

# Experimenting with Drugs (and Topic Models)



Michael Paul and Mark Dredze

Johns Hopkins University



# Online Drug Communities

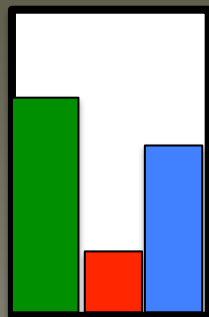
- **Drugs-Forum.com**
  - “Drugs-forum is an information hub of high-standards and a platform where people can freely discuss recreational drugs in a mature, intelligent manner. Drugs-Forum offers a wealth of quality information and discussion of drug-related politics, in addition to assistance for members 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

- For new and emerging drugs, information can be difficult to obtain through traditional means
  - e.g. lab experiments, surveys
- Modern source of information: Internet forums
  - Always curated manually by humans
- A step toward automation: **topic modeling**
  - Corpus exploration
  - Coarse-grained information extraction

# 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
- Topics are unobserved data (latent variables) which must be inferred

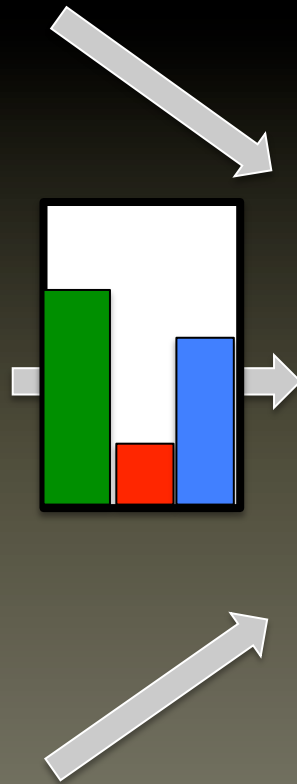


# Topic Modeling

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
...	...

# Performance Enhanced Factorial LDA

- Joint model of 3 factors:
  - Drug type
  - Method of intake (delivery)
  - Aspect

Drug (24 total)	Delivery	Aspect
<ul style="list-style-type: none"><li>• Alcohol</li><li>• Amphetamine</li><li>• Cannabis</li><li>• Cocaine</li><li>• ...</li><li>• Salvia</li><li>• Tobacco</li></ul>	<ul style="list-style-type: none"><li>• Injection</li><li>• Oral</li><li>• Smoking</li><li>• Snorting</li></ul>	<ul style="list-style-type: none"><li>• Chemistry</li><li>• Culture</li><li>• Effects</li><li>• Health</li><li>• Usage</li></ul>



# Performance Enhanced Factorial LDA

- Joint model of 3 dimensions:
  - Drug type
  - Method of intake (delivery)
  - 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 (Cannabis, Smoking, Effects)

## Cannabis


weed  
cannabis  
thc  
marijuana  
stoned  
bowl  
bud  
joint  
blunt  
herb  
bong  
pot  
sativa  
blaze  
indica  
smoking  
blunts  
strains  
hemp  
...

## Oral

capsules  
consumes  
toast  
stomach  
chewing  
ambien  
digestion  
juice  
absorbed  
ingestion  
meal  
tiredness  
chew  
juices  
gelatin  
yogurt  
fruit  
oj  
digest  
...

## Chemistry

solvent  
extraction  
evaporate  
evaporated  
solvents  
evaporation  
yield  
chloride  
alkaloids  
tek  
compounds  
evaporating  
atom  
aromatic  
non-polar  
purified  
jar  
methyl  
ethanol  
....



Each dimension  
has a weight vector  
over the vocabulary

exp(

## Cannabis

weed  
cannabis  
thc  
marijuana  
stoned  
bowl  
bud  
joint  
blunt  
herb  
bong  
pot  
sativa  
blaze  
indica  
smoking  
blunts  
strains  
hemp  
...



## Oral

capsules  
consumes  
toast  
stomach  
chewing  
ambien  
digestion  
juice  
absorbed  
ingestion  
meal  
tiredness  
chew  
juices  
gelatin  
yogurt  
fruit  
oj  
digest  
...



## Chemistry

solvent  
extraction  
evaporate  
evaporated  
solvents  
evaporation  
yield  
chloride  
alkaloids  
tek  
compounds  
evaporating  
atom  
aromatic  
non-polar  
purified  
jar  
methyl  
ethanol  
....



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

word distribution for triple

( Cannabis  
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  
...

word distribution for triple

( Cannabis  
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 seed words
  - More on this soon

# Model Parameters

- Where do the posteriors come from?
  - Gibbs sampling
- Our inference algorithm:
  - E step: 1 iteration of Gibbs sampling
  - M step: 1 iteration of gradient ascent



# Semi-Supervision

- Each thread in the corpus contains a “tag”

<a href="#">Culture</a> - <a href="#">Songs about cocaine</a> (  <a href="#">1</a> <a href="#">2</a> <a href="#">3</a> ... <a href="#">Last Page</a> ) Kittyofftitty
<a href="#">Experiences</a> - <a href="#">what to do on coke? cocaine activites</a> (  <a href="#">1</a> <a href="#">2</a> <a href="#">3</a> ... <a href="#">Last Page</a> ) madman316
<a href="#">Smoking</a> - <a href="#">Right way to smoke it?</a> ChristalVision
<a href="#">Effects</a> - <a href="#">Curious: crack vs IV intensity</a> MagicalOrangutan
<a href="#">Experiences</a> - <a href="#">You know you're a Crackhead..(add to it)</a> (  <a href="#">1</a> <a href="#">2</a> ) The Half Unlit
<a href="#">Effects</a> - <a href="#">Is Cocaine or Crack overrated ?</a> war209

- 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?

$$P(\text{word } w | \text{drug} = i, \text{factor } f = j) \\ = \frac{\exp(m_w + \eta_{iw}^{(1)} + \eta_{jw}^{(f)})}{\sum_{w'} \exp(m_{w'} + \eta_{iw'}^{(1)} + \eta_{jw'}^{(f)})}$$

- 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”

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

# Questions

- How are people using Salvia?



## ( Salvia Smoking Usage )

» "Best way is to use a torch lighter, bong or pipe (bong recommended) and hold in each hit 20-40 seconds."

## ( Salvia Oral Usage )

» "A dose of Salvia leaves is 2 grams. A dose of Salvia 5X extract is 0.4 grams (400mg) A dose of salvia 10X extract is 0.2 grams (200mg) A dose of Salvia 25X extract is 0.08 grams (80mg) A dose of Salvia 50X extract is 0.04 grams (40mg) A dose of Salvia 250X extract is impossible."

# Questions



- What are the effects of Salvia?

(**Salvia  
Smoking  
Effects**)

» "He then took one large hit and held it in and laid back and began to feel his body getting heavy and his vision started to get this dim orange brownish light to it and he closed his eyes and moved his body around and it shook both of the feelings off."

(**Salvia  
Oral  
Effects**)

» "When chewed, the first effects are felt after about 15 minutes. After about 30 minutes, the full effects should be realized. Typical Salvia experiences last 5 to 10 minutes on average, with noticeable after-effects lasting up to 1/2 hour."

# Conclusion

- Online communities yield a large amount of candid data on a subject that is traditionally difficult to study
- In the future, we plan to consider many analyses beyond the topic model experiments shown here
- The f-LDA code will be made available