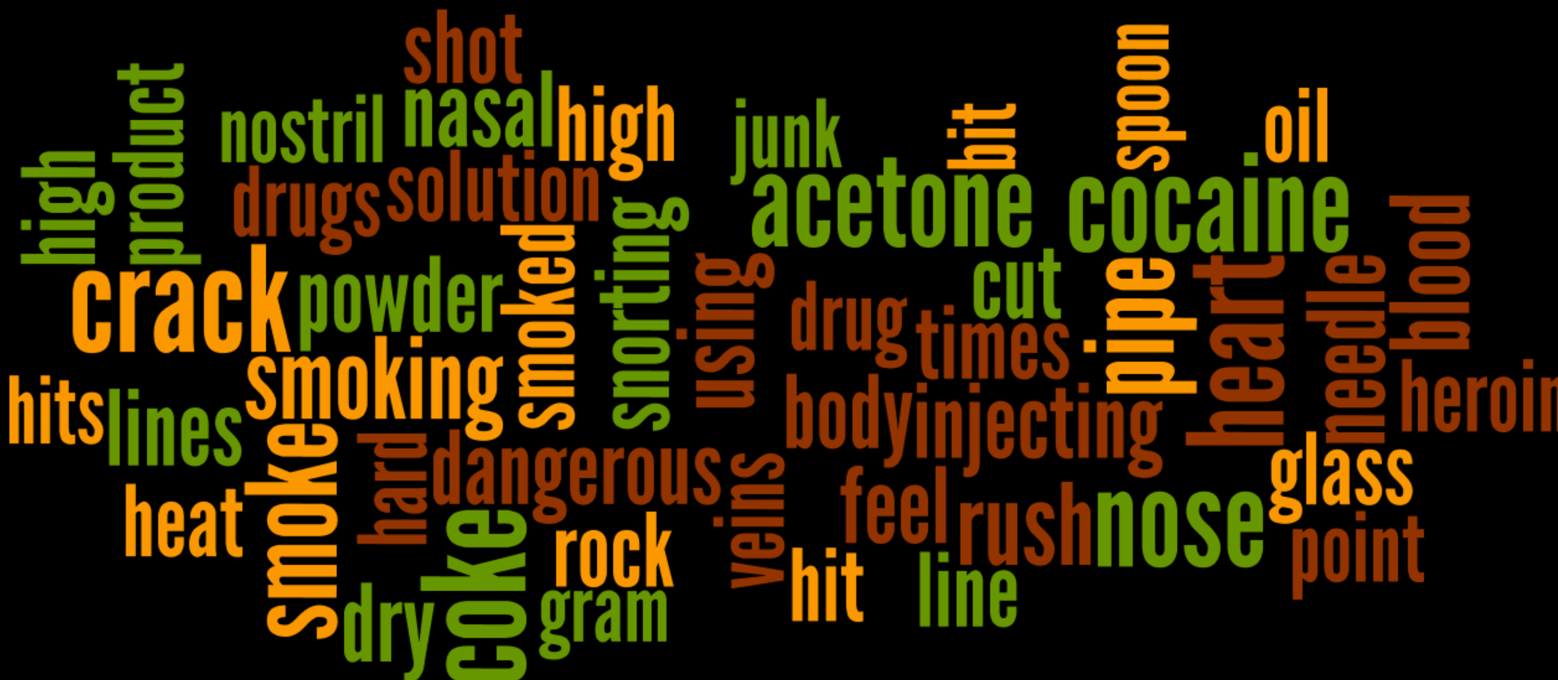
Summarizing Drug Experiences with Multi-Dimensional Topic Models



Michael Paul and **Mark Dredze**

Johns Hopkins University

Online Drug Communities

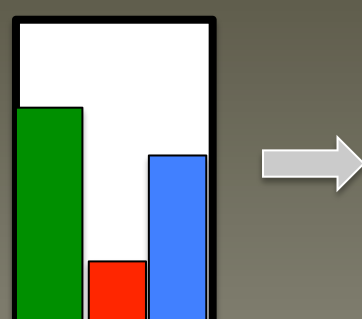
- **Drugs-Forum.com**
  - "Drugs-forum is an information hub of high-standards and a platform where people can freely discuss recreational drugs in a mature, intelligent manner. Drugs-Forum offers a wealth of quality information and discussion of drug-related politics, in addition to assistance for members struggling with addiction."
- Analyzed 100,000 messages
- Over 20,000 users in data set
  - 87% male
  - 50% American
  - 58% aged 20-29, 23% aged 30-39

Web-Based Drug Research

- Problem: novel drugs are created faster than researchers and officials can keep up; recent surge in new drugs
  - 49 new drugs detected in Europe in 2011 (a record)
- For new and emerging drugs, information can be difficult to obtain through traditional means
  - Modern source of information: Internet forums
  - Always curated manually by humans
- A step toward automation: **topic modeling**
  - Corpus exploration
  - Can be used for automatic **summarization** (later)

Topic Modeling

- Probabilistic model of text generation
  - e.g. Latent Dirichlet Allocation (Blei et al, 03)
- Each document has a distribution over *topics*
- Each topic has a distribution over words
- Each word token is associated with a latent topic variable



Jury Finds Baseball Star Roger Clemens Not Guilty On All Counts

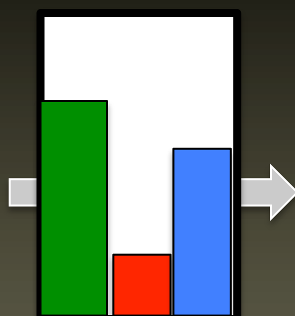
A jury found baseball star Roger Clemens not guilty on six charges against. Clemens was accused of lying to Congress in 2008 about his use of performance enhancing drugs.

Topic Modeling

football	0.03
team	0.01
hockey	0.01
baseball	0.005
…	…

charge	0.02
court	0.02
police	0.015
robbery	0.01
…	…

congress	0.02
president	0.02
election	0.015
senate	0.01
…	…



Jury Finds Baseball Star Roger Clemens Not Guilty On All Counts

A jury found baseball star Roger Clemens not guilty on six charges against. Clemens was accused of lying to Congress in 2008 about his use of performance enhancing drugs.

Factorial LDA (f-LDA)

- Multi-dimensional topic model
  - M.J. Paul and M. Dredze. Factorial LDA: Sparse Multidimensional Models of Text. NIPS 2012.
- Word tokens are associated with a *vector* of latent variables instead of a single topic variable
  - Can jointly model pairs of concepts like topic and perspective or sentiment
- Instead of a distribution over topics, each document has distribution over *tuples*
- Each tuple is associated with its own word distribution

Multi-Dimensional Topic Modeling

- Suppose we want to jointly model **topic** and editorial **perspective** in news articles
  - Could use f-LDA with 2 factors
- Each (topic,perspective) **pair** has its own word distribution
  - The same topic can be represented with different words, depending on the author perspective

democrats	0.035
obama	0.03
liberals	0.02
biden	0.005
…	…

republicans	0.02
romney	0.02
bush	0.015
republican	0.015
…	…

Factorial LDA for Drug Forums

- Joint model of 3 factors:
  - Drug type
  - Route of administration (i.e. method of intake)
  - Aspect

Drug (22 total)	Route	Aspect
• Alcohol • Amphetamine • Cannabis • Cocaine • ... • Salvia • Tobacco	• Injection • Oral • Smoking • Snorting	• Chemistry • Culture • Effects • Health • Usage

Factorial LDA for Drug Forums

- Joint model of 3 factors:
  - Drug type
  - Route of administration (i.e. method of intake)
  - Aspect
- Learn word distributions for triples such as:

(**Cocaine**,**Snorting**,**Health**)	(**Cocaine**,**Snorting**,**Usage**)
nose pain damage blood cocaine problem	coke line lines nose small cut

Model Parameters

- Why should the word distributions for triples make any sense?
- Parameters are tied across the priors of each word distribution
  - The prior for (**Cocaine**,**Snorting**,**Effects**) shares parameters with (**Cocaine**,**Smoking**,**Effects**) which shares parameters with the prior for (**Marijuana**,**Smoking**,**Effects**)

Marijuana	Oral	Chemistry
weed	capsules	solvent
cannabis	consumes	extraction
thc	toast	evaporate
marijuana	stomach	evaporated
stoned	chewing	solvents
bowl	ambien	evaporation
bud	digestion	yield
joint	juice	chloride
blunt	absorbed	alkaloids
herb	ingestion	tek
bong	meal	compounds
pot	tiredness	evaporating
sativa	chew	atom
blaze	juices	aromatic
indica	gelatin	non-polar
smoking	yogurt	purified
blunts	fruit	jar
strains	oj	methyl
hemp	digest	ethanol
...	...	....

Each dimension has a weight vector over the vocabulary

	Marijuana		Oral		Chemistry		
	weed		capsules		solvent		thc
	cannabis		consumes		extraction		method
	thc		toast		evaporate		extraction
	marijuana		stomach		evaporated		plant
	stoned		chewing		solvents		material
	bowl		ambien		evaporation		cannabis
	bud		digestion		yield		simple
	joint		juice		chloride		coffee
	blunt		absorbed		alkaloids		oil
exp(	herb	+	ingestion	+	tek	) =	contains
	bong		meal		compounds		jar
	pot		tiredness		evaporating		dried
	sativa		chew		atom		process
	blaze		juices		aromatic		dry
	indica		gelatin		non-polar		water
	smoking		yogurt		purified		extract
	blunts		fruit		jar		results
	strains		oj		methyl		salt
	hemp		digest		ethanol		available
	...		...		....		...

word distribution for triple

(**Marijuana** **Oral** **Chemistry**)

multinomial parameters sampled from Dirichlet

Posterior	Prior
oil	thc
water	method
butter	extraction
thc	plant
weed	material
hash	cannabis
cannabis	simple
alcohol	coffee
make	oil
milk	contains
high	jar
marijuana	dried
add	process
cup	dry
extract	water
...	extract
mixture	results
hours	salt
try	available
brownies	...

word distribution for triple

(**Marijuana**
Oral
Chemistry)

=

Posterior

oil
water
butter
thc
weed
hash
cannabis
alcohol
make
milk
high
marijuana
add
cup
extract
...
mixture
hours
try
brownies

multinomial parameters sampled from Dirichlet

Prior

thc
method
extraction
plant
material
cannabis
simple
coffee
oil
contains
jar
dried
process
dry
water
extract
results
salt
available
...

Model Parameters

- Where did the weight vectors come from?
- Parameter optimization
  - We learn from the data
- We would probably not learn anything sensible with zero supervision
  - Semi-supervised approach using informed priors
  - More on this soon

Model Parameters

- Where do the posteriors come from?
  – Gibbs sampling: basically identical to LDA sampler
- Our inference algorithm:
  – E step: 1 iteration of Gibbs sampling
  – M step: 1 iteration of gradient ascent
- Constraints:
  – Drug value fixed to subforum message came from
  – Route value restricted to values tagged by users

Semi-Supervision

- Each thread in the corpus contains a “tag”

- Can we leverage these tags to guide the model?

Semi-Supervision

- Our priors are log linear functions of weight vectors
- What if we trained a log linear model on documents with the tags as labels?

$$P(\text{word } w|\text{drug} = i, \text{factor} f = j) = \frac{\exp(\eta_w^{(0)} + \eta_{iw}^{(\text{drug})} + \eta_{jw}^{(f)})}{\sum_{w'} \exp(\eta_{w'}^{(0)} + \eta_{iw'}^{(\text{drug})} + \eta_{jw'}^{(f)})}$$

  – based on a model called SAGE (Eisenstein et al, '11)
- This gives us weight vectors that we could use in our model
  – But this model and the tags are both incomplete

Semi-Supervision

- The weights learned by training the log-linear model serve as a **Gaussian prior** over the weights in our f-LDA model

"Health"

"Health"	symptoms long-term depression disorder schizophrenia severe acute serotonin patients bodys psychosis psychological	$\sim N($	kidney hcv pains symptoms guidelines diet exercise hepatitis dreams disorder disease attack	$, \sigma^2)$

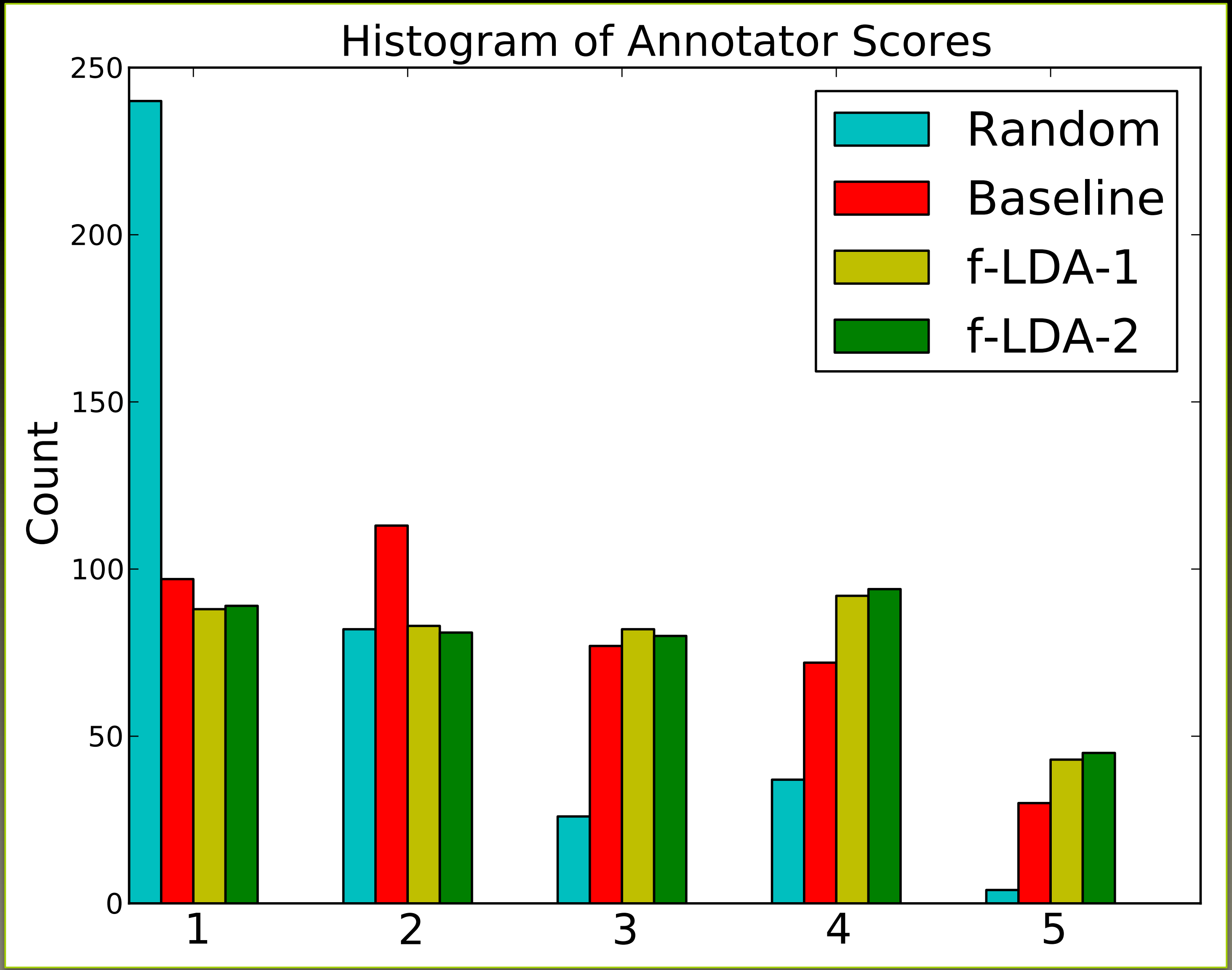
What can we learn by doing this?

- We can use the model to bring attention to relevant messages and snippets of text
  - Extractive summarization
- Pick the 5 best snippets for each triple word distribution
  - Snippets are spans of text of varying window size
  - Rank snippets by KL-divergence to each f-LDA distribution
  - Also considered distributions for pairs by marginalizing out the third factor

Evaluation

- Reference summaries for 5 new drugs
  - Technical reports from EU Psychonaut Project (Schifano et al, 2006)
  - These reports were created by reading similar web forums
  - We manually match some segments of reports to various triples/pairs
- 3 annotators asked give 1–5 score to each snippet
  - How well would the snippet inform the writing of the text segment?
- Systems
  - Baseline: unigram word model from tagged data
  - F-LDA-1: only messages with tags (25K)
  - F-LDA-2: includes messages without tags (100K)

Evaluation



Histogram of Annotator Scores
Count
0
50
100
150
200
250
1
2
3
4
5
Random
Baseline
f-LDA-1
f-LDA-2

Example Snippet

Reference Text:
» It is recommended by users that Mephedrone be taken on an empty stomach. Doses usually vary between 100mg – 1g.

F-LDA Text:
(Meph. Oral Usage)
» If it is [someone who isn't you]'s first time using Mephedrone [someone who isn't me] recommends a 100mg oral dose on an empty stomach.

Conclusion

- Online communities contain a large amount of candid data on a subject that is traditionally difficult to study
- Our experiments showed that we can automatically extract useful, targeted information
- Code for f-LDA with word priors will be available
  - http://cs.jhu.edu/~mpaul

Thank You

- Thanks to:
  - Meg Chisolm
  - Ryan Vandrey
  - Matt Johnson
  - Alex Lamb
  - Hieu Tran
  - NSF
- Johns Hopkins HLTCOE is hiring!
  - Research scientist and Postdoc positions
  - http://hltcoe.jhu.edu

human language technology
center of excellence

Example Snippets

Reference Text:

» "Dried leaves and/or salvia extract are smoked (using a butane lighter) either by pipe (considered to be the most effective but is considered to be quite painful) or water bong.

F-LDA Text:

(**Salvia** **Smoking** **Usage**)

» 2. Use a water pipe. Its harsh and needs to be smoked hot so this should be self explanatory. 3. Use a torch style lighter [...] Salvinorin A has a VERY high boiling point (around 700 degrees F I believe) so a regular bic just wont do it

Evaluation

- Estimated recall using ROUGE (Lin, 2004)
  - n-gram recall of reference text

	Random	Baseline	f-LDA-1	f-LDA-2
1-gram	.112	.326	**.355**	.327
2-gram	.023	.072	**.085**	.084

Evaluation

- Expert annotations
  - 2 faculty from the Johns Hopkins School of Medicine
  - rated snippets for two drugs: MDPV, Mephedrone
- Average ratings:
  - Random: 1.63
  - Baseline: 2.45
  - f-LDA: **2.57**