# Extracting Semantic Themes with Topic Models

Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine

Collaborators:
Tom Griffiths, UC Berkeley
Padhraic Smyth, UC Irvine

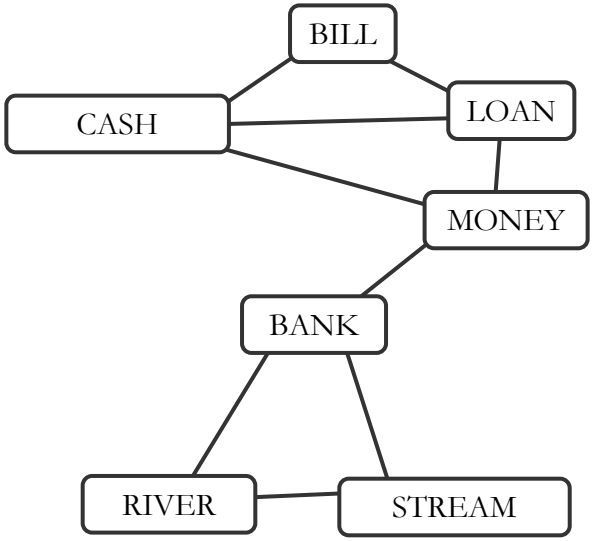
# Research on Semantic Representations

- Cognitive Science:
  - How is information represented in semantic memory?
  - How do we retrieve relevant semantic memories?

- Text-mining/ Information Retrieval/ Machine Learning:
  - How can computers represent semantic information?
  - How can computers automatically extract semantic information from text?

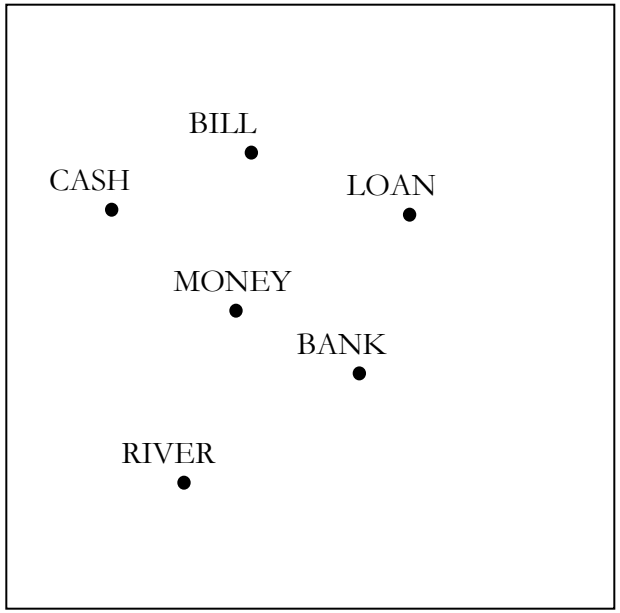
# Two approaches to semantic representation

## Semantic networks




*E.g. Spreading activation*
*Collins & Quillian (1969)*
*Collins & Loftus (1975)*

## Semantic Spaces




*Latent Semantic Analysis*
*(Landauer & Dumais, 1997)*

# New approach:
# Probabilistic Topic Models

# Topic Models

- Widely used model in machine learning and text mining
  - aka Latent Dirichlet Allocation (LDA) and pLSI

- *Automatic and unsupervised* extraction of semantic themes from large text collections.

- Essentially a Bayesian analysis of co-occurrence statistics. How are words correlated with other words across contexts?

# Topics provide quick summary of content

- What is in this corpus?
- What is in this document?
- What does this person/group of people write about?
- What are the topical trends over time?

# Example:
# Topic Analysis of Search Queries

# AOL dataset

- Dataset:
  - 20,000,000+ web queries
  - 650,000+ users
- Users were given “anonymous” user-id
- Publicly available (…but controversial)

# Example query log from user #2178

| ID | Query | Date/Time | URL clicked |
|---|---|---|---|
| 2178 | dog eats uncooked pasta | 2006-05-26 15:31:56 | |
| 2178 | inducing dog vomiting | 2006-05-26 15:32:46 | http://www.twodogpress.com |
| 2178 | inducing dog vomiting | 2006-05-26 15:32:46 | http://www.canismajor.com |
| 2178 | inducing dog vomiting | 2006-05-26 15:32:46 | http://kitchen.robbiehaf.com |
| 2178 | inducing dog vomiting | 2006-05-26 15:32:46 | http://www.dog-first-aid-101.com |
| 2178 | inducing dog vomiting | 2006-05-26 15:38:36 | |
| 2178 | walmart | 2006-05-12 12:39:52 | http://www.walmart.com |
| 2178 | sears | 2006-05-12 12:44:22 | http://www.sears.com |
| 2178 | target | 2006-05-12 17:05:36 | http://www.target.com |
| 2178 | babycenter.com | 2006-05-12 17:43:59 | http://www.babycenter.com |
| 2178 | google | 2006-05-16 10:54:39 | http://www.google.com |
| 2178 | fit pregnancy | 2006-05-16 15:34:23 | |
| 2178 | baby center | 2006-05-16 15:37:22 | |
| 2178 | yahoo.com | 2006-05-18 17:11:05 | http://www.yahoo.com |
| 2178 | applebee's carside | 2006-05-19 19:21:08 | http://www.applebees.com |
| 2178 | baby names | 2006-05-20 15:02:38 | http://www.babynames.com |
| 2178 | baby names | 2006-05-20 15:02:38 | http://www.babynamesworld.com |
| 2178 | baby names | 2006-05-20 15:02:38 | http://www.thinkbabynames.com |
| 2178 | mortgage calculator | 2006-05-24 14:39:05 | http://www.bankrate.com |
| 2178 | us zip codes | 2006-05-25 21:26:47 | http://www.usps.com |
| 2178 | us zip codes | 2006-05-25 21:26:47 | http://www.usps.com |

# Query Topic Model

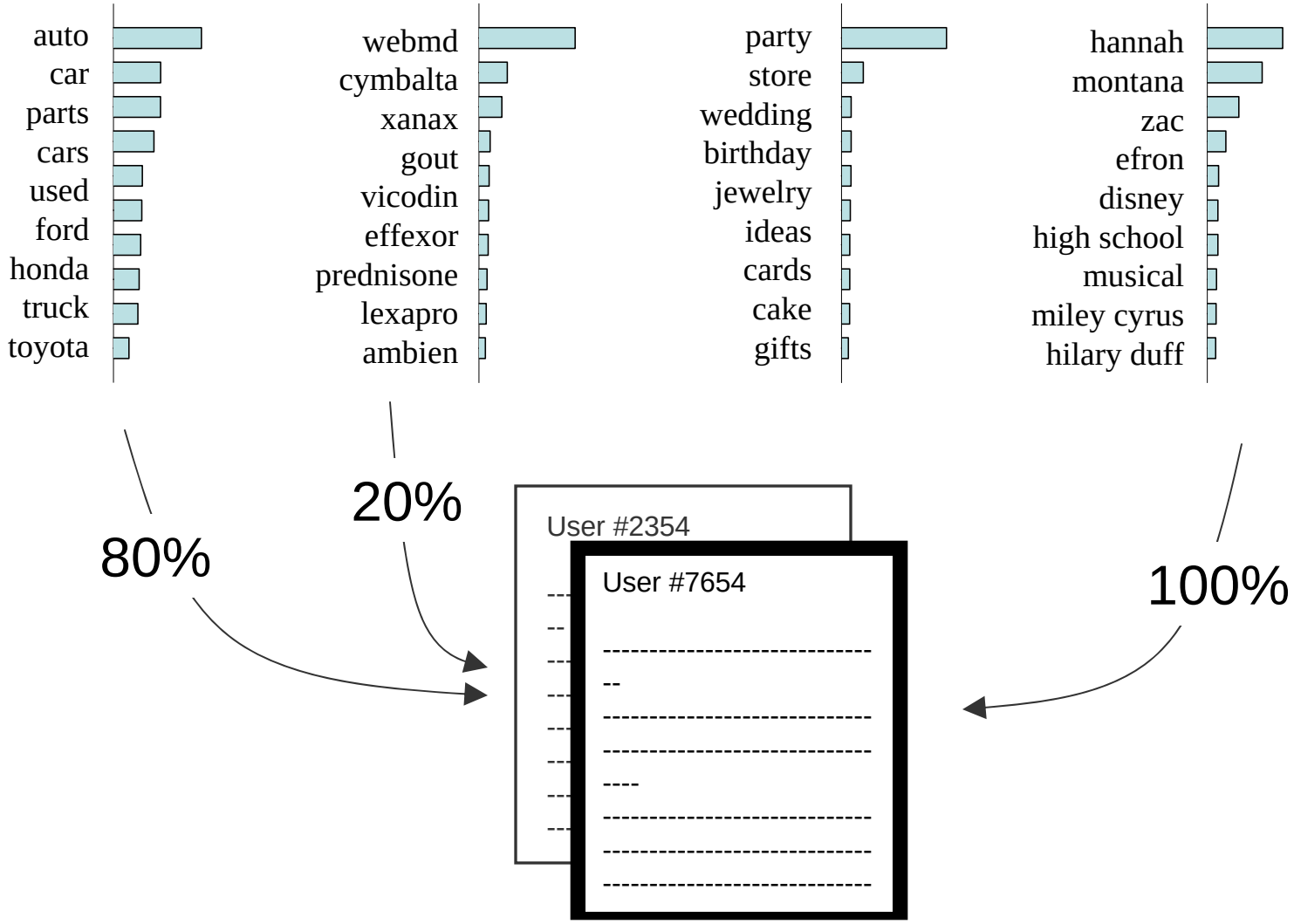
- Each *user* searches for a *mixture of topics*
- Each *topic* is a probability distribution over query words
- Simultaneously solve for all unknowns using efficient stochastic search (MCMC and Gibbs sampling)

# Topic = distribution over words

Four example topics (out of 200)


Probability distribution over words. Most likely words listed at the top

# User = mixture of topics

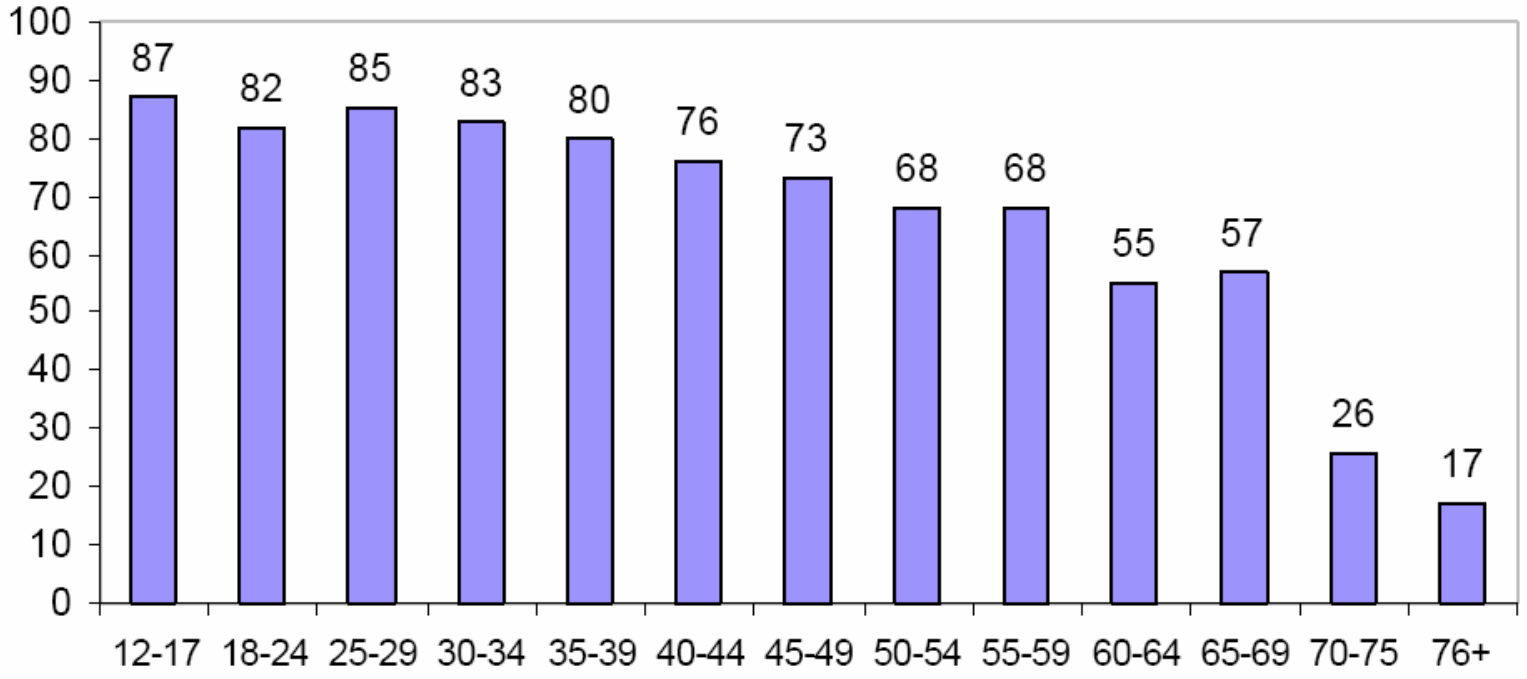


# Topic Model

- Model requires no background knowledge:
  - no dictionaries
  - no thesauri
  - no ontologies, etc

# Internet search behavior for different age groups

# Share of Americans online by age




Source: *PEW Internet & American Life Project* (2005)

# Survey of Online Activities by Age Group

| Generational Differences in Online Activities | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Online Teens[a] (12-17) | Gen Y (18-28) | Gen X (29-40) | Trailing Boomers (41-50) | Leading Boomers (51-59) | Matures (60-69) | After Work (70+) | All Online Adults[b] |
| Go online | 87% | 84% | 87% | 79% | 75% | 54% | 21% | 72% |
| *Teens and Gen Y are more likely to engage in the following activities compared with older users:* | | | | | | | | |
| Online games | 81 | 54 | 37 | 29 | 25 | 25 | 32 | 36 |
| School research | * | 73 | 60 | 61 | 48 | 33 | 14 | 57 |
| Instant message | 75 | 66 | 52 | 38 | 42 | 33 | 25 | 47 |
| Text message | * | 60 | 44 | 29 | 15 | 11 | 8 | 35 |
| Get info about a school | 57 | 59 | 42 | 50 | 40 | 30 | 14 | 45 |
| Download music | 51 | 45 | 28 | 16 | 14 | 8 | 5 | 25 |
| Read blogs | 38 | 41 | 30 | 20 | 21 | 19 | 16 | 27 |
| Download video | 31 | 27 | 22 | 14 | 8 | 8 | 1 | 18 |
| Create a blog | 19 | 20 | 9 | 3 | 9 | 3 | 4 | 9 |

Source: *PEW Internet & American Life Project* (2005)

# Alternative to Survey Analysis

- PEW surveys presuppose you already know the questions to ask for

- Alternative:
  - automatically topic analyze the search terms people enter into a search engine
  - Gives a more direct picture of the user goals and needs

(will incorporate result slides during presentation)

# Conclusions

- Topic modeling coupled with demographic analysis yields “windows” into the minds of different age groups

- Other potential applications:
  - clinical data, e.g. therapy discussions
  - open question surveys
  - internet behavior, e.g. chatting, SMS, email

# Open Questions

- How can research community get access to data from internet providers/ search engines?

- Privacy issues

# Additional Slides

# Example Topics from New York Times

| Terrorism | Wall Street Firms | Stock Market | Bankruptcy |
|---|---|---|---|
| SEPT_11 | WALL_STREET | WEEK | BANKRUPTCY |
| WAR | ANALYSTS | DOW_JONES | CREDITORS |
| SECURITY | INVESTORS | POINTS | BANKRUPTCY_PROTECTION |
| IRAQ | FIRM | 10_YR_TREASURY_YIELD | ASSETS |
| TERRORISM | GOLDMAN_SACHS | PERCENT | COMPANY |
| NATION | FIRMS | CLOSE | FILED |
| KILLED | INVESTMENT | NASDAQ_COMPOSITE | BANKRUPTCY_FILING |
| AFGHANISTAN | MERRILL_LYNCH | STANDARD_POOR | ENRON |
| ATTACKS | COMPANIES | CHANGE | BANKRUPTCY_COURT |
| OSAMA_BIN_LADEN | SECURITIES | FRIDAY | KMART |
| AMERICAN | RESEARCH | DOW_INDUSTRIALS | CHAPTER_11 |
| ATTACK | STOCK | GRAPH_TRACKS | FILING |
| NEW_YORK_REGION | BUSINESS | EXPECTED | COOPER |
| NEW | ANALYST | BILLION | BILLIONS |
| MILITARY | WALL_STREET_FIRMS | NASDAQ_COMPOSITE_INDEX | COMPANIES |
| NEW_YORK | SALOMON_SMITH_BARNEY | EST_02 | BANKRUPTCY_PROCEEDINGS |
| WORLD | CLIENTS | PHOTO_YESTERDAY | DEBTS |
| NATIONAL | INVESTMENT_BANKING | YEN | RESTRUCTURING |
| QAEDA | INVESTMENT_BANKERS | 10 | CASE |
| TERRORIST_ATTACKS | INVESTMENT_BANKS | 500_STOCK_INDEX | GROUP |

# Algorithm input/output

**INPUT:** word-document counts (word order is irrelevant)

**OUTPUT:**

| | |
|---|---|
| topic assignments to each word | $P( z_i )$ |
| likely words in each topic | $P( w \mid z )$ |
| likely topics in each document (“gist”) | $P( \theta \mid d )$ |

# Generative Process

- For each document, choose a mixture of topics

  $\theta \sim \text{Dirichlet}(\alpha)$

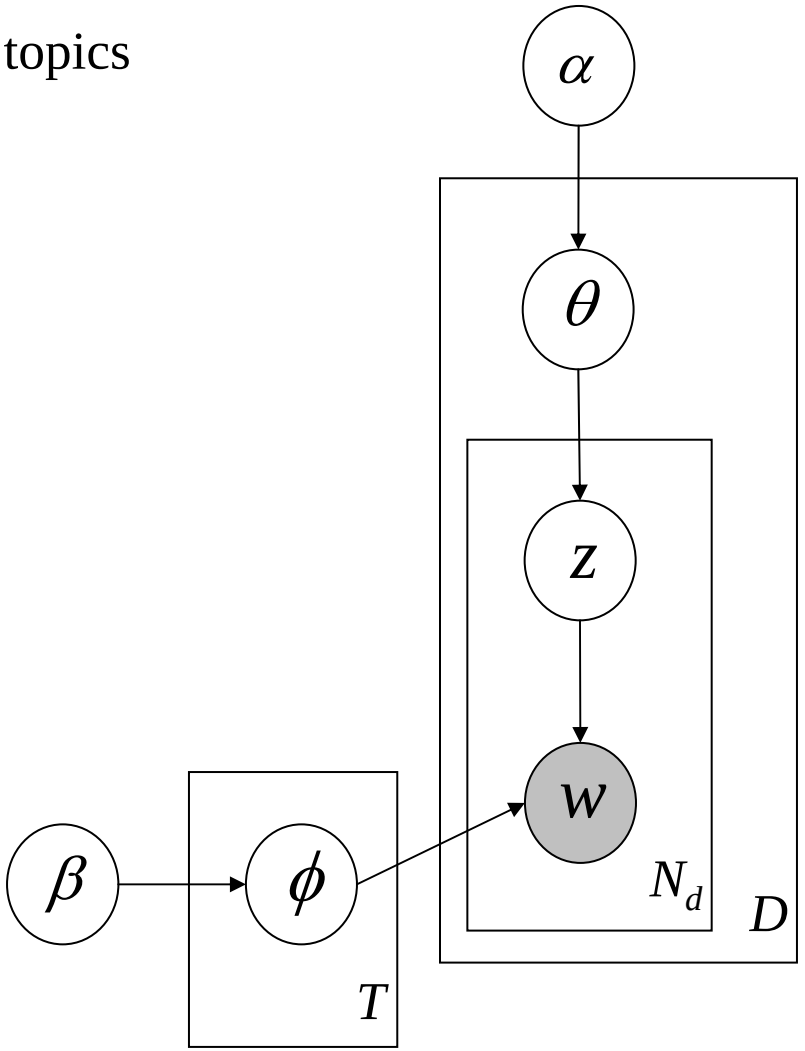
- Sample a topic [1..T] from the mixture

  $z \sim \text{Multinomial}(\theta)$

- Sample a word from the topic

  $w \sim \text{Multinomial}(\phi^{(z)})$

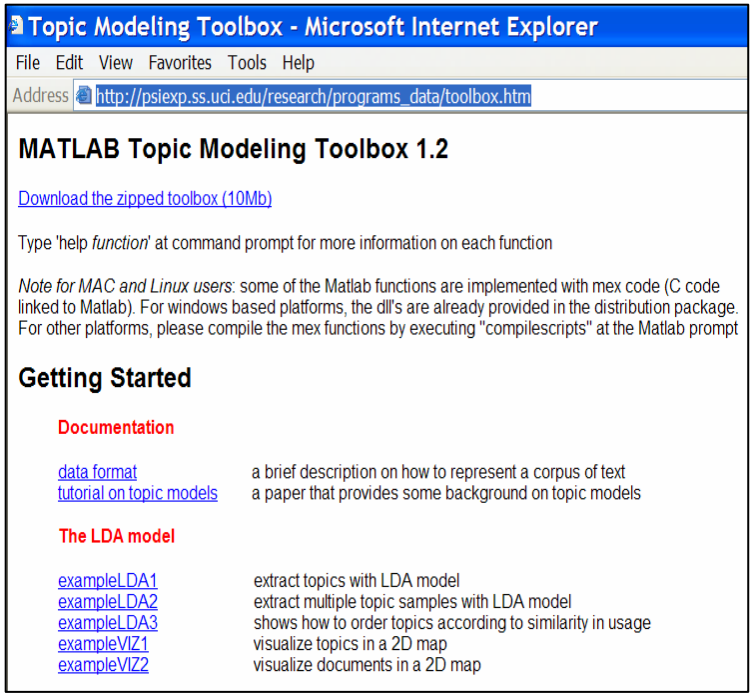
  $\phi \sim \text{Dirichlet}(\beta)$



# Software

Public-domain MATLAB toolbox for topic modeling on the Web:

http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm



# Choosing number of topics

- Bayesian model selection
- Generalization test
  - e.g., perplexity on out-of-sample data
- Non-parametric Bayesian approach
  - Number of topics grows with size of data
  - E.g. Hierarchical Dirichlet Processes (HDP)

# Polysemy

PRINTING
PAPER
PRINT
PRINTED
TYPE
PROCESS
INK
PRESS
IMAGE
PRINTER
PRINTS
PRINTERS
COPY
COPIES
FORM
OFFSET
GRAPHIC
SURFACE
PRODUCED
**CHARACTERS**

**PLAY**
PLAYS
STAGE
AUDIENCE
THEATER
ACTORS
DRAMA
SHAKESPEARE
ACTOR
THEATRE
PLAYWRIGHT
PERFORMANCE
DRAMATIC
COSTUMES
COMEDY
TRAGEDY
**CHARACTERS**
SCENES
OPERA
PERFORMED

TEAM
GAME
BASKETBALL
PLAYERS
PLAYER
**PLAY**
PLAYING
SOCCER
PLAYED
BALL
TEAMS
BASKET
FOOTBALL
SCORE
**COURT**
GAMES
TRY
COACH
GYM
SHOT

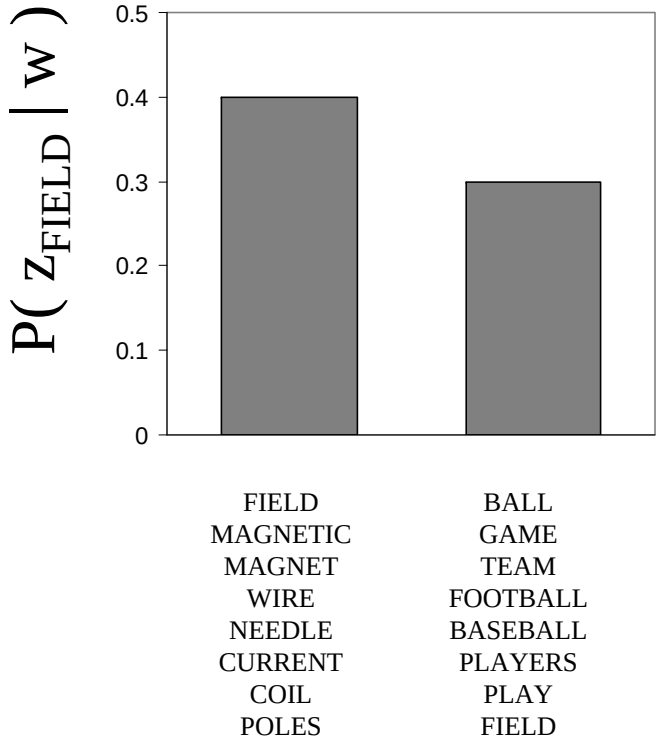
JUDGE
TRIAL
**COURT**
CASE
JURY
ACCUSED
GUILTY
DEFENDANT
JUSTICE
**EVIDENCE**
WITNESSES
CRIME
LAWYER
WITNESS
ATTORNEY
HEARING
INNOCENT
DEFENSE
CHARGE
CRIMINAL

HYPOTHESIS
EXPERIMENT
SCIENTIFIC
OBSERVATIONS
SCIENTISTS
EXPERIMENTS
SCIENTIST
EXPERIMENTAL
**TEST**
METHOD
HYPOTHESES
TESTED
**EVIDENCE**
BASED
OBSERVATION
SCIENCE
FACTS
DATA
RESULTS
EXPLANATION

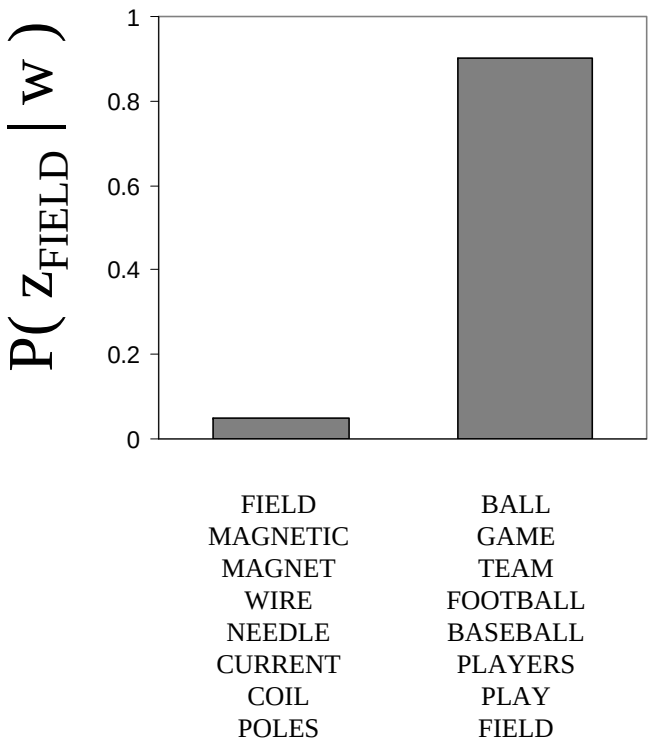
STUDY
**TEST**
STUDYING
HOMEWORK
NEED
CLASS
MATH
TRY
TEACHER
WRITE
PLAN
ARITHMETIC
ASSIGNMENT
PLACE
STUDIED
CAREFULLY
DECIDE
IMPORTANT
NOTEBOOK
REVIEW

# Disambiguation

"FIELD"




"FOOTBALL FIELD"



# Recent Papers


- Steyvers, M., Griffiths, T.L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences, 10(7), 327-334.*

- Griffiths, T.L., Steyvers, M., & Tenenbaum, J.B.T. (2007). Topics in Semantic Representation. *Psychological Review*, 114(2), 211-244.

- Griffiths, T.L., Steyvers, M., & Firl, A. (in press). Google and the mind: Predicting fluency with PageRank. *Psychological Science*.

- Steyvers, M. & Griffiths, T.L. (2008). Rational Analysis as a Link between Human Memory and Information Retrieval. In N. Chater and M Oaksford (Eds.) *The Probabilistic Mind: Prospects from Rational Models of Cognition*. Oxford University Press.

- Chemudugunta, C., Smyth, P., & Steyvers, M. (2007). Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In: *Advances in Neural Information Processing Systems, 19.*