets. Also, while new White House rule limits staffers' previous lobbying activities, exceptions were made for at least two senior administration officials.
" It's always delicate task to maintain your coalition and try to expand it," said George Edwards, political science professor at Texas A & M University." He's making the moves in the right direction to please his
Obama breaks from Bush, avoids divisive stands
Sat Jan 24, 2009 10:29 AM EST
politics, obama, barack-obama, week, first
Liz Sidoti, Associated Press Writer
WASHINGTON& mdash; Barack Obama opened his presidency by breaking sharply from George W. Bush's unpopular administration, but he mostly avoided divisive partisan and ideological stands. He focused instead on fixing the economy, repairing battered world image and cleaning up government.
Images (showing1 of 8 photos)
" What an opportunity we have to change this country," the Democrat told his senior staff after his inauguration." The American people are really counting on us now. Let's make sure we take advantage of it."
In the highly scripted first days of his administration, Obama overturned slew of Bush policies with great fanfare. He largely avoided cultural issues; the exception was reversing one abortion-related policy, predictable move done in very low-profile way.
The flurry of activity was intended to show that Obama was making good on his promise to bring change. Yet domestic and international challenges continue to pile up, and it's doubtful that life will be dramatically different for much of the ailing country anytime soon.
Obama's biggest agenda items — stabilizing the economy and ending the Iraq war — are complex tasks with results not expected soon. Even as Obama made broad pronouncements and signed stream of executive orders to usher in new governing era, his actions leave unanswered or unresolved questions, including how he will close the Guantanamo Bay prison camp for suspected terrorists.
In other cases, Obama set out new policy, only to signal it could be applied selectively.
He decreed that interrogators must follow techniques outlined in the Army Field Manual when questioning terrorism suspects, even as he ordered review that could allow CIA interrogators to use other methods for high-value targets. Also, while new White House rule limits staffers' previous lobbying activities, exceptions were made for at least two senior administration officials.
" It's always delicate task to maintain your coalition and try to expand it," said George Edwards, Texas

EXAMPLE (DETAILS)

" What an opportunity we have to change this country," the Democrat told his senior staff the day after his inauguration." The American people are really counting on us now. Let's make sure we take advantage of it."

On Thursday and Friday, Obama, with an executive pen in his left hand, overruled eight years of Bush administration policies, signing several executive orders on national security and abortion funding.

Obama also focused on fixing the economy, repairing battered world image and cleaning up government.

Yet domestic and international challenges continue to pile up, and it's doubtful that life will be dramatically different for much of the ailing country anytime soon.

The biggest agenda items-- stabilizing the economy and ending the Iraq war-- are complex tasks with results not expected this week, let alone this month. Obama's move to reverse Bush's policy on the treatment of detainees and interrogation techniques still leaves unanswered or unresolved questions, including how he will close the Guantanamo Bay prison camp for suspected terrorists.

In other cases, Obama set out new policy, only to signal it could be applied selectively.

" What an opportunity we have to change this country," the Democrat told his senior staff after his inauguration." The American people are really counting on us now. Let's make sure we take advantage of it."

In the highly scripted first days of his administration, Obama overturned slew of Bush policies with great fanfare. He largely avoided cultural issues; the exception was reversing one abortion-related policy, predictable move done in very low-profile way.

The flurry of activity was intended to show that Obama was making good on his promise to bring change. Yet domestic and international challenges continue to pile up, and it's doubtful that life will be dramatically different for much of the ailing country anytime soon.

Obama's biggest agenda items — stabilizing the economy and ending the Iraq war — are complex tasks with results not expected soon. Even as Obama made broad pronouncements and signed stream of executive orders to usher in new governing era, his actions leave unanswered or unresolved questions, including how he will close the Guantanamo Bay prison camp for suspected terrorists.

In other cases, Obama set out new policy, only to signal it could be applied selectively.

SPOTSIGS

- Intuition that common words would prefix phrases in content portion of a news document
- Binary weight applied after antecedent
- Antecedent sets range from “Is” to 571 term SMART stopword list
- Selects terms along some chain length, skips antecedents
- Potential improvements:
  - Not all documents will generate samples
  - Evaluated on a small test collection

OUR APPROACH

- Generalize SpotSigs idea that phrases beginning with common words are more likely to be in the important portions of a document
  - Weight phrases using the document frequency of the first term in the window
  - Scale weight by IDF of the phrase or filter to remove “common” phrases
  - Apply function to calculate final weight
- Approximate Weighted Jaccard via a sample based computation
  - Weighted Jaccard for weights W_1 and W_2 and phrases *U*
  - $0 \leq \frac{\sum_{u \in U} \min(W_1(u), W_2(u))}{\sum_{u \in U} \max(W_1(u), W_2(u))} = \frac{\|\min(W_1, W_2)\|_1}{\|\max(W_1, W_2)\|_1} \leq \frac{\|W_1\|_1}{\|W_2\|_1} \leq 1$
- Leverage Ioffe’s constant time sampling approach

Uniform	$\log^2$ DF	DF^2
DF	$\log^2$ DF	DF^3
log DF	$\log^2$ DF	DF^4
log (DocCount/DF)	$\log^{10}$ DF	

CONSTRUCTING THE EVALUATION CORPUS

- Start with parsed ClueWeb ‘09 Category A English – 500 million web pages
- Filter to include only web pages from Open Directory Project (DMOZ) News category
  - 11.8 million web pages from 7,261 distinct hostnames
- Remove Wikipedia content and duplicate pages
- Resulting corpus is 5.5 million web pages

GENERATING LABELED DATA

- Extract the potential news article from the document using the Maximum Subsequence Segmentation library
- Group documents that share at least 1 phrase of 7 words where the IDF of the phrase is in [0.2, 0.85]
- Uniformly draw samples (pairs of documents) proportional to the size of group
- Calculate unweighted Jaccard for each sampled pair
- Use a histogram of these Jaccard values to sample pairs distributed evenly across all Jaccard values
  - makes problem harder; most pairs have near-zero Jaccard similarity

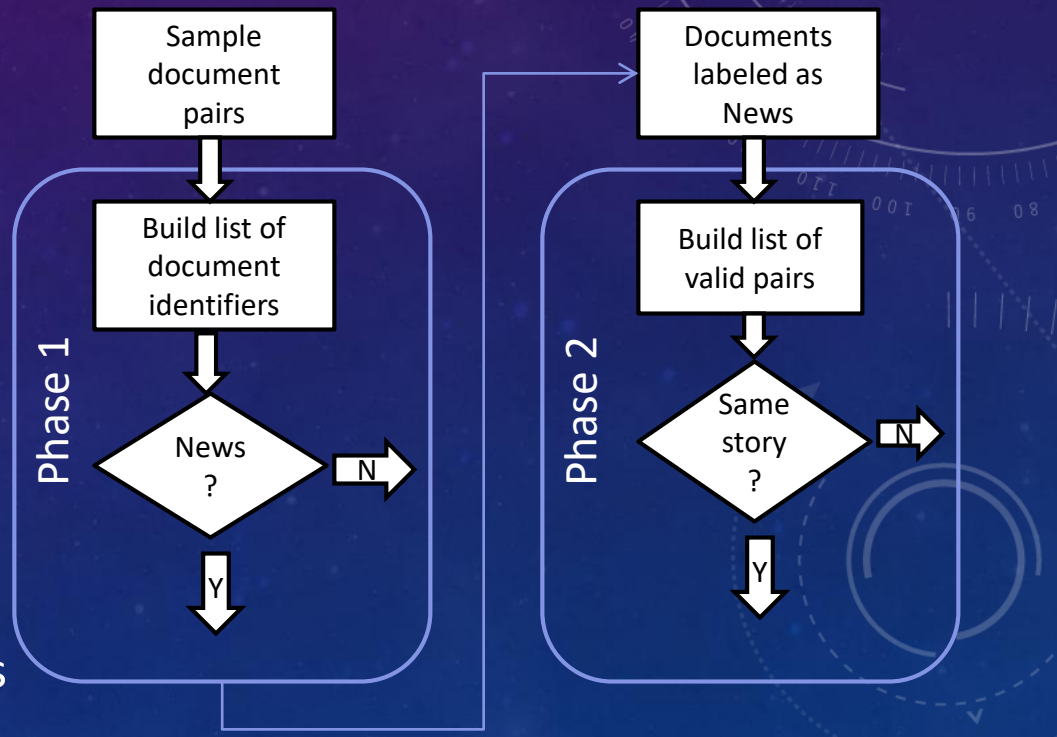
LABELING SAMPLED PAIRS

- Built a simple document comparison tool
  - Identify duplicate content, extracted news article
  - Lack of text formatting added difficulty
- Labeled 456 pairs of documents as one of:
  - Containment, Duplicate, Non-duplicate
  - Duplicate Irrelevant, and Non-duplicate Irrelevant
  - Manual resolution of disagreement
- Initial evaluation was promising

Label	Frequency
Duplicate	42
Containment	10
Non-duplicate	252
Duplicate, Irrelevant	10
Non-duplicate, Irrelevant	142

SCALING VIA CROWDSOURCING

- Leverage crowdsourcing platform to generate labels
  - Microsoft Universal Human Relevance System (UHRS)
  - Amazon's Mechanical Turk
- Manual assessment is a time consuming activity
  - More accessible approach for assessors
- Leverage sanitized HTML markup
- Two phase approach
  - Assess documents in sampled pairs as news/not news
  - Duplicate or non-duplicate assessment for news pairs



Please read the following document and let us know if the article is a news article in English

The following are **considered** news:

- Generic news articles (e.g., finance, sports, politics, etc.)
- Press releases by a company or institution
- Government data reports (example) or election returns (example)

The following are **not considered** news:

- Weather reports (conditions)
- Headlines with a single or no sentences/paragraph or photo captions (example)
- Event calendars (example)
- Article or blog post that describes another article with a teaser quoting a small amount of content from the original article.
- Library collections (example) or course listings (example)

Politics Home

Obama Sticks to the Script in First Week of Presidency

In the highly scripted first days of Obama's administration, the flurry of activity was intended to show that he was making good on his promise to bring change.

FOXNews. com

Saturday, January 24, 2009

- Photos

President Obama

Does the document above contain a **news** article written in English?

- **Yes**. It is a news article.
- **No**. It is not a news article.
- **I don't know**. I can't tell if the document is a news article.
- **Other**. Web page didn't load/error message/etc.
- **Non English**. This document is not in English.

Please only consider the article itself (not the surrounding template). Ignore ads, images and formatting. We are only considering the core text of the articles. Note that headlines can be different.

1. Are these 2 news articles about the same event/topic?

- **Yes**. These news articles are about the same.
- **No**. These news articles are not the same.
- **I don't know**. I can't tell if the news articles are the same or not.
- **Other**. Web page didn't load/error message/etc.
- **Non English**. This document is not in English.

2. Does one document cover more detail than the other?

- Document A covers more detail than document B.
- Document B covers more detail than document A.
- No

Document A

Obama Sticks to the Script in First Week of Presidency

In the highly scripted first days of Obama's administration, the flurry of activity was intended to show that he was making good on his promise to bring change.

FOXNews. com

Saturday, January 24, 2009

- Photos

President Obama

Jan. 21: President Obama places his first round of phone calls to Middle East leaders inside the Oval Office(White House photo by Pete Souza).

Document B

Put a Seed Newsvine link on your own site. Here's how...

- 6 Votes
- 19 Comments
- Print

Obama breaks from Bush, avoids divisive stands

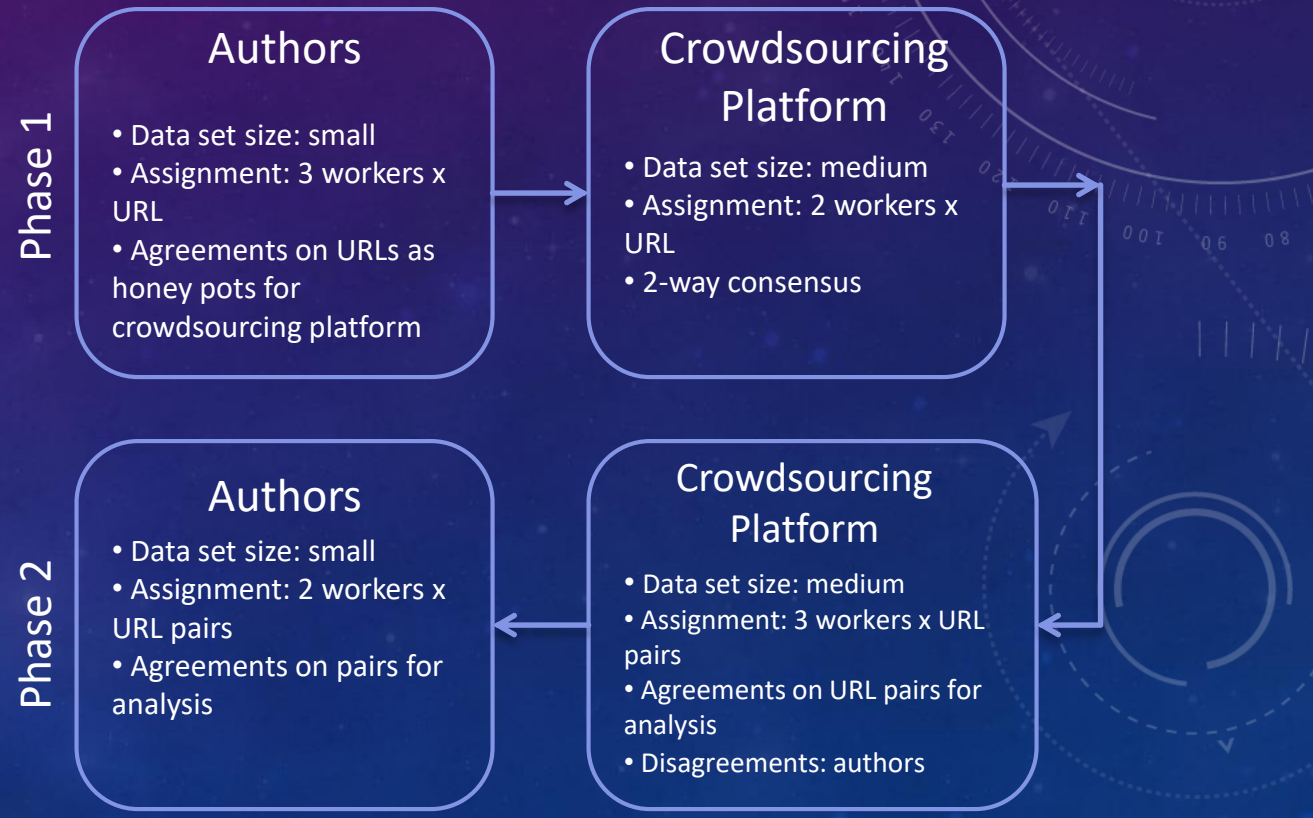
Sat Jan 24, 2009 10:29 AM EST
politics, obama, barack-obama, week, first
Liz Sidoti, Associated Press Writer

WASHINGTON& mdash; Barack Obama opened his presidency by breaking sharply from George W. Bush's unpopular administration, but he mostly avoided divisive partisan and ideological stands. He focused instead on fixing the economy, repairing battered

Images ◀ ▶ (showing1 of 8 photos)

CROWDSOURCING QUALITY CONTROL

- Leverage different assessors at each phase
- Generate known good labels for each phase
  - Used as honey pots for crowd labels
- Phase 1 documents with consensus as news are used in Phase 2 pairs
- Phase 2 disagreements are resolved by the authors
- Authors used same platform



UHRS LABELS

Raw Phase 1 Labels

Label	Frequency
Yes	4,235
No	7,877
I don’t know	208
Other	412
Non-English	45

Consensus Phase 2 Labels

Label	Frequency
Yes	323
No	386
I don’t know	0
Other	1

RARE PHRASES

- Goal: Filter out boilerplate phrases that occur in many documents
- Set phrase weight to 0 if phrase occurred in more than n% documents
  - n: {1, 5, 25, 50, 75, 95, 99}
- For Clueweb'09 with a phrase length of 7 words, this is at most 847,281 elements
- All variants perform well, 50% performs best on our collection
  - Better than all other techniques
- Statistical significance when comparing to SpotSigs for accuracy of prediction on our dataset
  - Measure difference in agreement with ground truth

EVALUATION – F1

- Recall – fraction of detected positives $\frac{TP}{TP+FN}$
- Precision – fraction of positives which are correct $\frac{TP}{TP+FP}$
- F1 measure – $2\frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} = \frac{2TP}{2TP+FP+FN}$

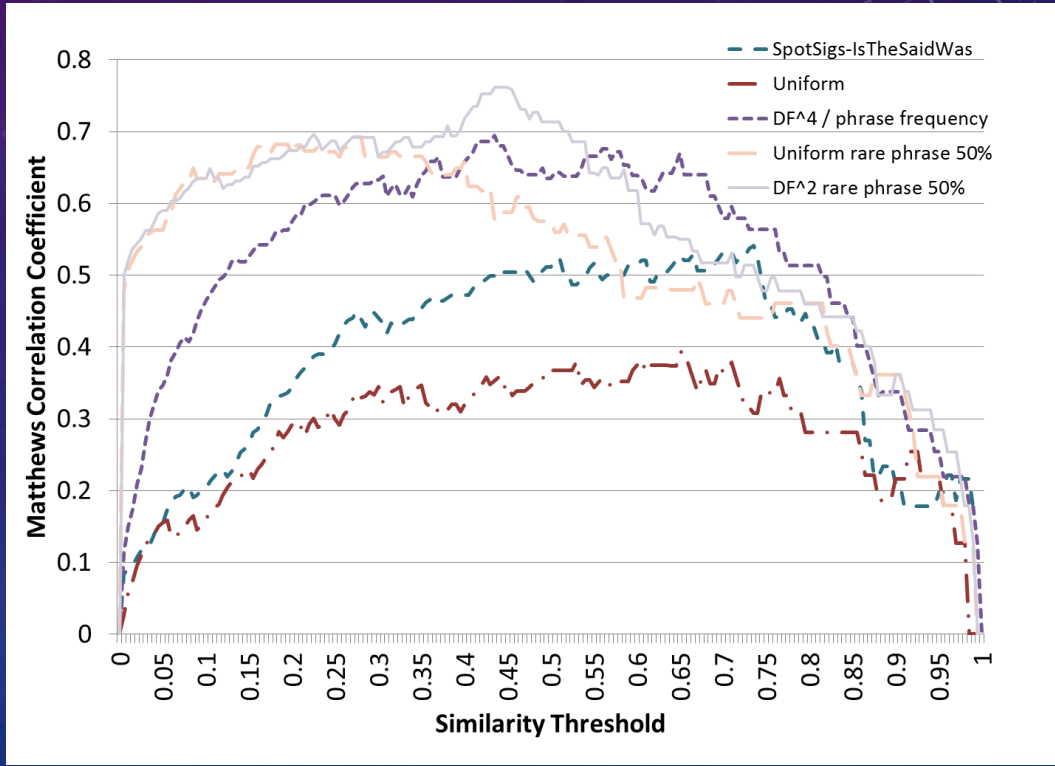
RARE PHRASE RESULTS

Antecedent List	Max F1
Is	0.6720
The	0.7868
Is, The	0.7851
Is,The,Said	0.7901
Is,The,Said,Was	0.7896
Is,The,Said,Was,There	0.7886
Is,The,Said,Was,There,A	0.7918
Is,The,Said,Was,There,A,It	0.7967
The,A,Can,Be,Will,Have,Do	0.7734
SMART Stopword List (571 words)	0.7572

Weighting Function	Max F1
Uniform	0.8352
DF	0.8275
log DF	0.8429
log IDF	0.8093
DF^2	0.8505
DF^3	0.8449
DF^4	0.8470
$\log^2$ DF	0.8352
$\log^3$ DF	0.8288
$\log^4$ DF	0.8257
$\log^{10}$ DF	0.8307

MATTHEWS CORRELATION COEFFICIENT

- Defined as $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
- Interval [-1,1]
- Value of 1 indicates perfect classification
- Best performing methods in figure
- Uniform weighting approximates shingling
- Uniform with rare phrases depicts impact of rare phrase filtering



CONCLUSIONS, FUTURE WORK, AND AVAILABILITY

- Effectively detect near duplicate news articles
- Generalization of SpotSigs using term and phrase frequencies
- Significant performance improvement
- Two phase crowdsourced pipeline to economically gather labels
- Could apply technique to languages other than English

- The document collection is available via
  - http://research.microsoft.com/en-us/projects/newsdupedetect/