Semantic Representations with Probabilistic Topic Models

Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine

Joint work with:
Tom Griffiths, UC Berkeley
Padhraic Smyth, UC Irvine

Topic Models in Machine Learning

- *Unsupervised* extraction of topics from large text collection
- Topics provide quick summary of "gist"
  - →What is in this corpus?
  - →What is in this document or paragraph?
  - →What are similar documents to a query?
  - →What are the topical trends over time?

Topic Models in Psychology

- Topic models address three computational problems for semantic memory system:

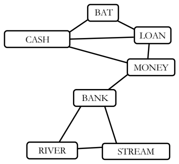
1) Gist extraction: what is this set of words about?

2) Disambiguation: what is the sense of this word?
- E.g. "football field" vs. "magnetic field"

3) Prediction: what fact, concept, or word is next?

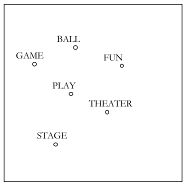
Two approaches to semantic representation

Semantic networks



How are these learned?

Semantic Spaces



Can be learned (e.g. Latent Semantic Analysis), but is this representation flexible enough?

Overview

I Probabilistic Topic Models
  generative model
  statistical inference: Gibbs sampling
II Explaining human memory
  word association
  semantic isolation
  false memory
III Information retrieval

Probabilistic Topic Models

- Extract topics from large text collections
  - → unsupervised
  - → generative
- Our modeling work is based on:
  - pLSI Model: Hoffman (1999)
  - LDA Model: Blei, Ng, and Jordan (2001, 2003)
  - Topics Model: Griffiths and Steyvers (2003, 2004)

Model input: "bag of words"

- Matrix of number of times words occur in documents

documents

words

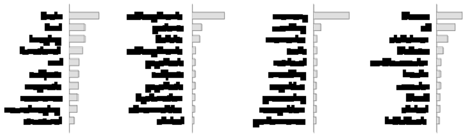
	Doc1	Doc2	Doc3 …
RIVER	34	0	0
STREAM	12	0	0
BANK	5	19	6
MONEY	0	16	1
…	…	…	…

[redacted]

- Note: some function words are deleted: "the", "a", "and", etc

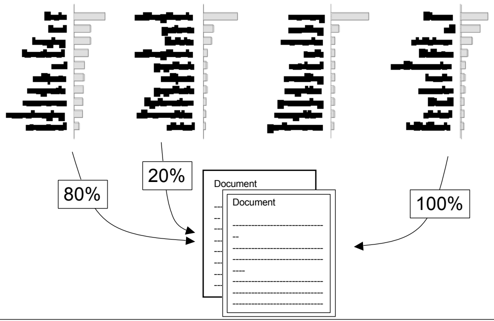
Probabilistic Topic Models

- A topic represents a probability distribution over words
  - Related words get high probability in same topic
- Example topics extracted from NIH/NSF grants:



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

Document = mixture of topics



Generative Process

- For each document, choose a mixture of topics

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

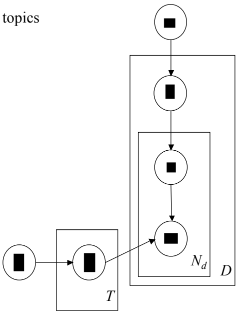
- Sample a topic [1..T] from the mixture

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

- Sample a word from the topic

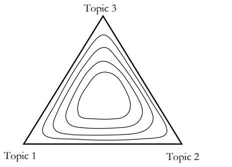
  $w \sim \text{Multinomial}(\phi^{(z)})$

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

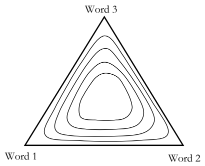


Prior Distributions

- Dirichlet priors encourage sparsity on topic mixtures and topics



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



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

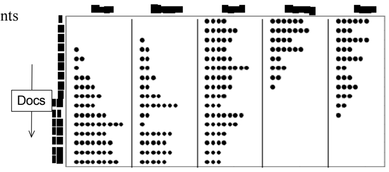
(darker colors indicate lower probability)

Creating Artificial Dataset

Two topics

	topic 1	topic 2
River	0.33	0
Stream	0.33	0
Bank	0.33	0.33
Money	0	0.33
Loan	0	0.33

16 documents



Can we recover the original topics and topic mixtures from this data?

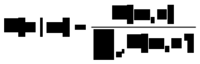
Statistical Inference

- Three sets of latent variables
  - topic mixtures θ
  - word mixtures ϕ
  - topic assignments z
- Estimate posterior distribution over topic assignments
  - P(z | **w**)

(we can later infer θ and ϕ)

Statistical Inference

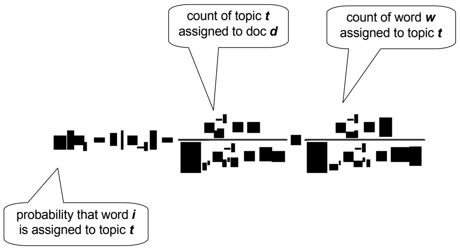
- Exact inference is impossible



Sum over T^n terms

- Use approximate methods:
  - Markov chain Monte Carlo (MCMC) with Gibbs sampling

Gibbs Sampling



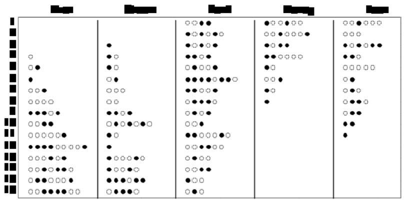
count of topic ***t*** assigned to doc ***d***

count of word ***w*** assigned to topic ***t***

probability that word ***i*** is assigned to topic ***t***

Example of Gibbs Sampling

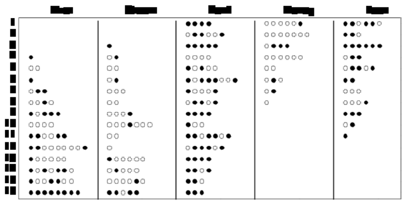
- Assign word tokens randomly to topics:



(●=topic 1; ○=topic 2)

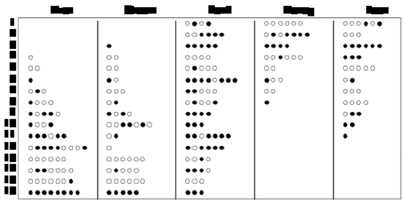
After 1 iteration

- Apply sampling equation to each word token:



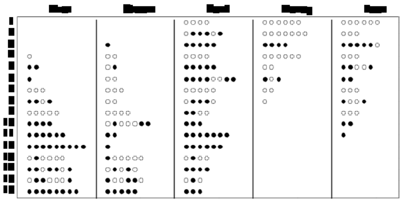
(●=topic 1; ○=topic 2)

After 4 iterations



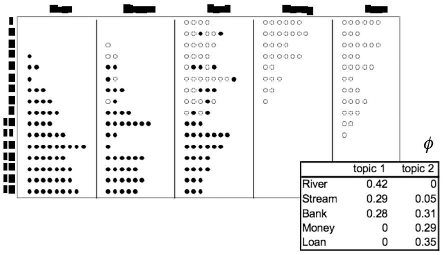
(●=topic 1; ○=topic 2)

After 8 iterations



(●=topic 1; ○=topic 2)

After 32 iterations



ϕ

	topic 1	topic 2
River	0.42	0
Stream	0.29	0.05
Bank	0.28	0.31
Money	0	0.29
Loan	0	0.35

(●=topic 1; ○=topic 2)

Algorithm input/output

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

OUTPUT:

topic assignments to each word $P(z_i)$

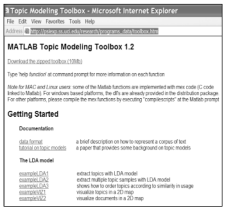
likely words in each topic $P(w \mid z)$

likely topics in each document ("gist") $P(\theta \mid d)$

Software

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

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



Topic Modeling Toolbox - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address

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

MATLAB Topic Modeling Toolbox 1.2

Download the zipped toolbox (10Mb)

Type 'help *function*' at command prompt for more information on each function

Note for MAC and Linux users: some of the Matlab functions are implemented with mex code (C code linked to Matlab). For windows based platforms, the dll's are already provided in the distribution package. For other platforms, please compile the mex functions by executing "compilescripts" at the Matlab prompt

Getting Started

Documentation

data format	a brief description on how to represent a corpus of text
tutorial on topic models	a paper that provides some background on topic models

The LDA model

exampleLDA1	extract topics with LDA model
exampleLDA2	extract multiple topic samples with LDA model
exampleLDA3	shows how to order topics according to similarity in usage
exampleVIZ1	visualize topics in a 2D map
exampleVIZ2	visualize documents in a 2D map

Examples 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

Example topics from an educational corpus

PRINTING	PLAY	TEAM	JUDGE	HYPOTHESIS	STUDY
PAPER	PLAYS	GAME	TRIAL	EXPERIMENT	TEST
PRINT	STAGE	BASKETBALL	COURT	SCIENTIFIC	STUDYING
PRINTED	AUDIENCE	PLAYERS	CASE	OBSERVATIONS	HOMEWORK
TYPE	THEATER	PLAYER	JURY	SCIENTISTS	NEED
PROCESS	ACTORS	PLAY	ACCUSED	EXPERIMENTS	CLASS
INK	DRAMA	PLAYING	GUILTY	SCIENTIST	MATH
PRESS	SHAKESPEARE	SOCCER	DEFENDANT	EXPERIMENTAL	TRY
IMAGE	ACTOR	PLAYED	JUSTICE	TEST	TEACHER

Example topics from psych review abstracts

SIMILARITY	STIMULUS	MEMORY	GROUP	EMOTIONAL
CATEGORY	CONDITIONING	RETRIEVAL	INDIVIDUAL	EMOTION
CATEGORIES	LEARNING	RECALL	GROUPS	BASIC
RELATIONS	RESPONSE	ITEMS	OUTCOMES	EMOTIONS
DIMENSIONS	STIMULI	INFORMATION	INDIVIDUALS	AFFECT
FEATURES	RESPONSES	TERM	GROUPS	STATES
STRUCTURE	AVOIDANCE	RECOGNITION	OUTCOMES	EXPERIENCES
SIMILAR	REINFORCEMENT	ITEMS	INDIVIDUALS	AFFECTIVE
REPRESENTATION	CLASSICAL	LIST	DIFFERENCES	AFFECTS
ONJECTS	DISCRIMINATION	ASSOCIATIVE	INTERACTION	RESEARCH

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)

Applications to Human Memory

Computational Problems for Semantic Memory System

- Gist extraction
  - What is this set of words about?

$P(w_2 | w_1)$

- Disambiguation
  - What is the sense of this word?

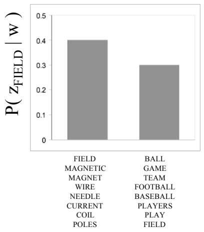
$P(z | w)$

- Prediction
  - what fact, concept, or word is next?

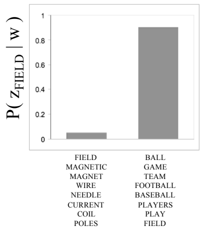
$P(z | w, context)$

Disambiguation

“FIELD”



“FOOTBALL FIELD”



Modeling Word Association

Word Association

(norms from Nelson et al. 1998)

CUE: PLANET

Word Association

(norms from Nelson et al. 1998)

CUE: PLANET

associate number	people
1	EARTH
2	STARS
3	SPACE
4	SUN
5	MARS
6	UNIVERSE
7	SATURN
8	GALAXY

(vocabulary = 5000+ words)

Word Association as a Prediction Problem

- Given that a single word is observed, predict what other words might occur in that context
- Under a single topic assumption:

$$P(w_2 | w_1) = \sum_z P(w_2 | z) P(z | w_1)$$

Response

Cue

Word Association

(norms from Nelson et al. 1998)

CUE: PLANET

associate number	people	model
1	EARTH	STARS
2	STARS	STAR
3	SPACE	SUN
4	SUN	EARTH
5	MARS	SPACE
6	UNIVERSE	SKY
7	SATURN	PLANET
8	GALAXY	UNIVERSE

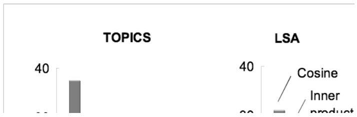
First associate "EARTH" has rank 4 in model

Median rank of first associate

TOPICS

Median rank of first associate

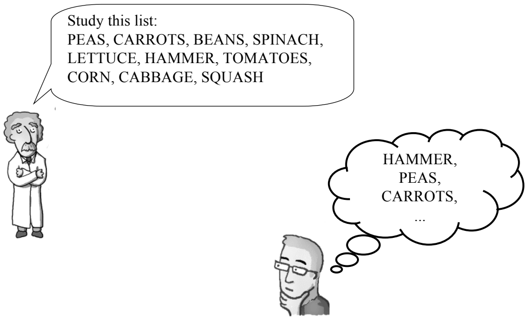
TOPICS LSA



Episodic Memory

Semantic Isolation Effects
False Memory

Semantic Isolation Effect



Semantic isolation effect / Von Restorff effect

- Finding: contextually unique words are better remembered
- Verbal explanations:
  - Attention, surprise, distinctiveness
- Our approach:
  - assume memories can be accessed and encoded at multiple levels of description
    - Semantic/ Gist aspects – generic information
    - Verbatim – specific information

Computational Problem

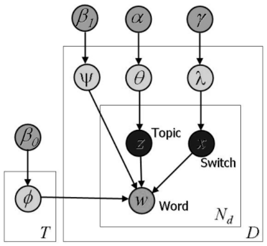
- How to tradeoff specificity and generality?
  - Remembering detail and gist
- Dual route topic model = topic model + encoding of specific words

Dual route topic model

- Two ways to generate words:
  - Topic Model
  - Verbatim word distribution (unique to document)
- Each word comes from a single route
  - Switch variable x_i for every word i:
    - x_i =0 → topics
    - x_i =1 → verbatim
- Conditional prob. of a word under a document:

[redacted]

Graphical Model



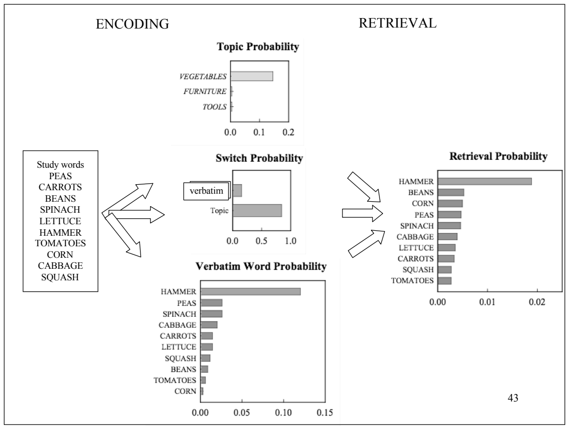
Variable x is a switch :

x=0 → sample from topic
x=1 → sample from verbatim word distribution

Applying Dual Route Topic Model to Human Memory

- Train model on educational corpus (TASA)
  - 37K documents, 1700 topics
- Apply model to list memory experiments
  - Study list is a "document"
  - Recall probability based on model

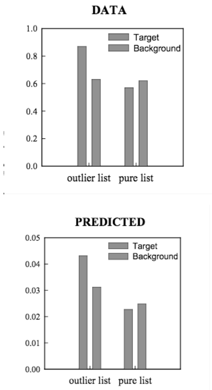
[redacted]



Study words
PEAS
CARROTS
BEANS
SPINACH
LETTUCE
HAMMER
TOMATOES
CORN
CABBAGE
SQUASH

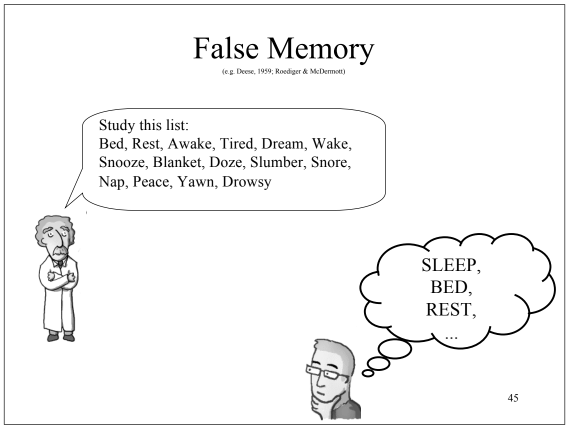
Hunt & Lamb (2001 exp. 1)

OUTLIER LIST	CONTROL LIST
PEAS	SAW
CARROTS	SCREW
BEANS	CHISEL
SPINACH	DRILL
LETTUCE	SANDPAPER
HAMMER	**HAMMER**
TOMATOES	NAILS
CORN	BENCH
CABBAGE	RULER
SQUASH	ANVIL



False Memory

(e.g. Deese, 1959; Roediger & McDermott)



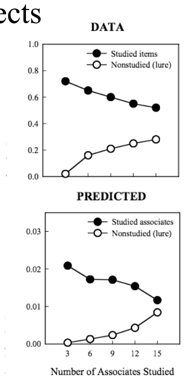
Study this list:
Bed, Rest, Awake, Tired, Dream, Wake, Snooze, Blanket, Doze, Slumber, Snore, Nap, Peace, Yawn, Drowsy

False memory effects

Number of Associates

3	6	9
MAD	MAD	MAD
FEAR	FEAR	FEAR
HATE	HATE	HATE
SMOOTH	RAGE	RAGE
NAVY	TEMPER	TEMPER
HEAT	FURY	FURY
SALAD	SALAD	WRATH
TUNE	TUNE	HAPPY
COURTS	COURTS	FIGHT
CANDY	CANDY	CANDY
PALACE	PALACE	PALACE
PLUSH	PLUSH	PLUSH
TOOTH	TOOTH	TOOTH
BLIND	BLIND	BLIND
WINTER	WINTER	WINTER

(lure = ANGER)



Robinson & Roediger (1997)

Modeling Serial Order Effects in Free Recall

Problem

- Dual route model predicts no sequential effects
  - Order of words is important in human memory experiments
- Standard Gibbs sampler is psychologically implausible:
  - Assumes list is processed in parallel
  - Each item can influence encoding of each other item

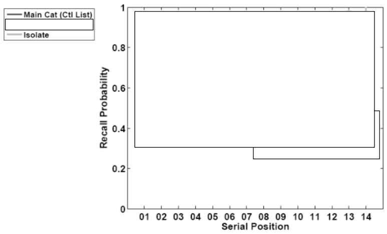
Semantic isolation experiment to study order effects

- Study lists of 14 words long
  - 14 isolate lists (e.g. A A A B A A ... A A)
  - 14 control lists (e.g. A A A A A A ... A A)
- Varied serial position of isolate (any of 14 positions)

Immediate Recall Results

Control list: A A A A A ... A

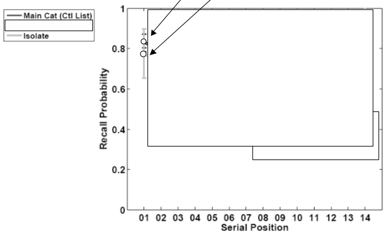
Isolate list: B A A A A ... A



Immediate Recall Results

Control list: A A A A A ... A

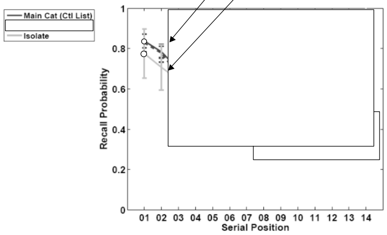
Isolate list: B A A A A ... A



Immediate Recall Results

Control list: A A A A A ... A

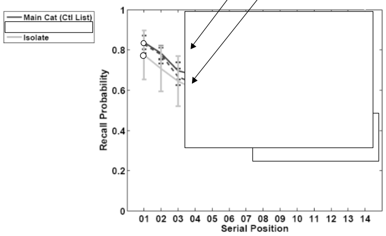
Isolate list: A B A A A ... A



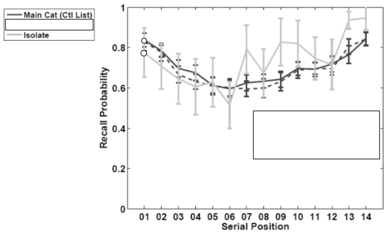
Immediate Recall Results

Control list: A A A A A ... A

Isolate list: A A B A A ... A

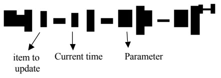


Immediate Recall Results



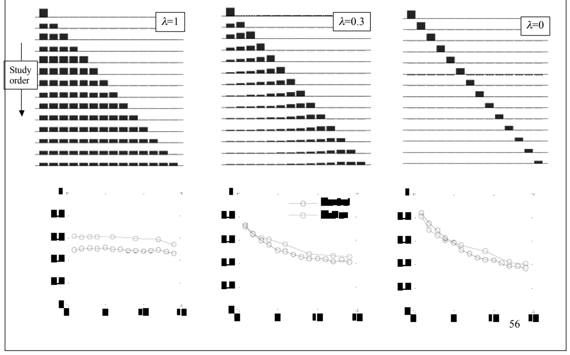
Modified Gibbs Sampling Scheme

- Update items non-uniformly in Gibbs sampler
- Probability of updating item i after observing words 1..*t*

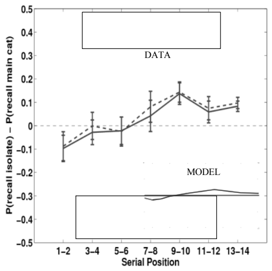


→ Words further back in time are less likely to be re-assigned

Effect of Sampling Scheme



Normalized Serial Position Effects

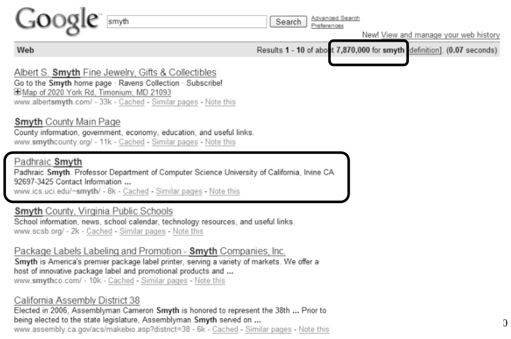


Information Retrieval & Human Memory

Example

- Searching for information on Padhraic Smyth:

Query = "Smyth"



Results **1** - **10** of about **7,870,000** for **smyth** [definition]. (**0.07** seconds)

Albert S. **Smyth** Fine Jewelry, Gifts & Collectibles
Go to the **Smyth** home page · Ravens Collection · Subscribe!
⊞Map of 2020 York Rd, Timonium, MD 21093
www.albert**smyth**.com/ - 33k - Cached - Similar pages - Note this

Smyth County Main Page
County information, government, economy, education, and useful links.
www.**smyth**county.org/ - 11k - Cached - Similar pages - Note this

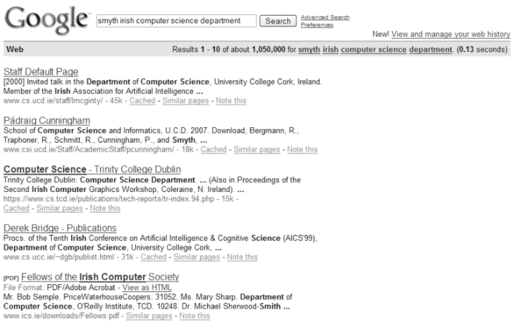
Padhraic **Smyth**
Padhraic **Smyth**. Professor Department of Computer Science University of California, Irvine CA 92697-3425 Contact Information **...**
www.ics.uci.edu/~**smyth**/ - 8k - Cached - Similar pages - Note this

Smyth County, Virginia Public Schools
School information, news, school calendar, technology resources, and useful links.
www.scsb.org/ - 2k - Cached - Similar pages - Note this

Package Labels Labeling and Promotion - **Smyth** Companies, Inc.
Smyth is America's premier package label printer, serving a variety of markets. We offer a host of innovative package label and promotional products and **...**
www.**smyth**co.com/ - 10k - Cached - Similar pages - Note this

California Assembly District 38
Elected in 2006, Assemblyman Cameron **Smyth** is honored to represent the 38th **...** Prior to being elected to the state legislature, Assemblyman **Smyth** served on **...**
www.assembly.ca.gov/acs/makebio.asp?district=38 - 6k - Cached - Similar pages - Note this

Query = "Smyth irish computer science department"



Google
smyth irish computer science department
Search
Advanced Search
Preferences
New! View and manage your web history
Web

Results **1** - **10** of about **1,050,000** for smyth irish computer science department. (**0.13** seconds)

Staff Default Page
[2000] Invited talk in the **Department** of **Computer Science**, University College Cork, Ireland. Member of the **Irish** Association for Artificial Intelligence **...**
www.cs.ucd.ie/staff/lmcginty/ - 45k - Cached - Similar pages - Note this

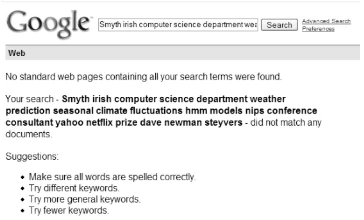
Pádraig Cunningham
School of **Computer Science** and Informatics, U.C.D. 2007. Download; Bergmann, R., Traphoner, R., Schmitt, R., Cunningham, P., and **Smyth**, **...**
www.csi.ucd.ie/Staff/AcademicStaff/pcunningham/ - 18k - Cached - Similar pages - Note this

Computer Science - Trinity College Dublin
Trinity College Dublin: **Computer Science Department**. **...** (Also in Proceedings of the Second **Irish Computer** Graphics Workshop, Coleraine, N. Ireland). **...**
https://www.cs.tcd.ie/publications/tech-reports/tr-index.94.php - 15k - Cached - Similar pages - Note this

Derek Bridge - Publications
Procs. of the Tenth **Irish** Conference on Artificial Intelligence & Cognitive **Science** (AICS'99), **Department** of **Computer Science**, University College Cork, **...**
www.cs.ucc.ie/~dgb/publist.html - 31k - Cached - Similar pages - Note this

[PDF] Fellows of the **Irish Computer** Society
File Format: PDF/Adobe Acrobat - View as HTML
Mr. Bob Semple. PriceWaterhouseCoopers. 31052. Ms. Mary Sharp. **Department** of **Computer Science**, O'Reilly Institute, TCD. 10248. Dr. Michael Sherwood-**Smith** **...**
www.ics.ie/downloads/Fellows.pdf - Similar pages - Note this

Query = "Smyth irish computer science department weather prediction seasonal climate fluctuations hmm models nips conference consultant yahoo netflix prize dave newman steyvers"



No standard web pages containing all your search terms were found.

Your search - **Smyth irish computer science department weather prediction seasonal climate fluctuations hmm models nips conference consultant yahoo netflix prize dave newman steyvers** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Problem

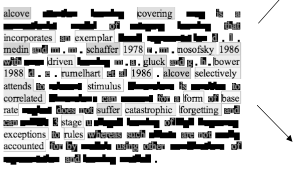
- More information in a query can lead to worse search results
- Human memory typically works better with more cues
- Problem: how can we better match queries to documents to allow for partial matches, and matches across documents?

Dual route model for information retrieval

- Encode documents with two routes
  - contextually unique words → verbatim route
  - Thematic words → topics route

Example encoding of a psych review abstract

Kruschke, J. K.. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22-44.



alcove covering is a
of that
incorporates an exemplar
medin and schaffer 1978 nosofsky 1986
with driven gluck and bower
1988 rumelhart et al 1986 alcove selectively
attends to stimulus is to
correlated can for a form of base
rate does not suffer catastrophic forgetting and
can 3 stage of
exceptions to rules whereas such are not
accounted for by using other of
and

Contextually unique words: ALCOVE, SCHAFFER, MEDIN, NOSOFSKY

Topic 1 (p=0.21): learning phenomena acquisition learn acquired ...

Topic 22 (p=0.17): similarity objects object space category dimensional categories spatial

Topic 61 (p=0.08): representations representation order alternative 1st higher 2nd descriptions problem form

Retrieval Experiments

- For each candidate document, calculate how likely the query was "generated" from the model's encoding

[illegible]

Information Retrieval Results

Evaluation Metric: precision for 10 highest ranked docs

APs

Method	Title	Desc	Concepts
TFIDF	.406	.434	.549
LSI	.455	.469	.523
LDA	.478	.463	.556
SW	.488	.468	**.561**
SWB	**.495**	**.473**	.558

FRs

Method	Title	Desc	Concepts
TFIDF	.300	.287	.483
LSI	.366	.327	.487
LDA	.428	.340	.487
SW	.448	**.407**	**.560**
SWB	**459**	.400	**.560**

Information retrieval systems in the mind & web

- Similar computational demands:
  - Both retrieve the most relevant items from a large information repository in response to external cues or queries.
- Useful analogies/ interdisciplinary approaches
- Many cognitive aspects in information retrieval
  - Internet content is produced by humans
  - Queries are formulated by humans

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. (in press). 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, in press). Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In: *Advances in Neural Information Processing Systems, 19.*

Text Mining Applications

Topics provide quick summary of content

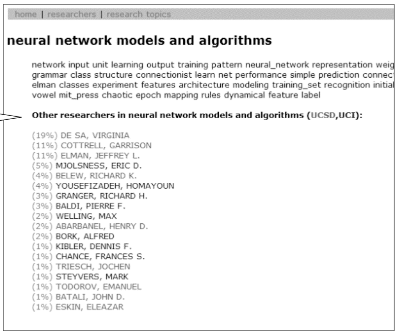
- Who writes on what topics?
- What is in this corpus? What is in this document?
- What are the topical trends over time?
- Who is mentioned in what context?

Faculty Browser

- System spiders UCI/UCSD faculty websites related to CalIT2 = California Institute for Telecommunications and Information Technology
- Applies topic model on text extracted from pdf files
- Browser demo: http://yarra.calit2.uci.edu/calit2/

one topic

most prolific researchers for this topic



home | researchers | research topics

neural network models and algorithms

network input unit learning output training pattern neural_network representation weig
grammar class structure connectionist learn net performance simple prediction connec
elman classes experiment features architecture modeling training_set recognition initial
vowel mit_press chaotic epoch mapping rules dynamical feature label

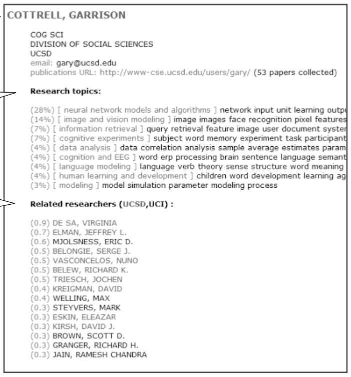
Other researchers in neural network models and algorithms (UCSD,UCI):

(19%) DE SA, VIRGINIA
(11%) COTTRELL, GARRISON
(11%) ELMAN, JEFFREY L.
(5%) MJOLSNESS, ERIC D.
(4%) BELEW, RICHARD K.
(4%) YOUSEFIZADEH, HOMAYOUN
(3%) GRANGER, RICHARD H.
(3%) BALDI, PIERRE F.
(2%) WELLING, MAX
(2%) ABARBANEL, HENRY D.
(2%) BORK, ALFRED
(1%) KIBLER, DENNIS F.
(1%) CHANCE, FRANCES S.
(1%) TRIESCH, JOCHEN
(1%) STEYVERS, MARK
(1%) TODOROV, EMANUEL
(1%) BATALI, JOHN D.
(1%) ESKIN, ELEAZAR

one researcher

topics this researcher works on

other researchers with similar topical interests



COTTRELL, GARRISON

COG SCI
DIVISION OF SOCIAL SCIENCES
UCSD
email: gary@ucsd.edu
publications URL: http://www-cse.ucsd.edu/users/gary/ (53 papers collected)

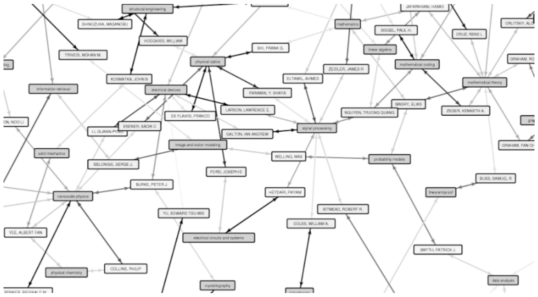
Research topics:

(28%) [neural network models and algorithms] network input unit learning outp
(14%) [image and vision modeling] image images face recognition pixel features
(7%) [information retrieval] query retrieval feature image user document syster
(7%) [cognitive experiments] subject word memory experiment task participant
(4%) [data analysis] data correlation analysis sample average estimates param
(4%) [cognition and EEG] word erp processing brain sentence language semant
(4%) [language modeling] language verb theory sense structure word meaning
(4%) [human learning and development] children word development learning ag
(3%) [modeling] model simulation parameter modeling process

Related researchers (UCSD,UCI) :

(0.9) DE SA, VIRGINIA
(0.7) ELMAN, JEFFREY L.
(0.6) MJOLSNESS, ERIC D.
(0.5) BELONGIE, SERGE J.
(0.5) VASCONCELOS, NUNO
(0.5) BELEW, RICHARD K.
(0.5) TRIESCH, JOCHEN
(0.4) KREIGMAN, DAVID
(0.4) WELLING, MAX
(0.3) STEYVERS, MARK
(0.3) ESKIN, ELEAZAR
(0.3) KIRSH, DAVID J.
(0.3) BROWN, SCOTT D.
(0.3) GRANGER, RICHARD H.
(0.3) JAIN, RAMESH CHANDRA

Inferred network of researchers connected through topics



Analyzing the New York Times

330,000 articles
2000-2002

Extracted Named Entities

Three investigations began Thursday into the **securities** and **exchange_commission**'s choice of **william_webster** to head a new board overseeing the accounting profession. **house** and **senate_democrats** called for the resignations of both **judge_webster** and **harvey_pitt**, the commission's chairman. The **white_house** expressed support for **judge_webster** as well as for **harvey_pitt**, who was harshly criticized Thursday for failing to inform other commissioners before they approved the choice of **judge_webster** that he had led the audit committee of a company facing fraud accusations. “The president still has confidence in **harvey_pitt**,” said **dan_bartlett**, **bush**'s communications director …

- Used standard algorithms to extract named entities:
  - People
  - Places
  - Organizations

Standard Topic Model with Entities

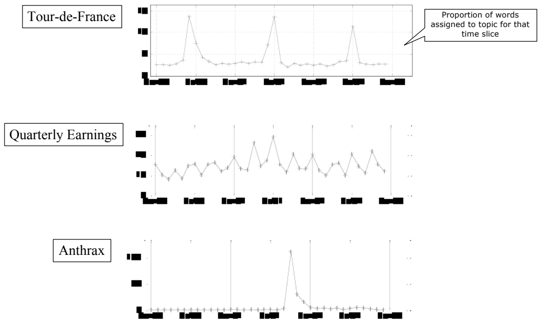
[redacted]	
team	0.028
play	0.015
game	0.013
season	0.012
final	0.011
games	0.011
point	0.011
series	0.011
player	0.010
coach	0.009
playoff	0.009
championship	0.007
playing	0.006
win	0.006
LAKERS	0.062
SHAQUILLE-O-NEAL	0.028
KOBE-BRYANT	0.028
PHIL-JACKSON	0.019
NBA	0.013
SACRAMENTO	0.007
RICK-FOX	0.007
PORTLAND	0.006
ROBERT-HORRY	0.006
DEREK-FISHER	0.006

[redacted]	
tour	0.039
rider	0.029
riding	0.017
bike	0.016
team	0.016
stage	0.014
race	0.013
won	0.012
bicycle	0.010
road	0.009
hour	0.009
scooter	0.008
mountain	0.008
place	0.008
LANCE-ARMSTRONG	0.021
FRANCE	0.011
JAN-ULLRICH	0.003
LANCE	0.003
U-S-POSTAL-SERVICE	0.002
MARCO-PANTANI	0.002
PARIS	0.002
ALPS	0.002
PYRENEES	0.001
SPAIN	0.001

[redacted]	
holiday	0.071
gift	0.050
toy	0.023
season	0.019
doll	0.014
tree	0.011
present	0.008
giving	0.008
special	0.007
shopping	0.007
family	0.007
celebration	0.007
card	0.007
tradition	0.006
CHRISTMAS	0.058
THANKSGIVING	0.018
SANTA-CLAUS	0.009
BARBIE	0.004
HANUKKAH	0.003
MATTEL	0.003
GRINCH	0.003
HALLMARK	0.002
EASTER	0.002
HASBRO	0.002

[redacted]	
award	0.026
film	0.020
actor	0.020
nomination	0.019
movie	0.015
actress	0.011
won	0.011
director	0.010
nominated	0.010
supporting	0.010
winner	0.008
picture	0.008
performance	0.007
nominees	0.007
OSCAR	0.035
ACADEMY	0.020
HOLLYWOOD	0.009
DENZEL-WASHINGTON	0.006
JULIA-ROBERT	0.005
RUSSELL-CROWE	0.005
TOM-HANK	0.005
STEVEN-SODERBERGH	0.004
ERIN-BROCKOVICH	0.003
KEVIN-SPACEY	0.003

Topic Trends



Proportion of words assigned to topic for that time slice

Example of Extracted
Entity-Topic Network

Prediction of Missing Entities in Text

Shares of **XXXX** slid 8 percent, or $1.10, to $12.65 Tuesday, as major credit agencies said the conglomerate would still be challenged in repaying its debts, despite raising $4.6 billion Monday in taking its finance group public. Analysts at **XXXX** Investors service in **XXXX** said they were keeping **XXXX** and its subsidiaries under review for a possible debt downgrade, saying the company "will continue to face a significant debt burden," with large slices of debt coming due, over the next 18 months. **XXXX** said …

Test article with entities removed

Prediction of Missing Entities in Text

Shares of **XXXX** slid 8 percent, or $1.10, to $12.65 Tuesday, as major credit agencies said the conglomerate would still be challenged in repaying its debts, despite raising $4.6 billion Monday in taking its finance group public. Analysts at **XXXX** Investors service in **XXXX** said they were keeping **XXXX** and its subsidiaries under review for a possible debt downgrade, saying the company "will continue to face a significant debt burden," with large slices of debt coming due, over the next 18 months. **XXXX** said …

Test article with entities removed

fitch goldman-sachs lehman-brother moody morgan-stanley new-york-stock-exchange standard-and-poor tyco tyco-international wall-street worldco

Actual missing entities

Prediction of Missing Entities in Text

Shares of **XXXX** slid 8 percent, or $1.10, to $12.65 Tuesday, as major credit agencies said the conglomerate would still be challenged in repaying its debts, despite raising $4.6 billion Monday in taking its finance group public. Analysts at **XXXX** Investors service in **XXXX** said they were keeping **XXXX** and its subsidiaries under review for a possible debt downgrade, saying the company "will continue to face a significant debt burden," with large slices of debt coming due, over the next 18 months. **XXXX** said …

Test article with entities removed

fitch goldman-sachs lehman-brother moody morgan-stanley new-york-stock-exchange standard-and-poor tyco tyco-international wall-street worldco

Actual missing entities

wall-street new-york nasdaq securities-exchange-commission sec merrill-lynch **new-york-stock-exchange** **goldman-sachs** **standard-and-poor**

Predicted entities given observed words (matches in **blue**)

Model Extensions

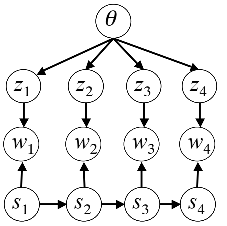
Model Extensions

- HMM-topics model
  - Modeling aspects of syntax
- Hierarchical topic model
  - Modeling relations between topics
- Collocation topic models
  - Learning collocations of words within topics

Hidden Markov Topic Model

Hidden Markov Topics Model

- Syntactic dependencies → short range dependencies
- Semantic dependencies → long-range

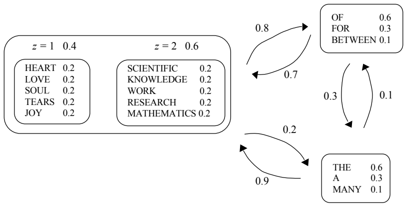


Semantic state: generate words from topic model

Syntactic states: generate words from HMM

(Griffiths, Steyvers, Blei, & Tenenbaum, 2004)

Transition between semantic state and syntactic states



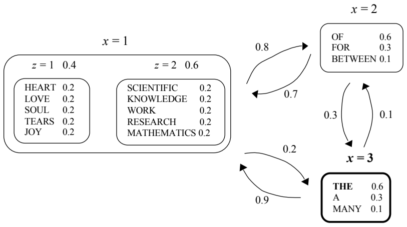
z = 1 0.4
HEART 0.2
LOVE 0.2
SOUL 0.2
TEARS 0.2
JOY 0.2

z = 2 0.6
SCIENTIFIC 0.2
KNOWLEDGE 0.2
WORK 0.2
RESEARCH 0.2
MATHEMATICS 0.2

OF 0.6
FOR 0.3
BETWEEN 0.1

THE 0.6
A 0.3
MANY 0.1

Combining topics and syntax



x = 1
z = 1 0.4
HEART 0.2
LOVE 0.2
SOUL 0.2
TEARS 0.2
JOY 0.2

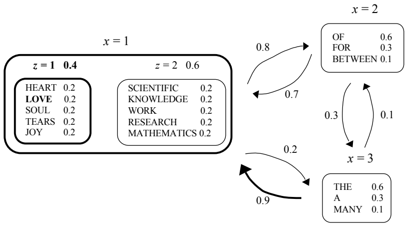
z = 2 0.6
SCIENTIFIC 0.2
KNOWLEDGE 0.2
WORK 0.2
RESEARCH 0.2
MATHEMATICS 0.2

x = 2
OF 0.6
FOR 0.3
BETWEEN 0.1

x = 3
THE 0.6
A 0.3
MANY 0.1

THE ……………………………

Combining topics and syntax



x = 1
z = 1 0.4
HEART 0.2
LOVE 0.2
SOUL 0.2
TEARS 0.2
JOY 0.2

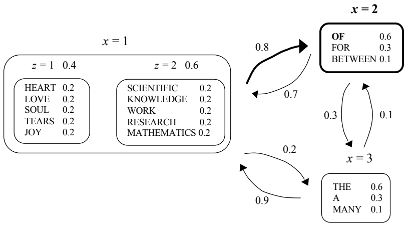
z = 2 0.6
SCIENTIFIC 0.2
KNOWLEDGE 0.2
WORK 0.2
RESEARCH 0.2
MATHEMATICS 0.2

x = 2
OF 0.6
FOR 0.3
BETWEEN 0.1

x = 3
THE 0.6
A 0.3
MANY 0.1

THE LOVE……………………

Combining topics and syntax



x = 2

OF 0.6
FOR 0.3
BETWEEN 0.1

x = 1

z = 1 0.4

HEART 0.2
LOVE 0.2
SOUL 0.2
TEARS 0.2
JOY 0.2

z = 2 0.6

SCIENTIFIC 0.2
KNOWLEDGE 0.2
WORK 0.2
RESEARCH 0.2
MATHEMATICS 0.2

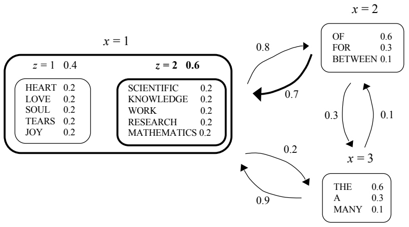
0.8
0.7
0.3
0.1
0.2
0.9

x = 3

THE 0.6
A 0.3
MANY 0.1

THE LOVE OF………………

Combining topics and syntax



x = 2

OF 0.6
FOR 0.3
BETWEEN 0.1

x = 1

z = 1 0.4

HEART 0.2
LOVE 0.2
SOUL 0.2
TEARS 0.2
JOY 0.2

z = 2 0.6

SCIENTIFIC 0.2
KNOWLEDGE 0.2
WORK 0.2
RESEARCH 0.2
MATHEMATICS 0.2

0.8
0.7
0.3
0.1
0.2
0.9

x = 3

THE 0.6
A 0.3
MANY 0.1

THE LOVE OF RESEARCH ……

Semantic topics

FOOD	MAP	DOCTOR	BOOK	GOLD	BEHAVIOR	CELLS	PLANTS
FOODS	NORTH	PATIENT	BOOKS	IRON	SELF	CELL	PLANT
BODY	EARTH	HEALTH	READING	SILVER	INDIVIDUAL	ORGANISMS	LEAVES
NUTRIENTS	SOUTH	HOSPITAL	INFORMATION	COPPER	PERSONALITY	ALGAE	SEEDS
DIET	POLE	MEDICAL	LIBRARY	METAL	RESPONSE	BACTERIA	SOIL
FAT	MAPS	CARE	REPORT	METALS	SOCIAL	MICROSCOPE	ROOTS
SUGAR	EQUATOR	PATIENTS	PAGE	STEEL	EMOTIONAL	MEMBRANE	FLOWERS
ENERGY	WEST	NURSE	TITLE	CLAY	LEARNING	ORGANISM	WATER
MILK	LINES	DOCTORS	SUBJECT	LEAD	FEELINGS	FOOD	FOOD
EATING	EAST	MEDICINE	PAGES	ADAM	PSYCHOLOGISTS	LIVING	GREEN
FRUITS	AUSTRALIA	NURSING	GUIDE	ORE	INDIVIDUALS	FUNGI	SEED
VEGETABLES	GLOBE	TREATMENT	WORDS	ALUMINUM	PSYCHOLOGICAL	MOLD	STEMS
WEIGHT	POLES	NURSES	MATERIAL	MINERAL	EXPERIENCES	MATERIALS	FLOWER
FATS	HEMISPHERE	PHYSICIAN	ARTICLE	MINE	ENVIRONMENT	NUCLEUS	STEM
NEEDS	LATITUDE	HOSPITALS	ARTICLES	STONE	HUMAN	CELLED	LEAF
CARBOHYDRATES	PLACES	DR	WORD	MINERALS	RESPONSES	STRUCTURES	ANIMALS
VITAMINS	LAND	SICK	FACTS	POT	BEHAVIORS	MATERIAL	ROOT
CALORIES	WORLD	ASSISTANT	AUTHOR	MINING	ATTITUDES	STRUCTURE	POLLEN
PROTEIN	COMPASS	EMERGENCY	REFERENCE	MINERS	PSYCHOLOGY	GREEN	GROWING
MINERALS	CONTINENTS	PRACTICE	NOTE	TIN	PERSON	MOLDS	GROW

Syntactic classes

SAID	THE	MORE	ON	GOOD	ONE	HE	BE
ASKED	HIS	SUCH	AT	SMALL	SOME	YOU	MAKE
THOUGHT	THEIR	LESS	INTO	NEW	MANY	THEY	GET
TOLD	YOUR	MUCH	FROM	IMPORTANT	TWO	I	HAVE
SAYS	HER	KNOWN	WITH	GREAT	EACH	SHE	GO
MEANS	ITS	JUST	THROUGH	LITTLE	ALL	WE	TAKE
CALLED	MY	BETTER	OVER	LARGE	MOST	IT	DO
CRIED	OUR	RATHER	AROUND	*	ANY	PEOPLE	FIND
SHOWS	THIS	GREATER	AGAINST	BIG	THREE	EVERYONE	USE
ANSWERED	THESE	HIGHER	ACROSS	LONG	THIS	OTHERS	SEE
TELLS	A	LARGER	UPON	HIGH	EVERY	SCIENTISTS	HELP
REPLIED	AN	LONGER	TOWARD	DIFFERENT	SEVERAL	SOMEONE	KEEP
SHOUTED	THAT	FASTER	UNDER	SPECIAL	FOUR	WHO	GIVE
EXPLAINED	NEW	EXACTLY	ALONG	OLD	FIVE	NOBODY	LOOK
LAUGHED	THOSE	SMALLER	NEAR	STRONG	BOTH	ONE	COME
MEANT	EACH	SOMETHING	BEHIND	YOUNG	TEN	SOMETHING	WORK
WROTE	MR	BIGGER	OFF	COMMON	SIX	ANYONE	MOVE
SHOWED	ANY	FEWER	ABOVE	WHITE	MUCH	EVERYBODY	LIVE
BELIEVED	MRS	LOWER	DOWN	SINGLE	TWENTY	SOME	EAT
WHISPERED	ALL	ALMOST	BEFORE	CERTAIN	EIGHT	THEN	BECOME

NIPS Semantics

IMAGE	DATA	STATE	MEMBRANE	EXPERTS	KERNEL	NETWORK
IMAGES	GAUSSIAN	POLICY	SYNAPTIC	EXPERT	SUPPORT	NEURAL
OBJECT	MIXTURE	VALUE	CELL	GATING	VECTOR	NETWORKS
OBJECTS	LIKELIHOOD	FUNCTION	*	HME	SVM	OUPUT
FEATURE	POSTERIOR	ACTION	CURRENT	ARCHITECTURE	KERNELS	INPUT
RECOGNITION	PRIOR	REINFORCEMENT	DENDRITIC	MIXTURE	#	TRAINING
VIEWS	DISTRIBUTION	LEARNING	POTENTIAL	LEARNING	SPACE	INPUTS
#	EM	CLASSES	NEURON	MIXTURES	FUNCTION	WEIGHTS
PIXEL	BAYESIAN	OPTIMAL	CONDUCTANCE	FUNCTION	MACHINES	#
VISUAL	PARAMETERS	*	CHANNELS	GATE	SET	OUTPUTS

NIPS Syntax

IN	IS	SEE	USED	MODEL	HOWEVER	#
WITH	WAS	SHOW	TRAINED	ALGORITHM	ALSO	*
FOR	HAS	NOTE	OBTAINED	SYSTEM	THEN	I
ON	BECOMES	CONSIDER	DESCRIBED	CASE	THUS	X
FROM	DENOTES	ASSUME	GIVEN	PROBLEM	THEREFORE	T
AT	BEING	PRESENT	FOUND	NETWORK	FIRST	N
USING	REMAINS	NEED	PRESENTED	METHOD	HERE	-
INTO	REPRESENTS	PROPOSE	DEFINED	APPROACH	NOW	C
OVER	EXISTS	DESCRIBE	GENERATED	PAPER	HENCE	F
WITHIN	SEEMS	SUGGEST	SHOWN	PROCESS	FINALLY	P

Random sentence generation

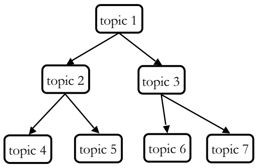
LANGUAGE:
[S] RESEARCHERS GIVE THE SPEECH
[S] THE SOUND FEEL NO LISTENERS
[S] WHICH WAS TO BE MEANING
[S] HER VOCABULARIES STOPPED WORDS
[S] HE EXPRESSLY WANTED THAT BETTER VOWEL

Nested Chinese Restaurant Process

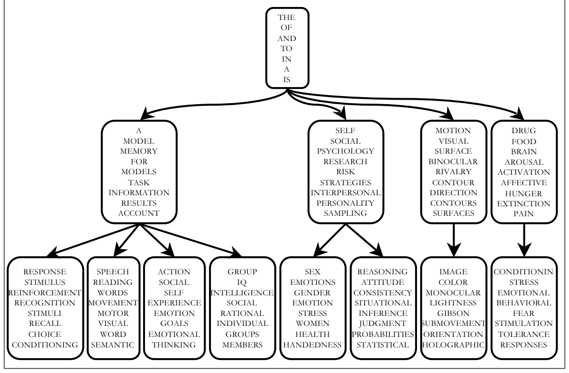
Topic Hierarchies

- In regular topic model, no relations between topics

- Nested Chinese Restaurant Process
  - Blei, Griffiths, Jordan, Tenenbaum (2004)
  - Learn hierarchical structure, as well as topics within structure



Example: Psych Review Abstracts



THE OF AND TO IN A IS

A MODEL MEMORY FOR MODELS TASK INFORMATION RESULTS ACCOUNT

SELF SOCIAL PSYCHOLOGY RESEARCH RISK STRATEGIES INTERPERSONAL PERSONALITY SAMPLING

MOTION VISUAL SURFACE BINOCULAR RIVALRY CONTOUR DIRECTION CONTOURS SURFACES

DRUG FOOD BRAIN AROUSAL ACTIVATION AFFECTIVE HUNGER EXTINCTION PAIN

RESPONSE STIMULUS REINFORCEMENT RECOGNITION STIMULI RECALL CHOICE CONDITIONING

SPEECH READING WORDS MOVEMENT MOTOR VISUAL WORD SEMANTIC

ACTION SOCIAL SELF EXPERIENCE EMOTION GOALS EMOTIONAL THINKING

GROUP IQ INTELLIGENCE SOCIAL RATIONAL INDIVIDUAL GROUPS MEMBERS

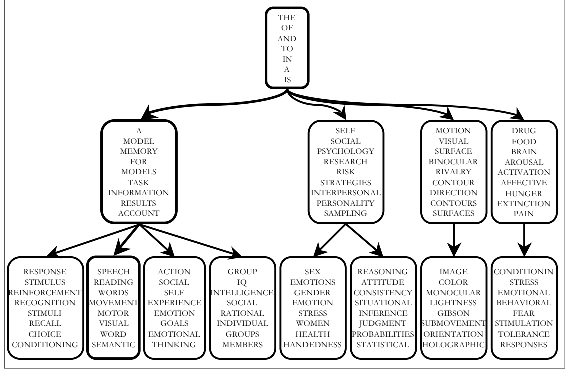
SEX EMOTIONS GENDER EMOTION STRESS WOMEN HEALTH HANDEDNESS

REASONING ATTITUDE CONSISTENCY SITUATIONAL INFERENCE JUDGMENT PROBABILITIES STATISTICAL

IMAGE COLOR MONOCULAR LIGHTNESS GIBSON SUBMOVEMENT ORIENTATION HOLOGRAPHIC

CONDITIONIN STRESS EMOTIONAL BEHAVIORAL FEAR STIMULATION TOLERANCE RESPONSES

Generative Process



THE OF AND TO IN A IS

A MODEL MEMORY FOR MODELS TASK INFORMATION RESULTS ACCOUNT

SELF SOCIAL PSYCHOLOGY RESEARCH RISK STRATEGIES INTERPERSONAL PERSONALITY SAMPLING

MOTION VISUAL SURFACE BINOCULAR RIVALRY CONTOUR DIRECTION CONTOURS SURFACES

DRUG FOOD BRAIN AROUSAL ACTIVATION AFFECTIVE HUNGER EXTINCTION PAIN

RESPONSE STIMULUS REINFORCEMENT RECOGNITION STIMULI RECALL CHOICE CONDITIONING

SPEECH READING WORDS MOVEMENT MOTOR VISUAL WORD SEMANTIC

ACTION SOCIAL SELF EXPERIENCE EMOTION GOALS EMOTIONAL THINKING

GROUP IQ INTELLIGENCE SOCIAL RATIONAL INDIVIDUAL GROUPS MEMBERS

SEX EMOTIONS GENDER EMOTION STRESS WOMEN HEALTH HANDEDNESS

REASONING ATTITUDE CONSISTENCY SITUATIONAL INFERENCE JUDGMENT PROBABILITIES STATISTICAL

IMAGE COLOR MONOCULAR LIGHTNESS GIBSON SUBMOVEMENT ORIENTATION HOLOGRAPHIC

CONDITIONIN STRESS EMOTIONAL BEHAVIORAL FEAR STIMULATION TOLERANCE RESPONSES

Collocation Topic Model

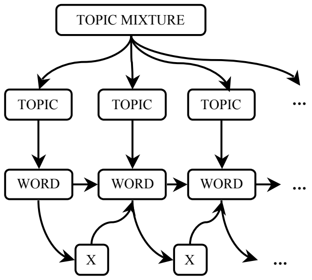
What about collocations?

- Why are these words related?
  - PLAY - GROUND
  - DOW - JONES
  - BUMBLE - BEE

- Suggests at least two routes for association:
  - Semantic
  - Collocation

→ Integrate collocations into topic model

Collocation Topic Model



If x=0, sample a word from the topic

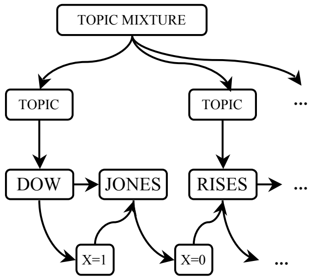
If x=1, sample a word from the distribution based on previous word

Collocation Topic Model

Example:
"DOW JONES RISES"

JONES is more likely explained as a word following DOW than as word sampled from topic

Result: DOW_JONES recognized as collocation



Examples Topics from New York Times

Terrorism

- SEPT_11
- WAR
- SECURITY
- IRAQ
- TERRORISM
- NATION
- KILLED
- AFGHANISTAN
- ATTACKS
- OSAMA_BIN_LADEN
- AMERICAN
- ATTACK
- NEW_YORK_REGION
- NEW
- MILITARY
- NEW_YORK
- WORLD
- NATIONAL
- QAEDA
- TERRORIST_ATTACKS

Wall Street Firms

- WALL_STREET
- ANALYSTS
- INVESTORS
- FIRM
- GOLDMAN_SACHS
- FIRMS
- INVESTMENT
- MERRILL_LYNCH
- COMPANIES
- SECURITIES
- RESEARCH
- STOCK
- BUSINESS
- ANALYST
- WALL_STREET_FIRMS
- SALOMON_SMITH_BARNEY
- CLIENTS
- INVESTMENT_BANKING
- INVESTMENT_BANKERS
- INVESTMENT_BANKS

Stock Market

- WEEK
- DOW_JONES
- POINTS
- 10_YR_TREASURY_YIELD
- PERCENT
- CLOSE
- NASDAQ_COMPOSITE
- STANDARD_POOR
- CHANGE
- FRIDAY
- DOW_INDUSTRIALS
- GRAPH_TRACKS
- EXPECTED
- BILLION
- NASDAQ_COMPOSITE_INDEX
- EST_02
- PHOTO_YESTERDAY
- YEN
- 10
- 500_STOCK_INDEX

Bankruptcy

- BANKRUPTCY
- CREDITORS
- BANKRUPTCY_PROTECTION
- ASSETS
- COMPANY
- FILED
- BANKRUPTCY_FILING
- ENRON
- BANKRUPTCY_COURT
- KMART
- CHAPTER_11
- FILING
- COOPER
- BILLIONS
- COMPANIES
- BANKRUPTCY_PROCEEDINGS
- DEBTS
- RESTRUCTURING
- CASE
- GROUP