

Information Extraction in the Large – Part 2

CS 510 Winter 2007

1



Lydia news analysis system

- Lloyd, Kechagias, Skiena. *Lydia: A System for Large-Scale News Analysis*. SPIRE 2005.
- Lloyd, Bhagwan, Gurhl, Tomkins. *Disambiguation of References to Individuals*. IBM Research Report RJ10364.

CS 510 Winter 2007

2



Compare to NGD

- Complete system, not just a measure
- More focused: English-language news articles
- Processing of individual documents: classify articles, recognize entities
- However, still looking at corpus-wide property: co-occurrence frequency

CS 510 Winter 2007

3



Processing steps

1. Find, de-duplicate, classify articles
2. Recognize entities (people, places, titles, companies, ...)
3. Finding synonymous entities
 - I. Lewis Libby/Scooter Libby
 - But: George H.W. Bush/George W. Bush
4. Juxtaposition analysis: over-represented nearby entities
5. Analysis tools: linkage graphs, "heatmaps", temporal trends.

CS 510 Winter 2007

4

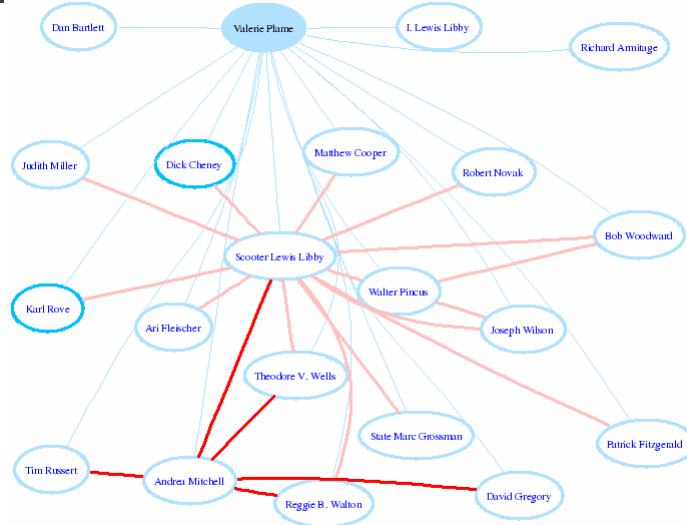


Seattle Seahawks (365 days)

Matt Hasselbeck
Shaun Alexander
San Francisco 49ers
Chicago Bears
Dallas Cowboys
Frank Gore
Seattle, WA
Dallas, TX
Jim Mora
Mike Holmgren

CS 510 Winter 2007

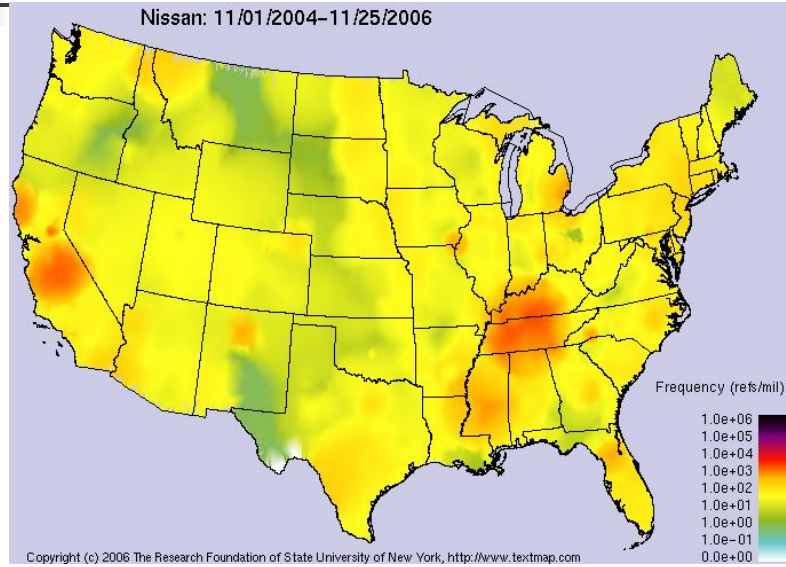
5



CS 510 Winter 2007

6

Heatmap: spatial distribution

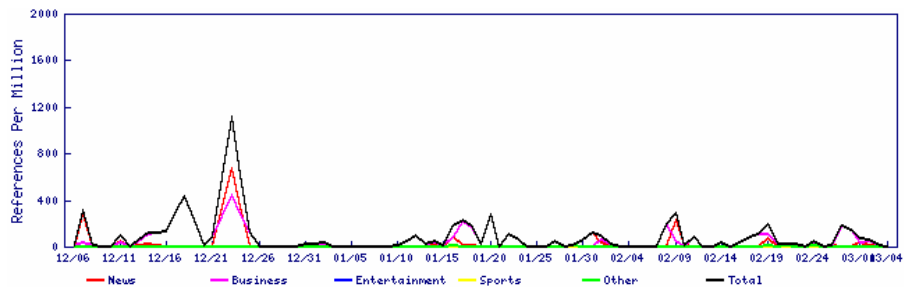


CS 510 Winter 2007

7

Temporal interest

Raul Castro (from textmap.com)



CS 510 Winter 2007

8



More on steps

- Spidering – Hard part: getting rid of parts of the page that aren't the article body but might influence linkage
 - author
 - ad (company name)
 - copyright (person or company name)
- Article classification – Bayesian classifier
 - Does okay on sports; entertainment confused often with news, business

CS 510 Winter 2007

9



Named entity resolution

- Extracting entity names
- Co-reference detection (within and between articles)
- Disambiguation
 - Howard Stern/Howard K. Stern
 - Paul Simon (senator/musician)

CS 510 Winter 2007

10



Part of speech tagging

- Break into sentences
- Use a part-of-speech tagger, based on vocabulary and rules
- Join sequences of proper nouns into proper noun phrases

CS 510 Winter 2007

11



Classifying proper noun phrases

Semantic classes: person, team, title

- Lookups
 - Place names (fairly stable)
 - Identifying suffixes, prefixes (Jr., Sr.; Prof.)
 - First, last name lists
(however, Lockheed Martin ends up a person)

CS 510 Winter 2007

12



Bayesian classifier

Only classifies small percent of occurrences with high confidence, but can combine predictions from multiple occurrences

- "Herman Miller announced today"
- "Furniture maker Howard Miller"
- "an employee of the Herman Miller furniture company"

CS 510 Winter 2007

13



Additional techniques

- Rules: "University of Portland"
- Aliasing in a document
 - "Dick Cheney" → "Cheny"
 - "Seattle Seahawks" → "Seahawks"
 - "Sun Microsystems" → "Sun"
- Rules for ambiguous geographic names, such as "Portland"
 - Is there a town local to the paper?
 - Which is the most populous

CS 510 Winter 2007

14



Features used in clustering

Seems to be used for both co-reference detection and disambiguation

1. Keywords in 100-word snippet
2. Other recognized names anywhere in the page
3. Descriptions (normalized)
"District Attorney Mark Hurlbert"
district attorney → lawyer (WordNet)

When used for classification, try to tune for precision.

CS 510 Winter 2007

15



Juxtaposition score

Log-inverse-likelihood of the number of co-occurrences appearing by chance given the frequency of each term.

CS 510 Winter 2007

16