News and Blog Analysis with Lydia

Steven Skiena
Dept. of Computer Science
SUNY Stony Brook
http://www.cs.sunysb.edu/~skiena

Opportunities in Text Analysis

- The increasing volume of online information coupled with decreasing costs of communications and computation creates exciting new opportunities in text mining.
- We can analyze all of the 1000+ online English-language newspapers daily on a single commodity computer!
- Our ultimate goal is to build a *computational model* of much of the world's knowledge through analysis of news media, reference texts, and primary sources.

Learning From Text Sources

Knowledge extraction becomes easier when you start with reliable sources:

- Online newspapers (both domestic and foreign)
  what is happening in the world?
- Scientific abstracts, e.g. Medline/Pubmed
  what is known about disease and medicine?
- Blogs
  what do people *think* is happening in the world?

www.textmap.com

TEXTMAP
THE ENTITY SEARCH ENGINE
Monitoring the World So You Don't Have To ...

ENTITIES SOURCES CONTACT

Search!

Text Map : TextMood TextBlog TextBiz Make this page : Shortcuts : Help

George W. Bush

PERSON
22853 references in 8325 articles [Show Articles]
[Web Query] [Popularity Time Series] [Heatmap]

Complete

Name	Corr.	Rel.
Iraq		
White House		
Republican		
Harriet Miers		
Democrats		
Bush		
Dick Cheney		
Karl Rove		
Al Gore		
Washington, DC		
Social Security		
Americans		
John Roberts		
Bill Clinton		

30 Days

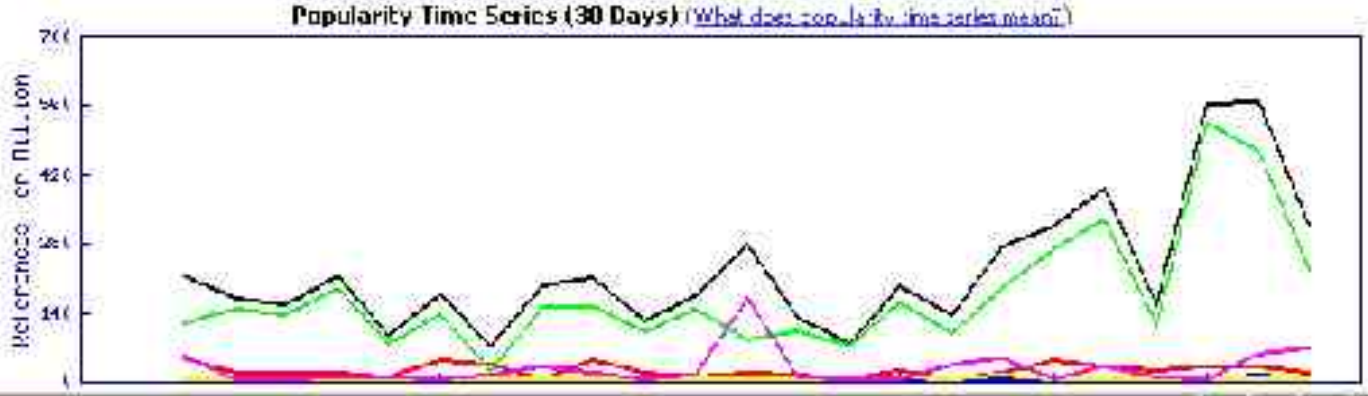
Name	Corr.	Rel.
Bush		
Democrats		
Iraq		
Republican		
Congress		
Washington, DC		
Senate Alito		
Americans		
Dick Cheney		
U.S.		
United States		
Social Security		
Supreme Court		
Clinton		
Al Gore		

7 Days

Name	Corr.	Rel.
Bush		
Democrats		
Iraq		
Congress		
Republican		
Washington, DC		
Social Security		
Americans		
Senate Alito		
Jesse Sekulow		
United States		
U.S.		
Bill Clinton		
America		
Clinton		

Also Known As
George W. Bush, George Bush

Popularity Time Series (30 Days) (What does popularity time series mean?)



Done
Internet

Search!

Jack Abramoff

PERSON
[illegible] references in 8601 articles [Show Articles]
[Web Query] [Popularity Time Series]

Complete

Name	Cum	Ref
Tom DeLay		
Bob Ney		
Congress		
DeLay		
Adam Kidan		
Republican		
Ney		
Indian		
Michael Scanlon		
House		
[illegible]		
[illegible]		
John Doolittle		
House Administration Committee		

30 Days

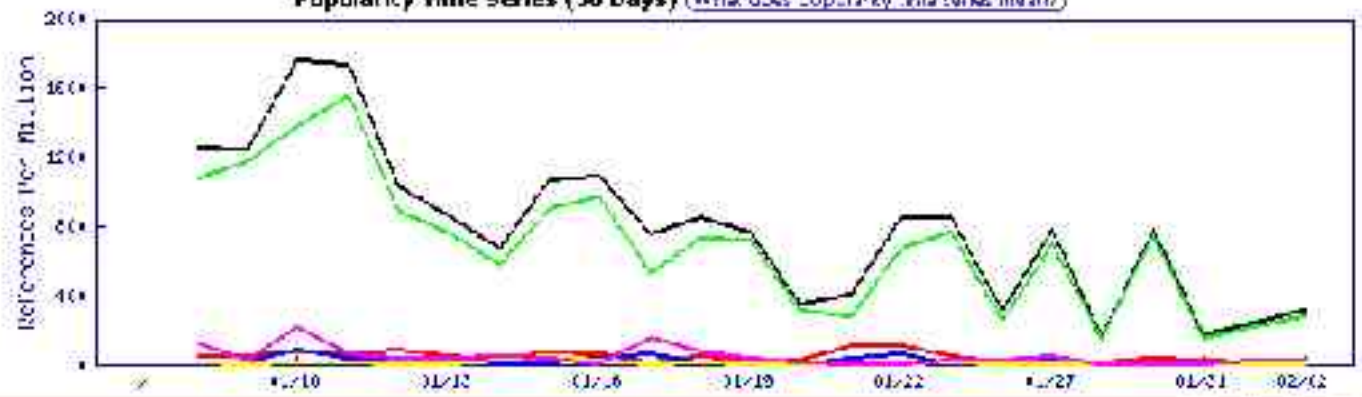
Name	Cum	Ref
Tom DeLay		
Republican		
DeLay		
Bob Ney		
Congress		
Bush		
Ney		
Indian		
Washington, DC		
House		
Democrats		
John Doolittle		
House Administration Committee		
Speaker Dennis Hastert		

7 Days

Name	Cum	Ref
Republican		
Tom DeLay		
David H. Safavian		
John Boehner		
DeLay		
Bush		
Democrats		
John Doolittle		
Boehner		
Indian		
Congress		
Ralph Reed		
House Republicans		
Greenberg Traurig		

Also Known As
Jack Abramoff, Abramoff, Lobbyist Jack Abramoff

Popularity Time Series (30 Days) (What does popularity time series mean?)



Newsday
https://www.newsday.com

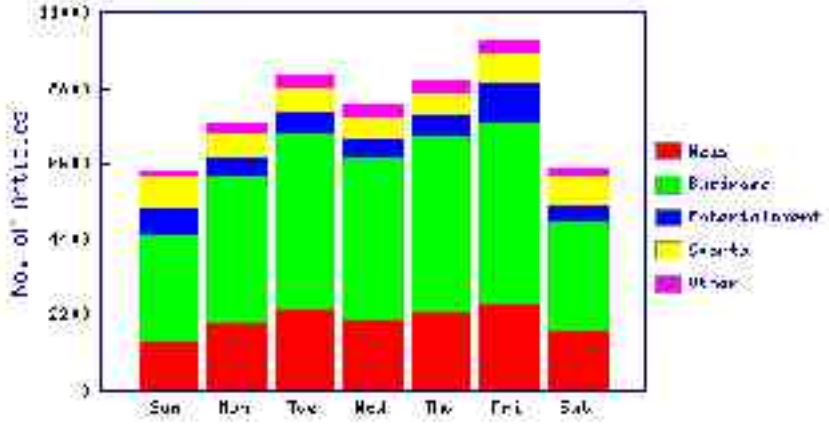
No. of Files: 280
No. of Articles: 5,181

Over Populated Entities
Entity Name Frequency Standard Deviation

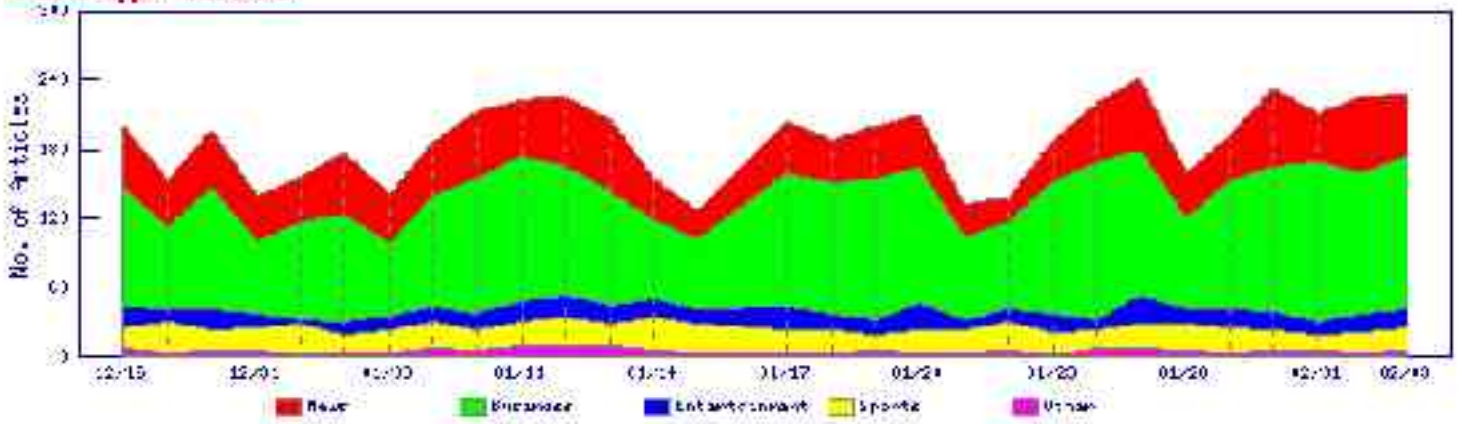
Under Populated Entities
Entity Name Frequency Standard Deviation

Content Type Distribution

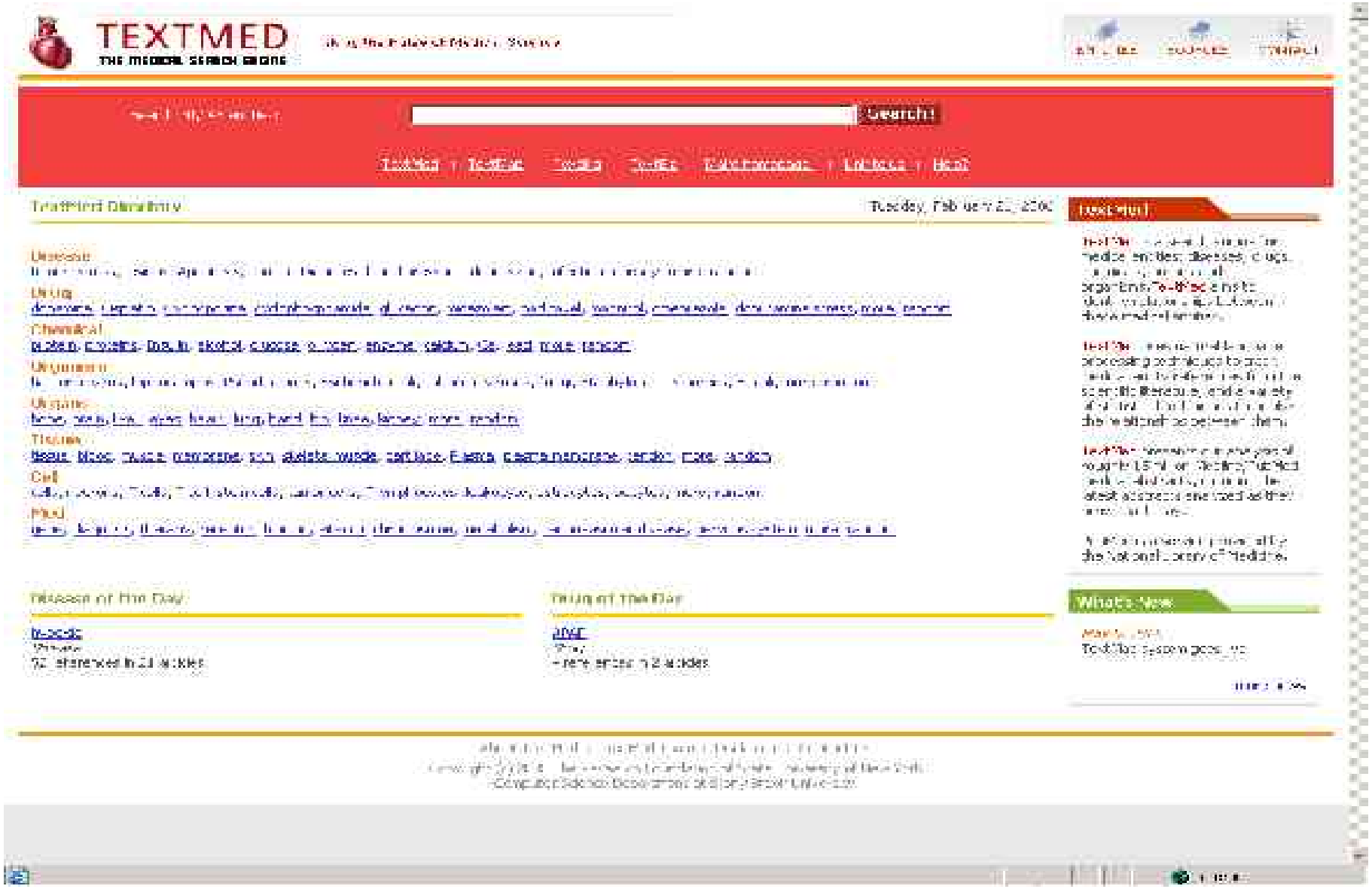
Daily Content Type Distribution



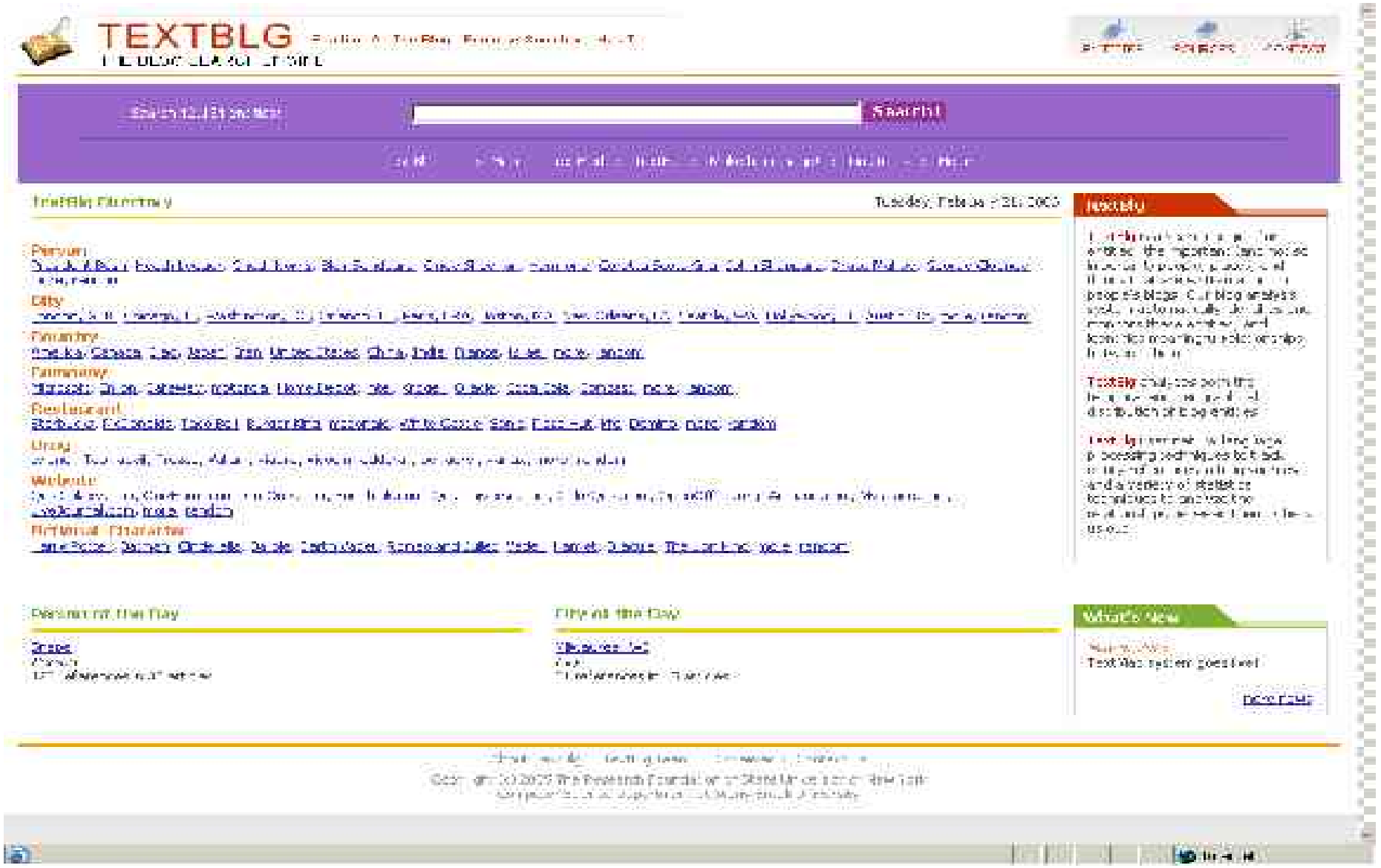
Content Type Time Series



www.textmed.com



www.textblg.com



President Bush

PERSON
2139 references in 1972 articles [Show Articles]
[Web Query] [Popularity Time Series]

Complete

Name	Cont. / Ref.
Bush	
Congress	
Iraq	
Americans	
Vice President	
United States	
Jack Abramoff	
State	
Republican	
Constitution	
America	
Washington, DC	
U.S.	
Confederate States	

30 Days

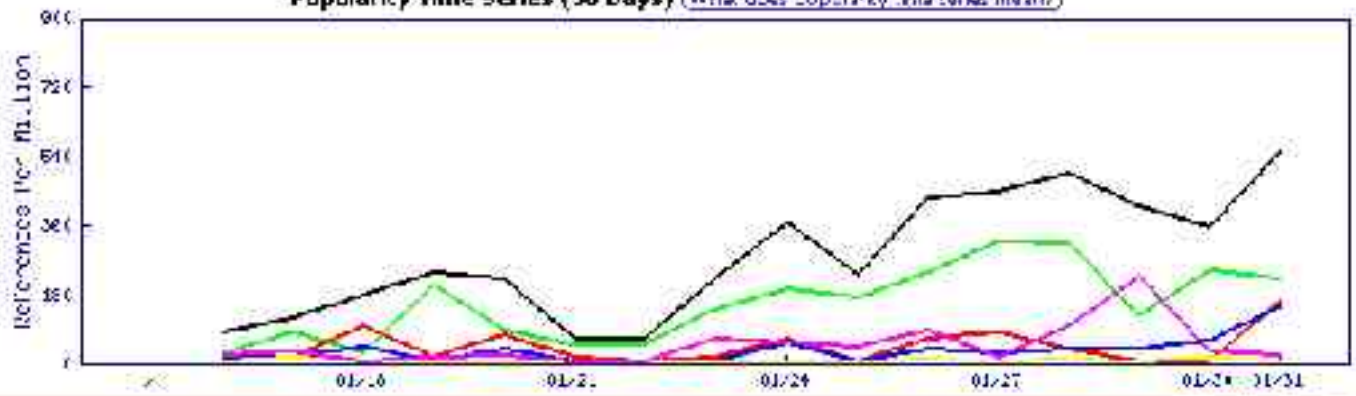
Name	Cont. / Ref.
Bush	
Congress	
State	
Jack Abramoff	
Vice President	
Iraq	
Americans	
Republican	
Confederate States	
United States	
Union	
America	
Constitution	
Washington, DC	

7 Days

Name	Cont. / Ref.
Bush	
Union	
State	
America	
Iraq	
Condoleezza Rice	
United States	
Congress	
Republican	
Social Security	
Americans	
Washington, DC	
Syrian Al-Baath Party	
Constitution	

Also Known As
President Bush, President

Popularity Time Series (30 Days) (What does popularity time series mean?)



Related Systems

- IBM's WebFountain
  - Framework for web-scale text analytics
- Columbia's Newsblaster
  - Provides summaries of online newspaper articles
- Google News
  - Automatic aggregation and presentation of online news.

Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions

System Architecture

Spidering – text is retrieved from a given site on a daily basis using semi-custom spidering agents.

Normalization – clean text is extracted with semi-custom parsers and formatted for our pipeline

Text Markup – annotates parts of the source text for storage and analysis.

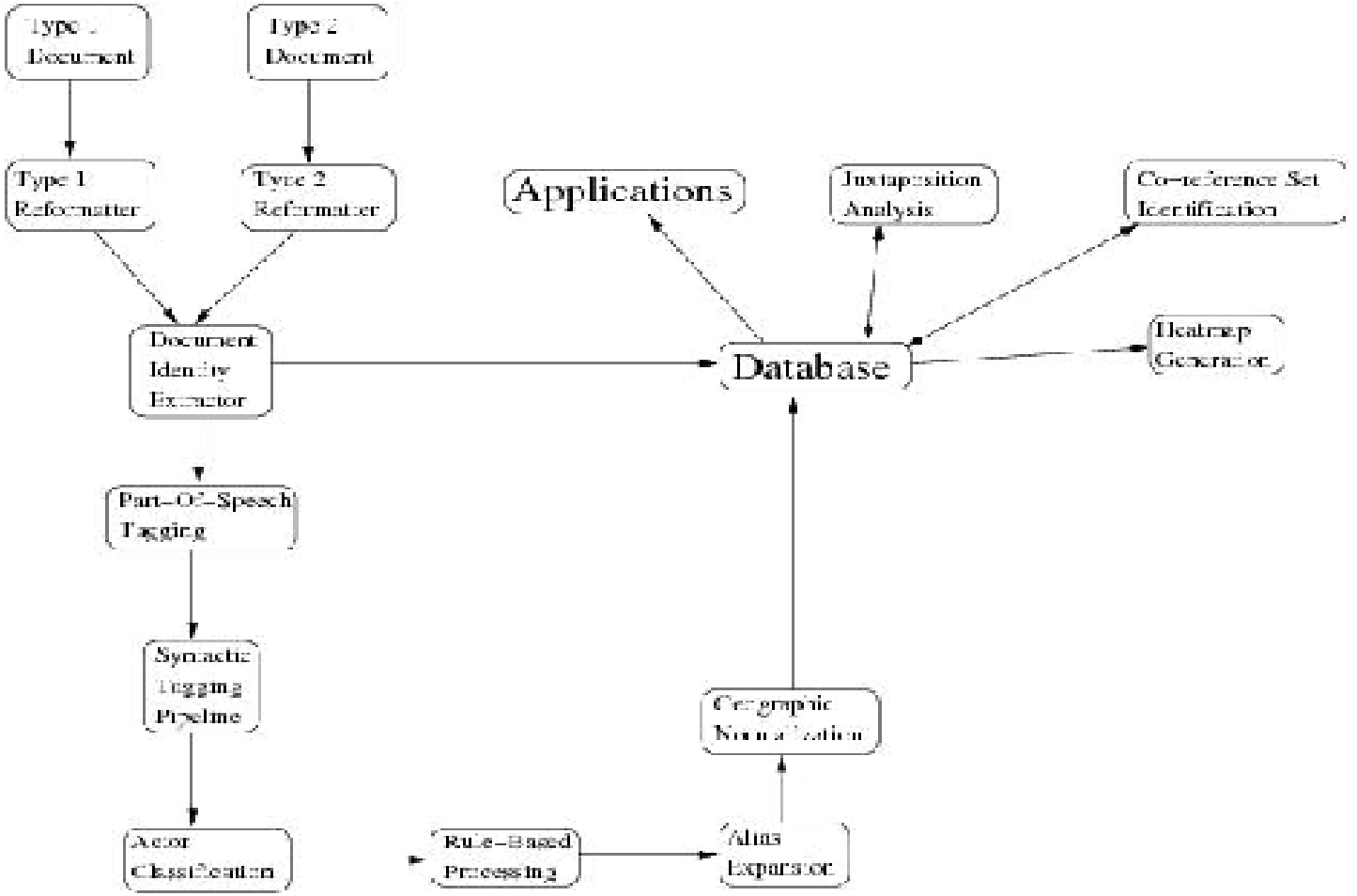
Back Office Operations – we aggregate entity frequency and relational data for a variety of statistical analyses.

[1] Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. Lydia: A System for Large-Scale News Analysis. In *String Processing and Information Retrieval: 12th International Conference (SPIRE 2005)*.

Text Markup

- We apply natural language processing (NLP) techniques to annotate interesting features of the document.
- Full parsing techniques are too slow to keep up with our volume of text, so we employ shallow parsing instead.
- We can currently markup approximately 2000 newspapers per day per CPU.
- Analysis phases include…

Flowchart


Type 1 Document
Type 2 Document
Type 1 Reformatter
Type 2 Reformatter
Applications
Juxtaposition Analysis
Co-reference Set Identification
Document Identifier Extractor
Database
Heatmap Generator
Part-Of-Speech Tagging
Syntactic Tagging Pipeline
Geographic Normalization
Actor Classification
Rule-Based Processing
Alias Expansion

Input

Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York.

She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president. Mr. Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.

Sentence and Paragraph Identification

Dr. Judith Rodin, the former president of the University of Pennsylvania, will become president of the Rockefeller Foundation next year, the foundation announced yesterday in New York.

She will take over in March 2005, succeeding Gordon Conway, the foundation's first non-American president.
Mr. Conway announced last year that he would retire at 66 in December and return to Britain, where his children and grandchildren live.

Part Of Speech Tagging

Dr./NNP Judith/NNP Rodin/NNP ,/, the/DT former/JJ president/NN of/IN the/DT University/NNP of/IN Pennsylvania/NNP ,/, will/MD become/VB president/NN of/IN the/DT Rockefeller/NNP Foundation/NN next/JJ year/NN ,/, the/DT foundation/NN announced/VBD yesterday/RB in/IN New/NNP York/NNP./.

She/PRP will/MD take/VB over/IN in/IN March/NNP 2005/CD ,/, succeeding/VBG Gordon/NNP Conway/NNP ,/, the/DT foundation/NN 's/POS first/JJ non-American/JJ president/NN ./.
Mr./NNP Conway/NNP announced/VBD last/JJ year/NN that/IN he/PRP would/MD retire/VB at/IN 66/CD in/IN December/NNP and/CC return/NN to/TO Britain/NNP ,/, where/WRB his/PRP$ children/NNS and/CC grandchildren/NNS live/VBP ./.

Proper Noun Extraction

<pn> Dr./NNP Judith/NNP Rodin/NNP </pn> ,/, the/DT former/JJ president/NN of/IN the/DT <pn> University/NNP </pn> of/IN <pn> Pennsylvania/NNP </pn> ,/, will/MD become/VB president/NN of/IN the/DT <pn> Rockefeller/NNP </pn> Foundation/NN next/JJ year/NN ,/, the/DT foundation/NN announced/VBD yesterday/RB in/IN <pn> New/NNP York/NNP </pn> ./.

She/PRP will/MD take/VB over/IN in/IN March/NNP 2005/CD ,/, succeeding/VBG <pn> Gordon/NNP Conway/NNP </pn> ,/, the/DT foundation/NN 's/POS first/JJ non-American/JJ president/NN ./.
<pn> Mr./NNP Conway/NNP </pn> announced/VBD last/JJ year/NN that/IN he/PRP would/MD retire/VB at/IN 66/CD in/IN December/NNP and/CC return/NN to/TO <pn> Britain/NNP </pn> ,/, where/WRB his/PRP$ children/NNS and/CC grandchildren/NNS live/VBP ./.

Date and Number Extraction

<pn> Dr./NNP Judith/NNP Rodin/NNP </pn> ,/, the/DT former/JJ president/NN of/IN the/DT <pn> University/NNP </pn> of/IN <pn> Pennsylvania/NNP </pn> ,/, will/MD become/VB president/NN of/IN the/DT <pn> Rockefeller/NNP </pn> Foundation/NN next/JJ year/NN ,/, the/DT foundation/NN announced/VBD yesterday/RB in/IN <pn> New/NNP York/NNP </pn> ./.

She/PRP will/MD take/VB over/IN in/IN <embedded_date> March/NNP 2005/CD </embedded_date> ,/, succeeding/VBG <pn> Gordon/NNP Conway/NNP </pn> ,/, the/DT foundation/NN 's/POS <num type = "ORDINAL"> first/JJ </num> non-American/JJ president/NN ./.
<pn> Mr./NNP Conway/NNP </pn> announced/VBD last/JJ year/NN that/IN he/PRP would/MD retire/VB at/IN <num type = "CARDINAL"> 66/CD </num> in/IN <embedded_date> December/NNP </embedded_date> and/CC return/NN to/TO <pn> Britain/NNP </pn> ,/, where/WRB his/PRP$ children/NNS and/CC grandchildren/NNS live/VBP ./.

Actor Classification

<pn category = "PERSON"> Dr./NNP Judith/NNP Rodin/NNP </pn> ,/, the/DT former/JJ president/NN of/IN the/DT <pn category = "UNKNOWN"> University/NNP </pn> of/IN <pn category = "STATE"> Pennsylvania/NNP </pn> ,/, will/MD become/VB president/NN of/IN the/DT <pn category = "UNKNOWN"> Rockefeller/NNP </pn> Foundation/NN next/JJ year/NN ,/, the/DT foundation/NN announced/VBD yesterday/RB in/IN <pn category = “CITY”> New/NNP York/NNP </pn> ./.

She/PRP will/MD take/VB over/IN in/IN <embedded_date> March/NNP 2005/CD </embedded_date> ,/, succeeding/VBG <pn category = "PERSON"> Gordon/NNP Conway/NNP </pn> ,/, the/DT foundation/NN 's/POS <num type = "ORDINAL"> first/JJ </num> non-American/JJ president/NN ./.
<pn category = "PERSON"> Mr./NNP Conway/NNP </pn> announced/VBD last/JJ year/NN that/IN he/PRP would/MD retire/VB at/IN <num type = "CARDINAL"> 66/CD </num> in/IN <embedded_date> December/NNP </embedded_date> and/CC return/NN to/TO <pn category = "COUNTRY"> Britain/NNP </pn> ,/, where/WRB his/PRP$ children/NNS and/CC grandchildren/NNS live/VBP ./.

Rewrite Rules

<appellation> Dr. </appellation> <pn category = "PERSON"> Judith Rodin </pn> , the former president of the <pn category = "UNIVERSITY"> University of Pennsylvania </pn> , will become president of the <pn category = "UNKNOWN"> Rockefeller Foundation </pn> next year , the foundation announced yesterday in <pn category = “CITY”> New York </pn> .

She will take over in <embedded_date> March 2005 </embedded_date> , succeeding <pn category = "PERSON"> Gordon Conway </pn> , the foundation 's <num type = "ORDINAL"> first </num> non-American president .
<appellation> Mr. </appellation> <pn category = "PERSON"> Conway </pn> announced last year that he would retire at <num type = "CARDINAL"> 66 </num> in <embedded_date> December </embedded_date> and return to <pn category = "COUNTRY"> Britain </pn> , where his children and grandchildren live .

Alias Expansion

<appellation> Dr. </appellation> <pn category = "PERSON"> Judith Rodin </pn> , the former president of the <pn category = "UNIVERSITY"> University of Pennsylvania </pn> , will become president of the <pn category = "UNKNOWN"> Rockefeller Foundation </pn> next year , the foundation announced yesterday in <pn category = “CITY”> New York </pn> .

She will take over in <embedded_date> March 2005 </embedded_date> , succeeding <pn category = "PERSON"> Gordon Conway </pn> , the foundation 's <num type = "ORDINAL"> first </num> non-American president .
<appellation> Mr. </appellation> <pn category = "PERSON"> Gordon Conway </pn> announced last year that he would retire at <num type = "CARDINAL"> 66 </num> in <embedded_date> December </embedded_date> and return to <pn category = "COUNTRY"> Britain </pn> , where his children and grandchildren live.

Geography Normalization

<appellation> Dr. </appellation> <pn category = "PERSON"> Judith Rodin </pn> , the former president of the <pn category = "UNIVERSITY"> University of Pennsylvania </pn> , will become president of the <pn category = "UNKNOWN"> Rockefeller Foundation </pn> next year , the foundation announced yesterday in <pn category = “CITY, STATE, COUNTRY”> New York City, New York, USA </pn> .

She will take over in <embedded_date> March 2005 </embedded_date> , succeeding <pn category = "PERSON"> Gordon Conway </pn> , the foundation 's <num type = "ORDINAL"> first </num> non-American president .
<appellation> Mr. </appellation> <pn category = "PERSON"> Gordon Conway </pn> announced last year that he would retire at <num type = "CARDINAL"> 66 </num> in <embedded_date> December </embedded_date> and return to <pn category = "COUNTRY"> Britain </pn> , where his children and grandchildren live.

Back Office Operations

- The most interesting analysis occurs after markup, using our MySQL database of all occurrences of interesting entities.
- Each day's worth of analysis yields about 10 million occurrences of about 1 million different entities, so efficiency matters...
- Linkage of each occurrence to source and time facilitates a variety of interesting analysis.

Duplicate Article Elimination

Supreme Court Justice David Souter suffered minor injuries when a group of young men assaulted him as he jogged on a city street, a court spokeswoman and Metropolitan Police said Saturday.

Supreme Court Justice David Souter suffered minor injuries when a group of young men assaulted him as he jogged on a city street, a court spokeswoman and Metropolitan Police said.

Hashing techniques can efficiently identify duplicate and near-duplicate articles appearing in different news sources.

Synonym Sets

- JFK, John Kennedy, John F. Kennedy, and John Fitzgerald Kennedy all refer to the same person.
- We need a mechanism to link multiple entities that have slightly different names but refer to the same thing.
- We say that two actors belong in the same synonym set if:
  - There names are morphologically compatible.
  - If the sets of entities that they are related to are similar.

Morphological Similarity

- Subsequence Similarity
  - *George Bush* = *George W. Bush*
- Stemming
  - *New York Yankees* = *New York Yankee*
- Pronunciation Similarity – Metaphone
  - *Victor Yanukovich* = *Viktor Yanukovych*
- Abbreviations
  - *JFK* = *John F. Kennedy* = *John Fitzgerald Kennedy*

Contextual Similarity

Virginia Woolf
- “The Hours”
- Michael Cunningham
- “Mrs. Dalloway”
- Nicole Kidman
- “Butterfield 8”
- Edward Albee
- Elizabeth Taylor

$\neq$

Virginia
- Alexandria
- Maryland
- North Carolina
- John Allen Muhammad
- Falls Church
- Lee Boyd Malvo
- Tennessee

Bush
- Iraq
- Al Gore
- Democrats
- Texas
- Ari Fleischer
- Florida
- Republican

$=$

George W. Bush
- Al Gore
- Republican
- GOP
- Texas
- Arizona Sen. John McCain
- Florida
- Dick Cheney

Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions

Juxtaposition Analysis

- We want to compute the significance of the co-occurrences between two entities
- Similar to *collaborative filtering*, determining which customers are most similar in order to predict future buying preferences
- Just counting the number of co-occurrences causes the most popular entities to be related to everyone

Scoring Juxtapositions

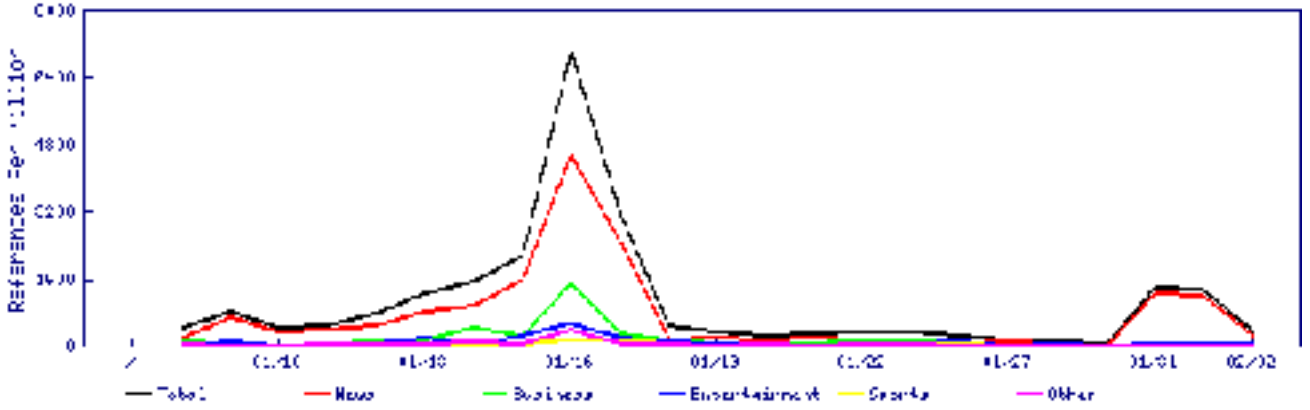
- In collaborative filtering, correlation coefficients and dot products are often used, but they did not work well for our problem
- We use a Chernoff Bound to bound the probability of the number of co-occurrences
  - $P(X > (1+\delta)E[X]) \leq (e^{\delta} / (1+\delta)^{(1+\delta)})E[X]$

Juxtapositions

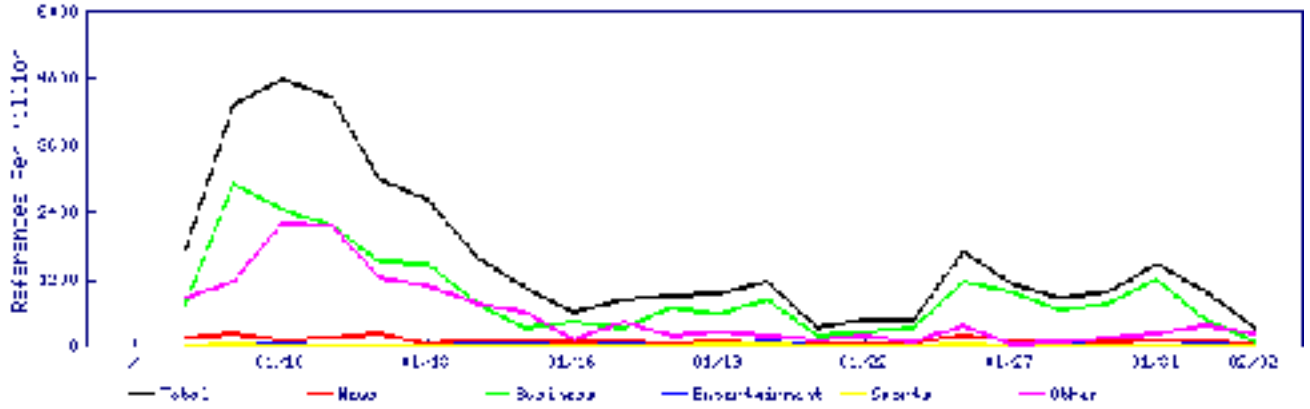
Hamas	Detroit, MI	Google
Israel	Super Bowl	Yahoo
Mahmoud Abbas	Pittsburgh Steelers	Microsoft
Fatah	Jerome Bettis	Internet
Gaza	Detroit Pistons	Web
Palestinian Authority	Pittsburgh, PA	AOL
West Bank	Seattle, WA	Kai-Fu Lee
Palestinians	Minnesota	Larry Page
Islamic Jihad	Chicago, IL	MSN
Islamic	Red Wings	Sergey Brin
Gaza City	NFL	Mountain View, CA

Time Series Analysis

Martin Luther King



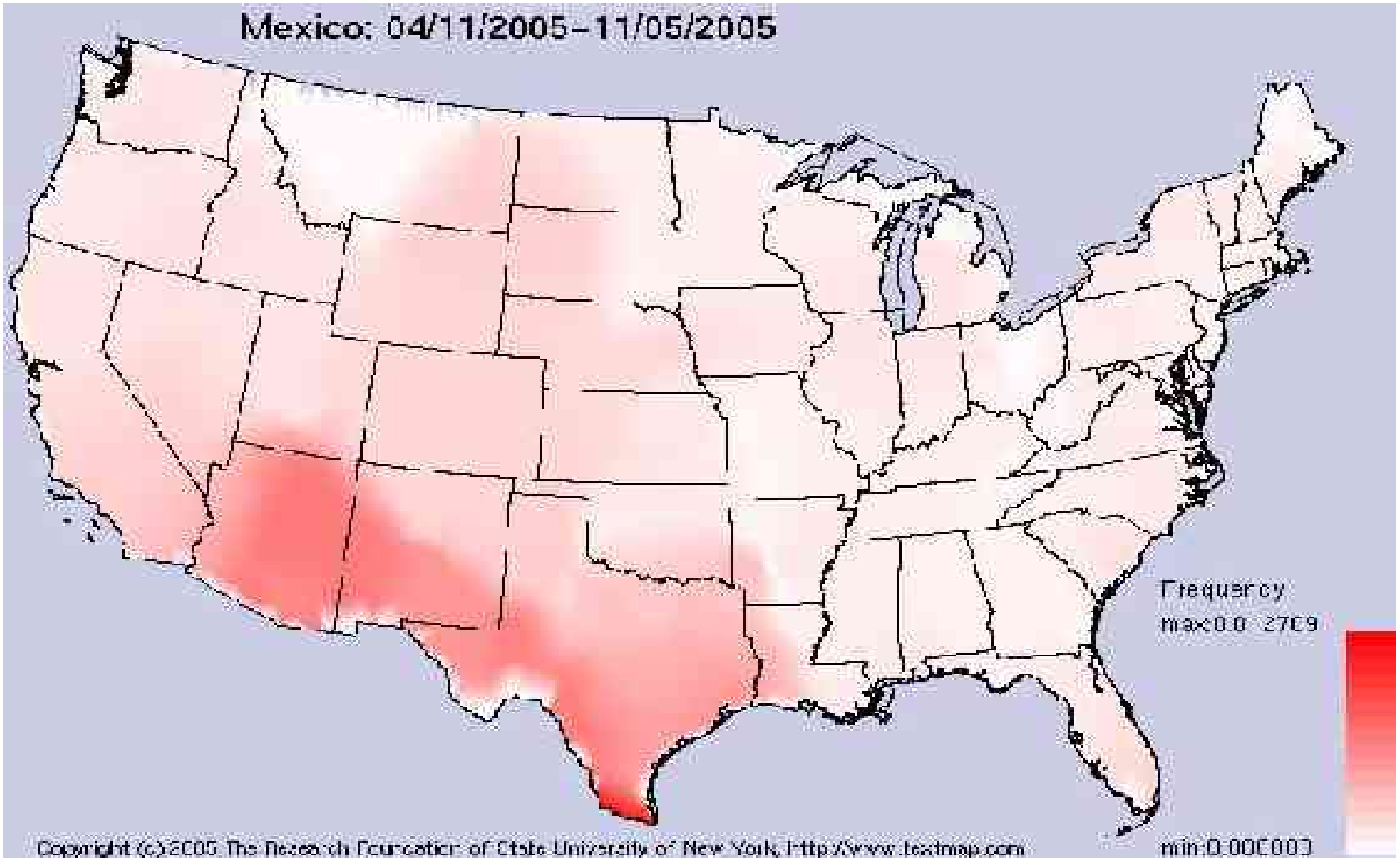
Samuel Alito



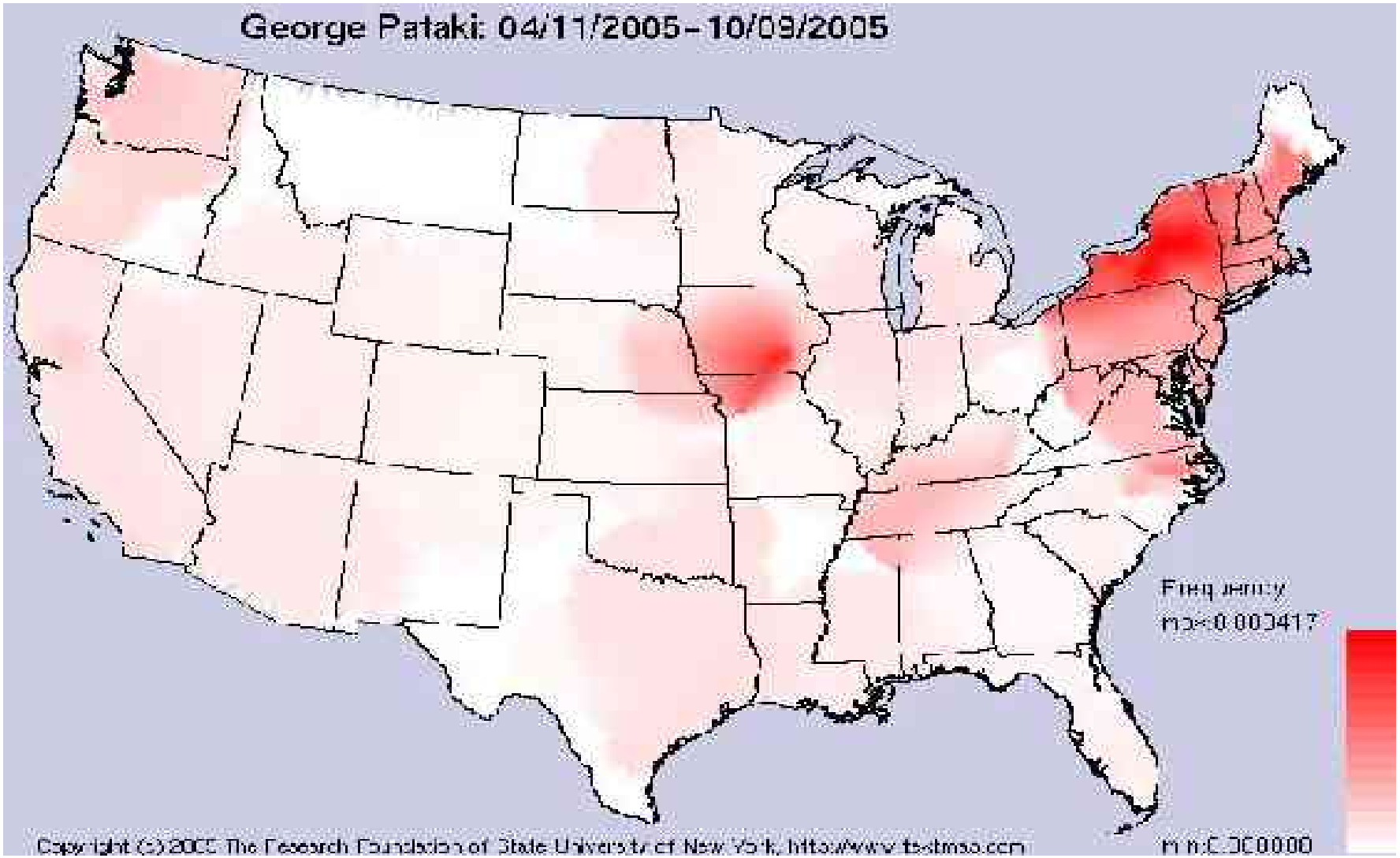
Heatmaps

- Where are people are talking about particular topics?
- Newspapers have a *sphere of influence* based on:
  - Power of the source – circulation, website popularity
  - Population density of surrounding cities
- The *heat* a given entity generates in a particular location is a function of the frequency it is mentioned in local sources

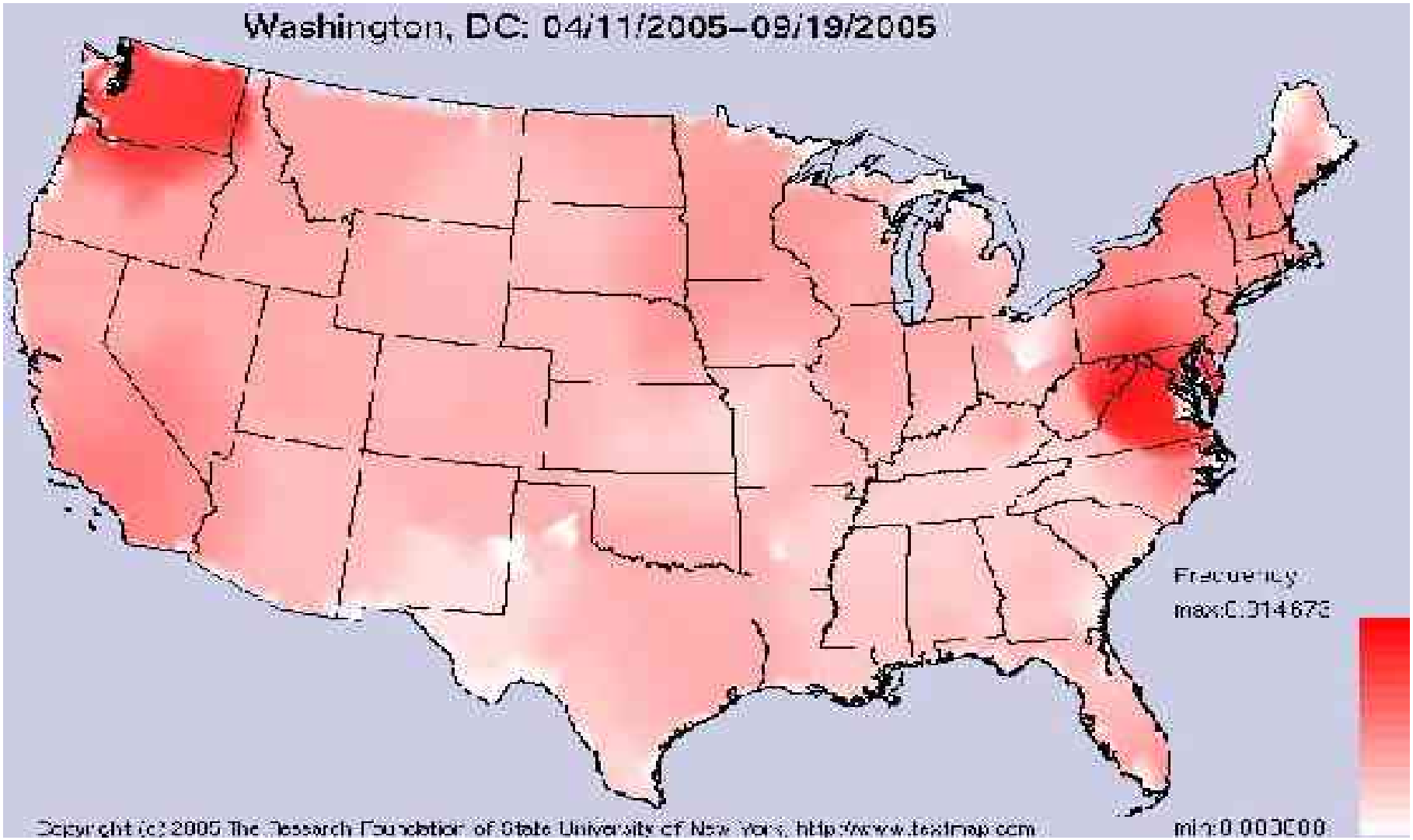
Donde Esta Mexico?



Who is running for president?



Where is Washington?



Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions

Blog Analysis with Lydia

- Blogs represent a different view of the world than newspapers.
  - Less objective
  - Greater diversity of topics
- We adapted Lydia to process *Livejournal* blogs, and compared blog content to that of newspapers.

[1] Levon Lloyd, Prachi Kaulgud, and Steven Skiena. News vs. Blogs: Who Gets the Scoop?. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs.*

Blog-specific Processing

- Blogs prove more difficult to analyze than professional news sources:
- Inconsistent capitalization
  - Some write in uniform case.
  - Some convey emotions through capitalization (I'm SO excited)
- Emoticons and unique abbreviations
  - ;), :(, etc.
  - b4 = before, 2nite = tonight, etc.

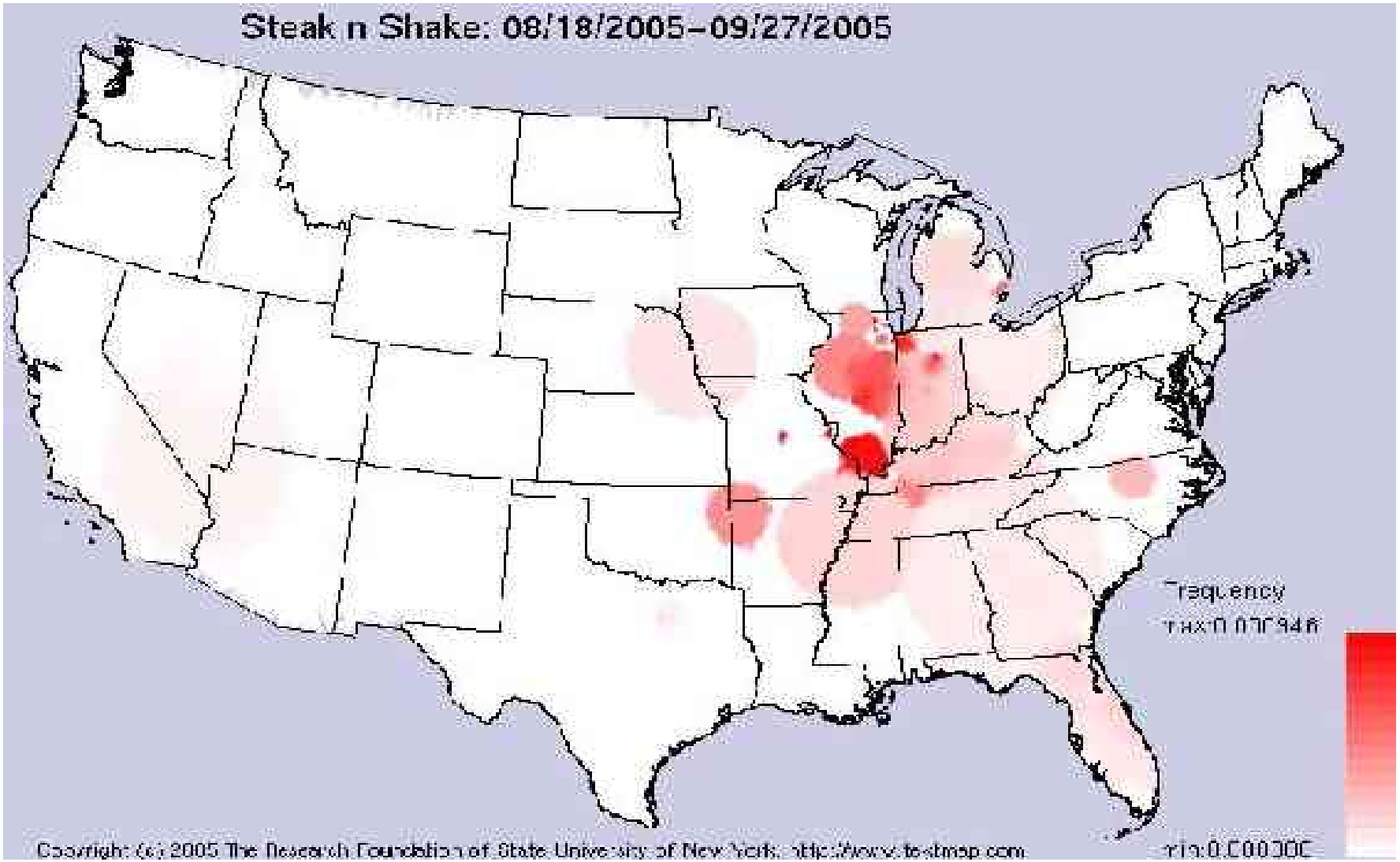
Co-reference Sets in Blogs

- Bloggers frequently misspell entity names.
- Using our co-reference set system, we can identify and correct such misspellings.
- Examples include
  - Britney Spears, Brittany Spears
  - Michael Jackson, Micheal Jackson
  - Stephen King, Steven King

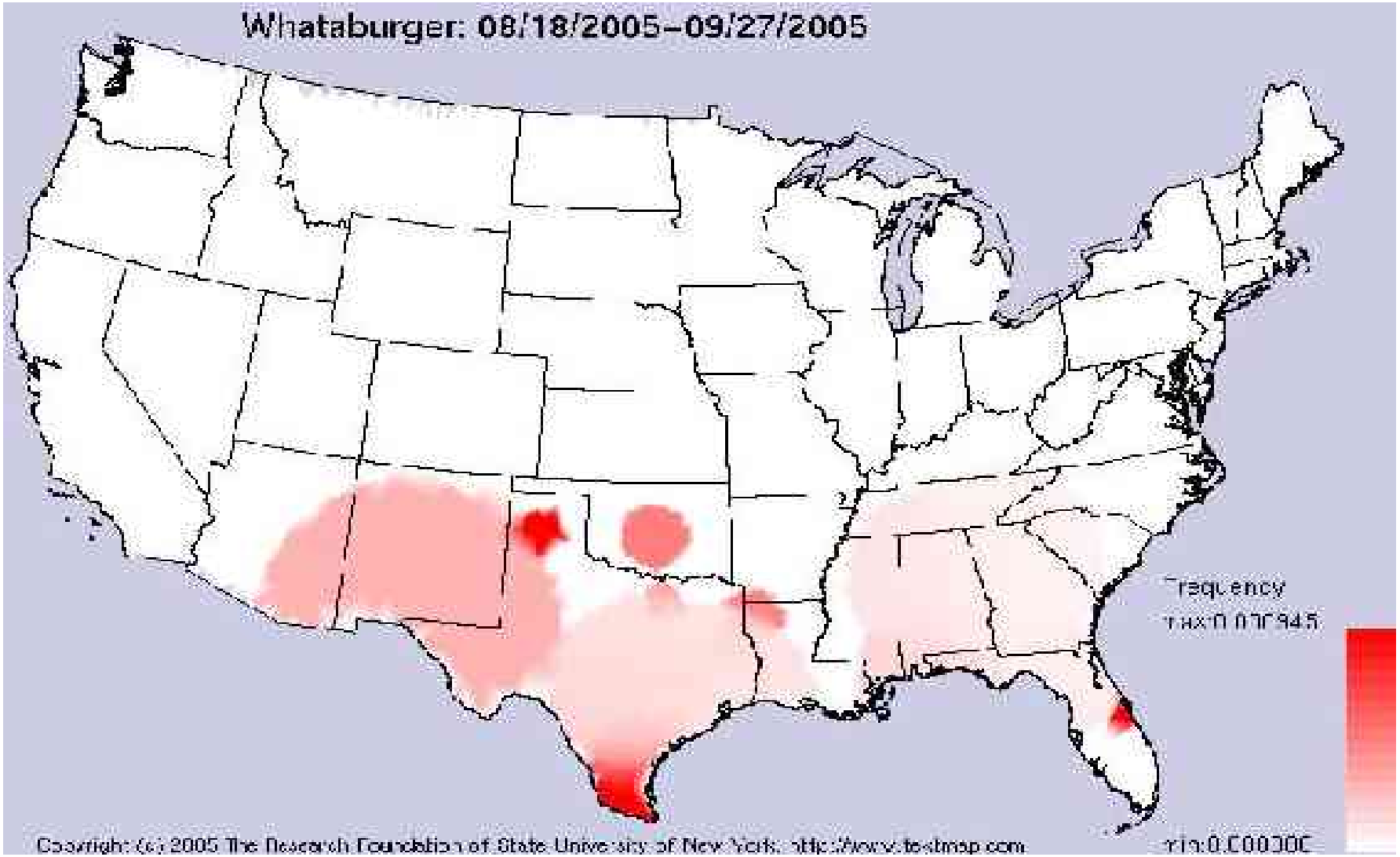
Heatmaps for Blogs

- Roughly half of all Livejournal users specify their geographic location.
- Modelling the geographic influence for blogs is harder than newspapers
  - Many more sources/locations
  - No published ratings of source power
  - Individuals need not have local constituencies

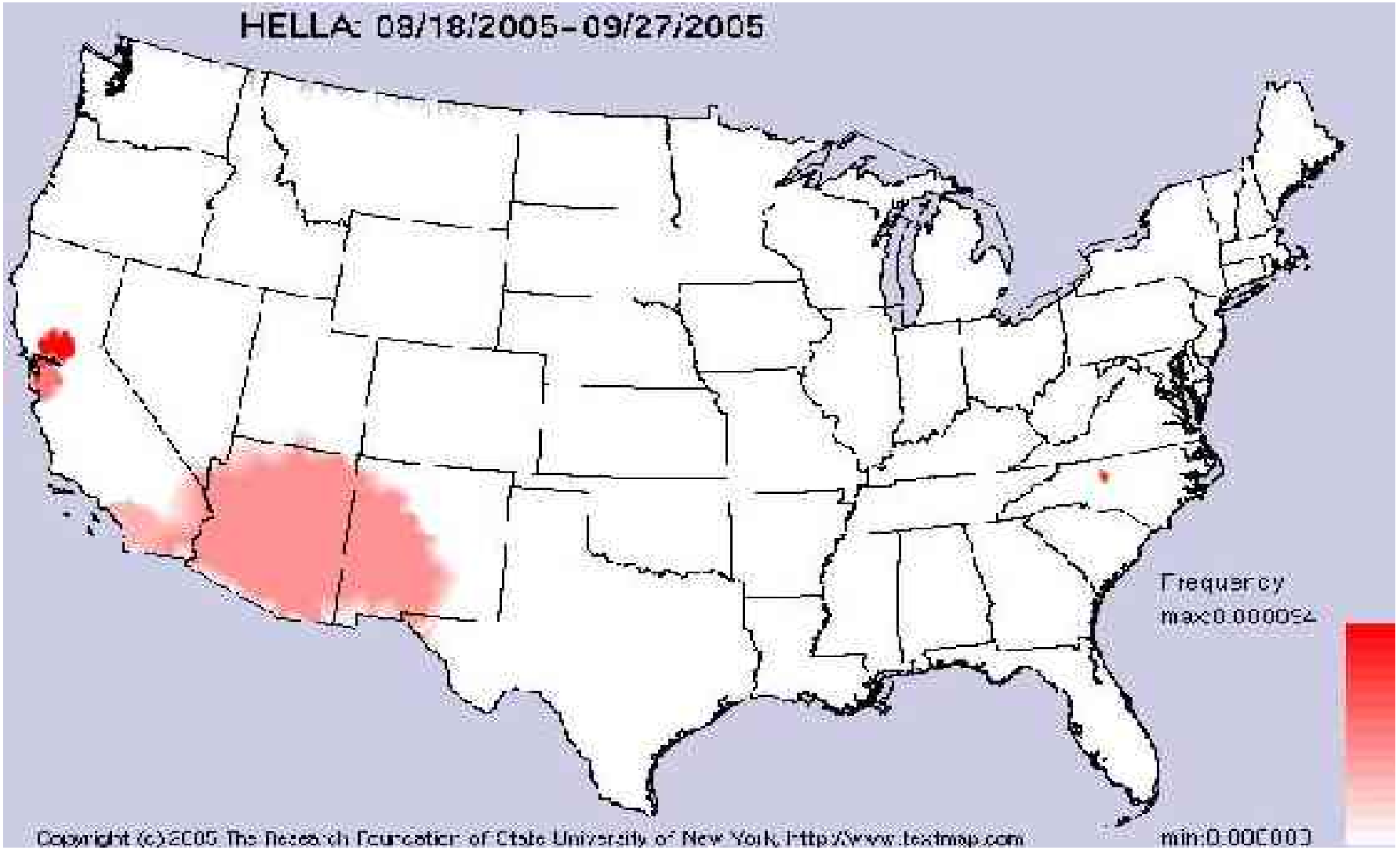
Where's the nearest Steak n Shake?



What about Whataburger?



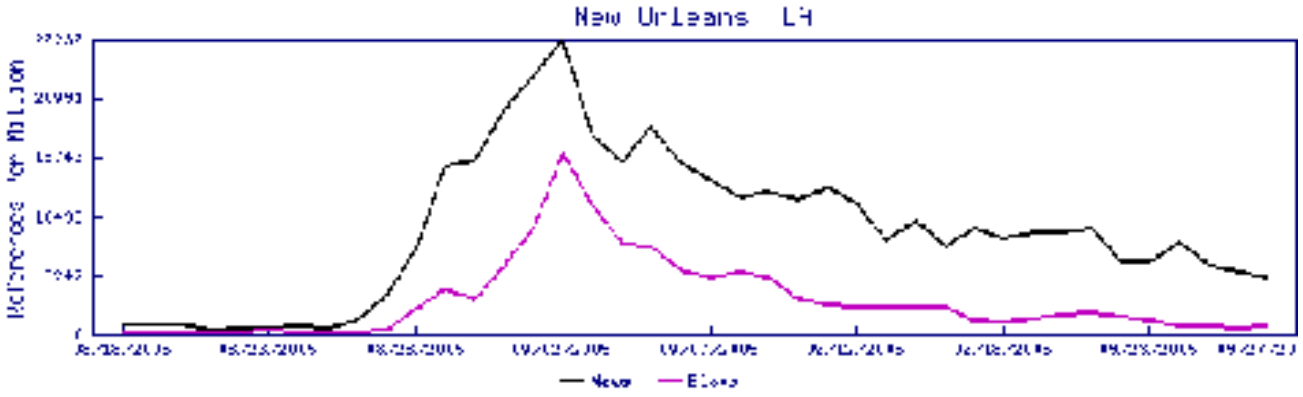
Local Dialect: Who says Hella?



Popular Entities in News vs. Blogs

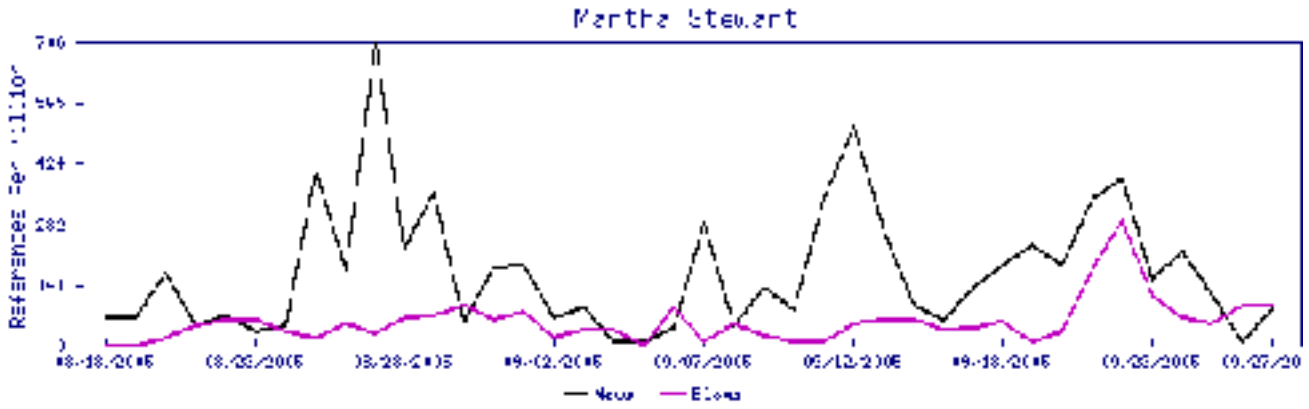
Top People in News			Top People In Blogs		
Rank In News	Rank In Blogs	Name	Rank In Blogs	Rank In News	Name
1	2	George Bush	1	380	Harry Potter
2	48	John Roberts	2	1	George Bush
3	2498	Ray Nagin	3	359	Britney Spears
4	7	Michael Brown	4	177	Michael Jackson
5	765	Arnold Schwarzenegger	5	421	Tim Burton
6	2975	Steve Spurrier	6	439	Kelly Clarkson
7	324	William Rehnquist	7	4	Michael Brown
8	192	Kathleen Blanco	8	16	Cindy Sheehan
9	N/A	Ariel Sharon	9	N/A	Brad Renfro
10	109	Pat Robertson	10	1921	Rick Perry

Blogs can track the news..

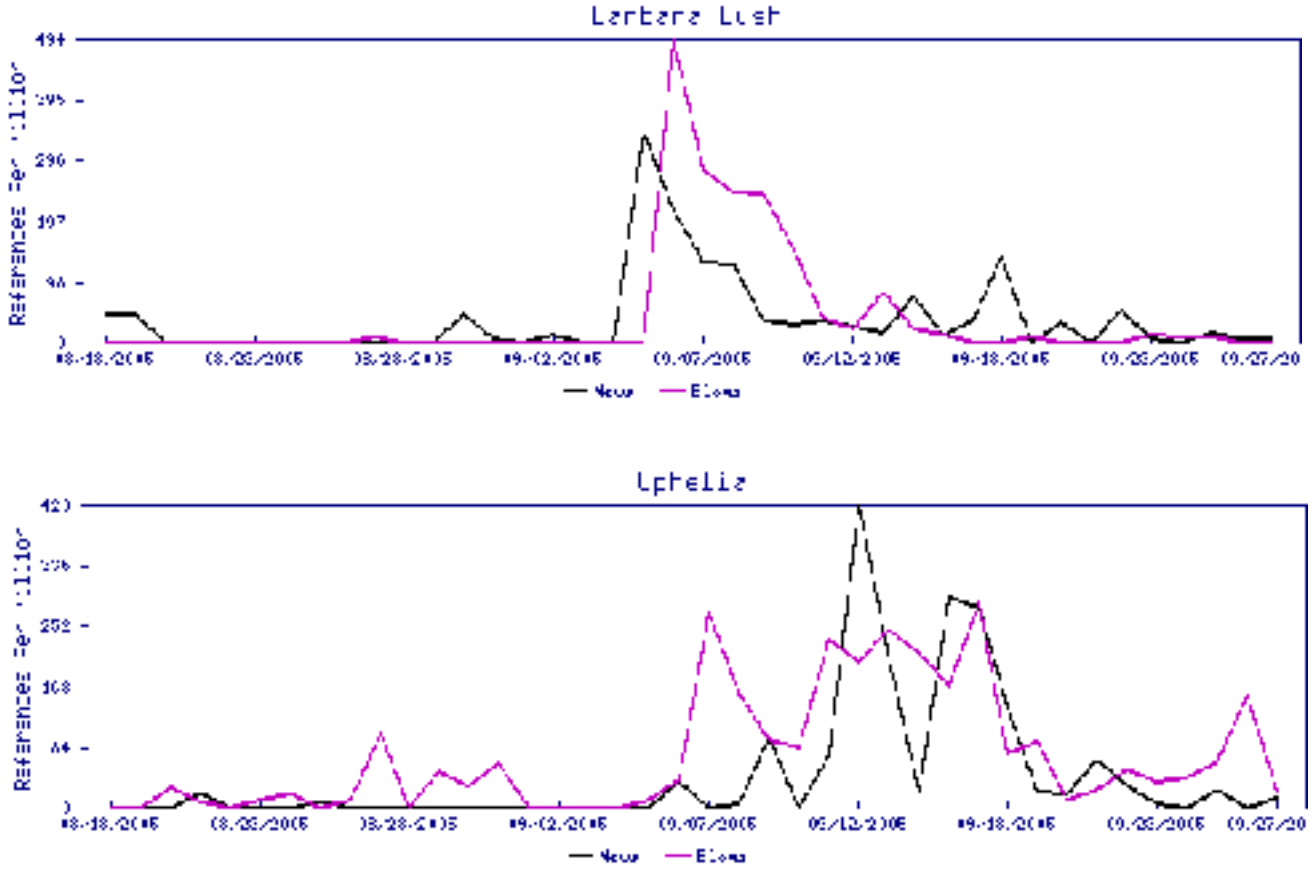


Or move independently..

- Only one of Martha Stewart's new TV shows caused a blip in blogspace.



Who Gets the Scoop?

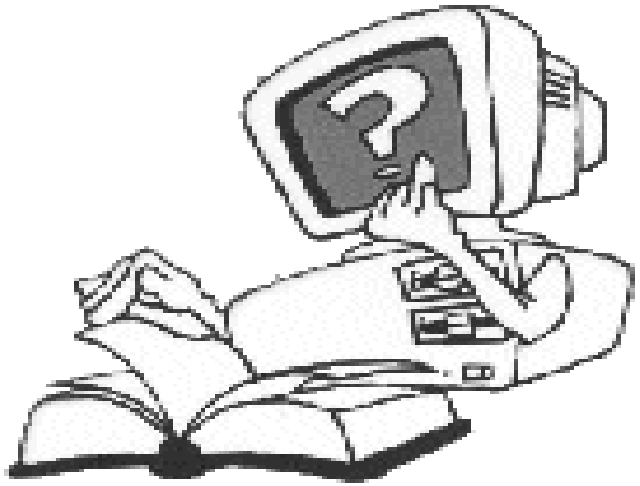


Outline of Talk

- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions

Question Answering

- Our Lydia question answering system competed in the TREC 2005 Question Answering Track, finishing with median scores using only 2000 lines of code.



Q: Where was Herbert Hoover born?

A: West Branch, Iowa

Q: Who is the Governor of Iowa?

A: Tom Vilsack

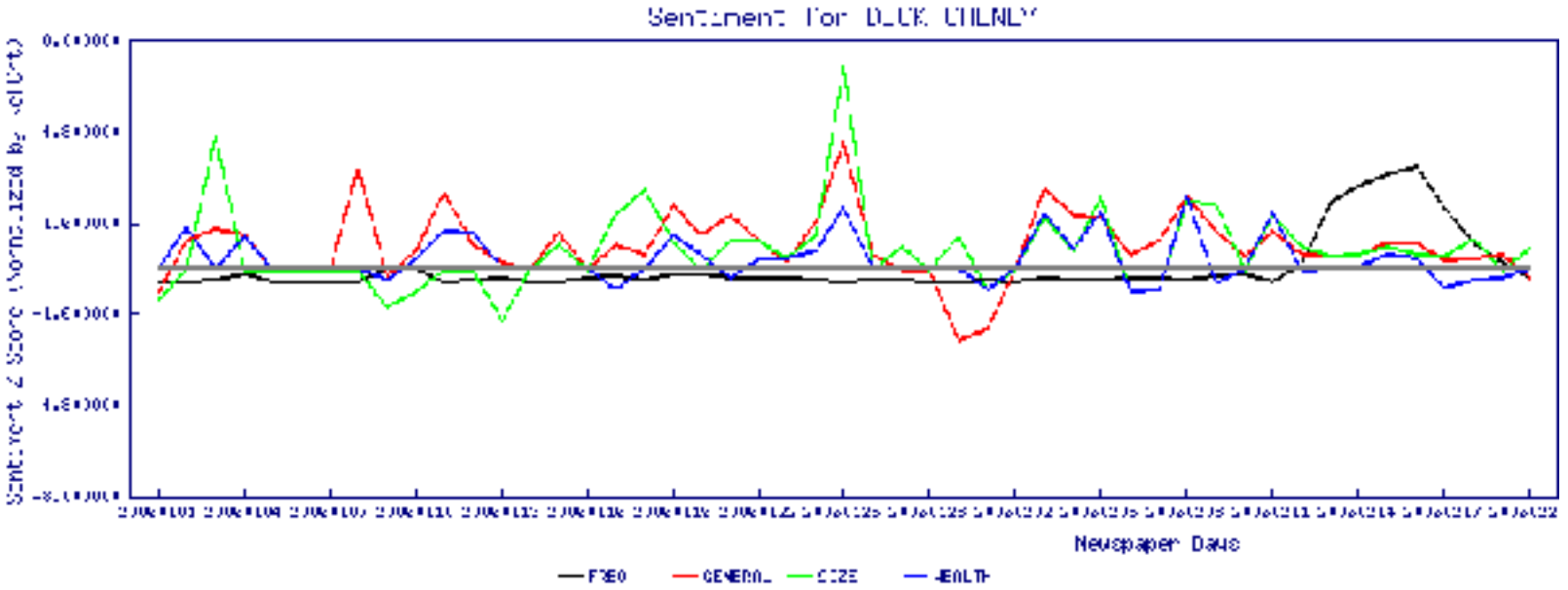
Jae Hong Kil, Levon Lloyd, and Steven Skiena. Question Answering with Lydia.
In *The Fourteenth Text Retrieval Conference Proceedings (TREC 2005).*

Description Extraction

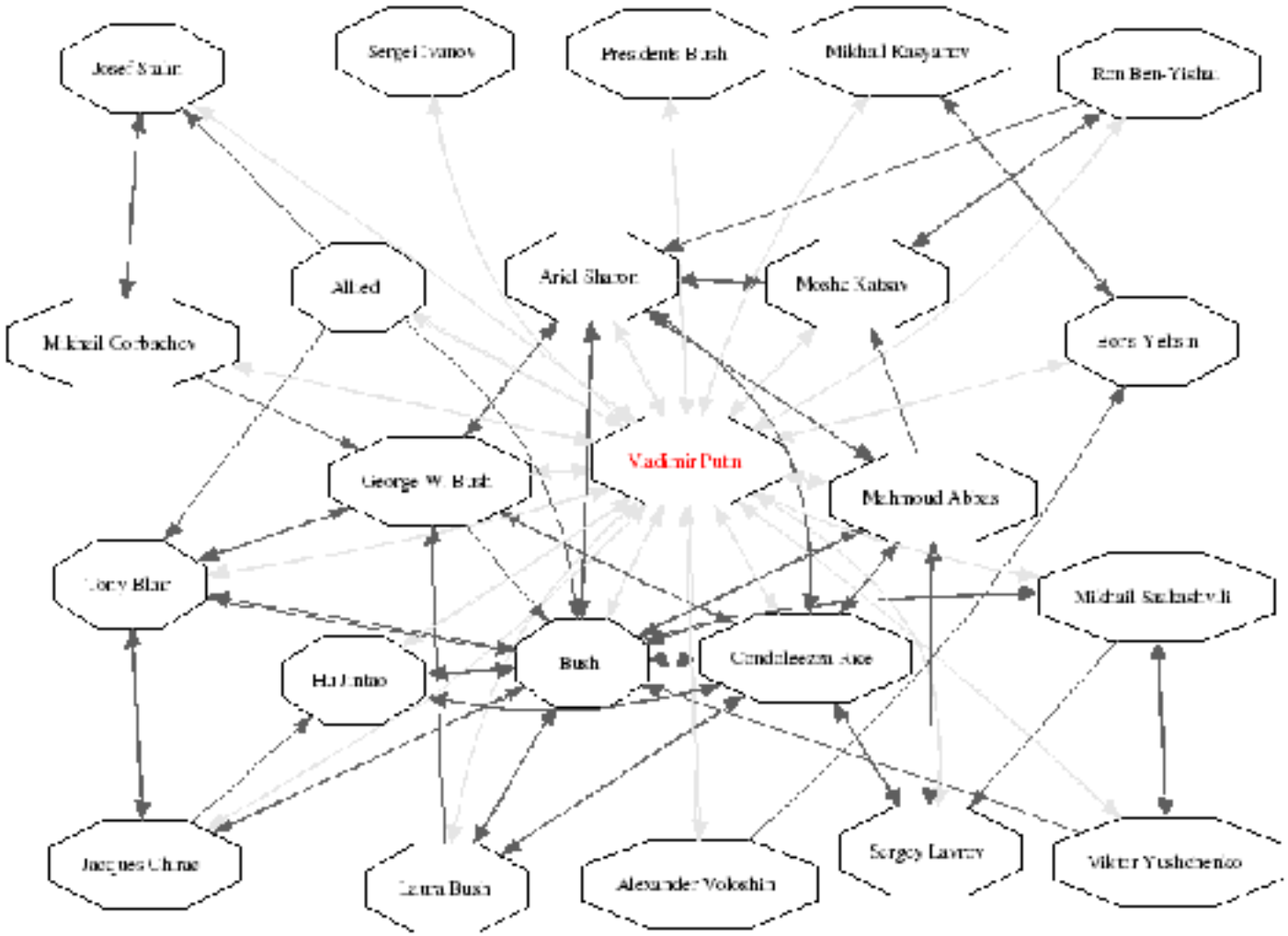
- We use template-based methods and WordNet sense analysis to extract meaningful descriptions, such as:
- Warren Buffett, billionaire investor
- Giacomo, Kentucky Derby winner
- Kim Jong Il, North Korean leader

Sentiment Analysis

- Can we measure positive/negative entity-specific vibes through adjective analysis?



Social Network Analysis



Relationship Identification

- We use verb-frames and template-based methods to try to identify the nature of statistically-significant relationships, e.g
- devastated <Hurricane Katrina:Louisiana>
- killed-in <Diana:Paris, FRA>
- became <Joseph Ratzinger:Pope Benedict XVI>
- not-watch <Dalai Lama:`` The Simpsons ">

Outline of Talk

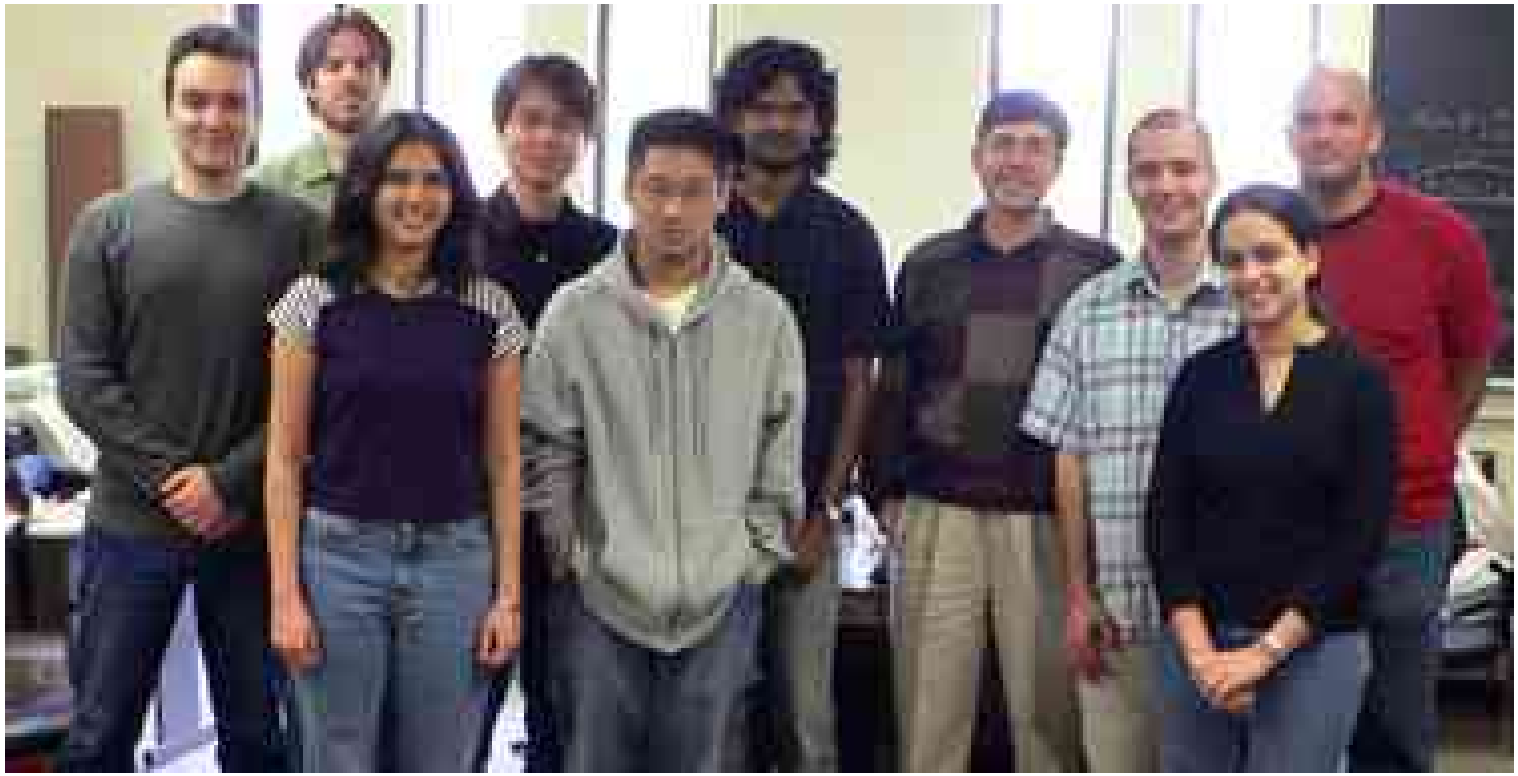
- Lydia NLP pipeline
- Spatial and temporal analysis
- Blogs vs. news
- Current research
- Future visions

Future Visions

- Entity-oriented (instead of document-based) search engines
- Market research: geographic/temporal analysis
- Legal document (deposition/filing) analysis
- Financial modelling and analysis
- Medical/Scientific applications
- Law enforcement/Homeland Security

We actively seek industrial collaboration..

The Lydia Team



The Lydia Team

- **Namrata Godbole** – NLP pipeline, sentiment analysis
- **Prachi Kaulgud** – NLP pipeline, web development
- **Dimitris Kechagias** – spidering
- **Jae Hong Kil** – question answering
- **Levon Lloyd** – systems architecture, co-reference sets
- **Andrew Mehler** – rule-based processing, heatmaps
- **Manja Srinivasaiah** – production, sentiment analysis
- **Izzet Zorlu** – web development

www.textmap.com