- Some slides on
  Paul and Dredze, 2012. Discovering Health Topics in Social Media Using Topic Models. PLOS ONE.
  http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0103408

- Brendan O'Connor, 2015-02-11
  for: http://people.cs.umass.edu/~brenocon/smacss2015/

Overview

- Goals
  - Measure public health from social media
  - (Background: Google flu trends)

- Contributions
  - Health-specific message filters.
  - Topic model (plus supervision!) to discover/infer people talking about ailments on Twitter
  - Exploratory analysis of correlations against ground-truth survey and illness tracking data from CDC
    - Keyword frequencies perform best!

Funnel underlying social-media-as-measurement

Inference we want

What actually happens

1. People segmented by geography or time have internal states (health, employment)
2. Some use social media service
3. Some write messages, conditioned on user's internal state
4. Researcher runs NLP algorithm (manually defined keywords? topic model?)
5. NLP output, segmented by geography or time

Or different models: media attention and common causes

What they did

- Collect tweets
  - Classify for health-relatedness.
    - P,R = 68,72
    - Are errors independent of QOI?
  - Geolocate: GPS plus user-supplied profile info
    - they open-sourced their system
- Topic model
- Correlate geo/temporal aggregates of tweet inferences, against CDC indicators

Geolocation: "Carmen"

- http://www.cs.jhu.edu/~mpaul/files/aaai13_geo.pdf
- Uses user-supplied "location" field in profile, compares to Yahoo Geolocation API (a placename => place entity linker)

Model ("ATAM")

- Switching (vector averaging) to combine word distributions
  - Contrast to multiplicative (log-additive) combinations. (other papers by Paul; Eisenstein; Roberts; Gourmley; etc.)
- Want to combine word distributions
  - Background
  - Ailment (symptom, treatment, other)
  - non-ailment Topic

Data collection/filtering

- There is no such thing as unsupervised analysis. Defining your dataset = critical human supervision. This part of the paper indicates extensive work and thought into the problem. Without this everything else would fail.

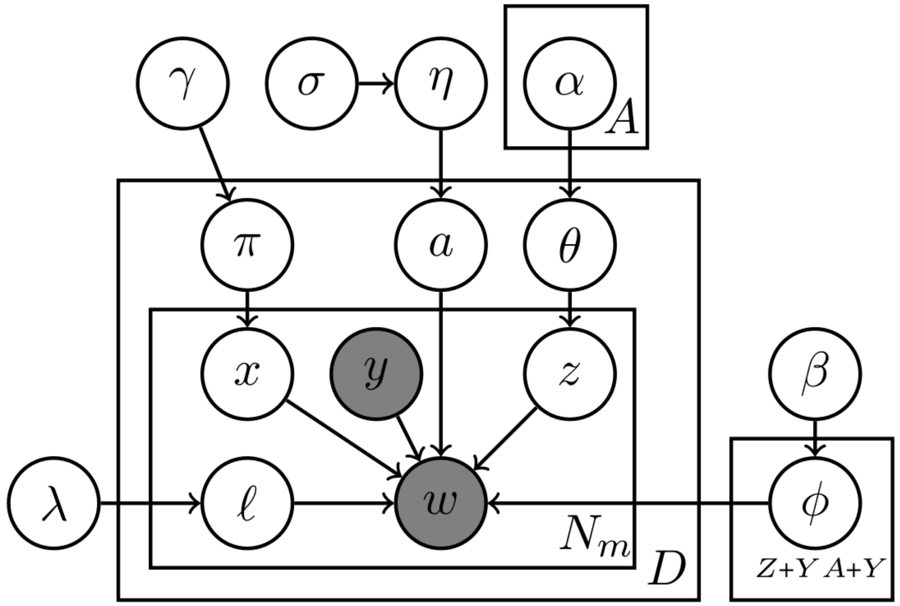
- General tweets, plus more selected (queried) by 20,000 health-related keyphrases from websites (plus “sick”, “doctor”)
- 20 health issues (“ailments”) from WebMD
  - Each issue has multiple articles about it (not clear .. the website defines a tagging/taxonomy?)
- Remove messages containing URLs (some of my papers do this too -- twitter {with, without} URLs are very different corpora)
- Message classifier: About the user’s health?
  - NOT: news, ads, non-English, or ambiguous
  - Human labeling (MTurk) => 5138 labeled messages

Model

- Briefly say that LDA conflates topics and ailments in prelim experiments
  - Made-up example: "damn flu, home with a fever watching TV"
- Ailment Topic-Aspect Model
  - Generative model of message texts, with latent variables
  - Every message has one *ailment*.
  - The set of ailments (i'll call the "ontology") is pre-defined from WebMD articles (not unsupervised!!!)
    - An ailment's unigram dists are prior-biased towards dist from a set of WebMD article about it
    - This prior is the only reason these have any interpretation as "ailments" !!!
      - Otherwise it's just a meaningless latent variable which may learn something meaningful, but you have to figure out -- like latent topics usually are
  - An ailment has 3 different worddists (three "aspects")
    - symptom worddist, treatment worddist, general/other worddist ... defined by the 20k keyphrases dictionary, which is from a different health website besides WebMD, it sounds like.
  - Words in a tweet are either from the background, or from a topic, or from an ailment vocabulary.
    - Extra twist: message ailment affects the topic selection for non-ailment words (e.g. flu => talk about TV?)

Compare

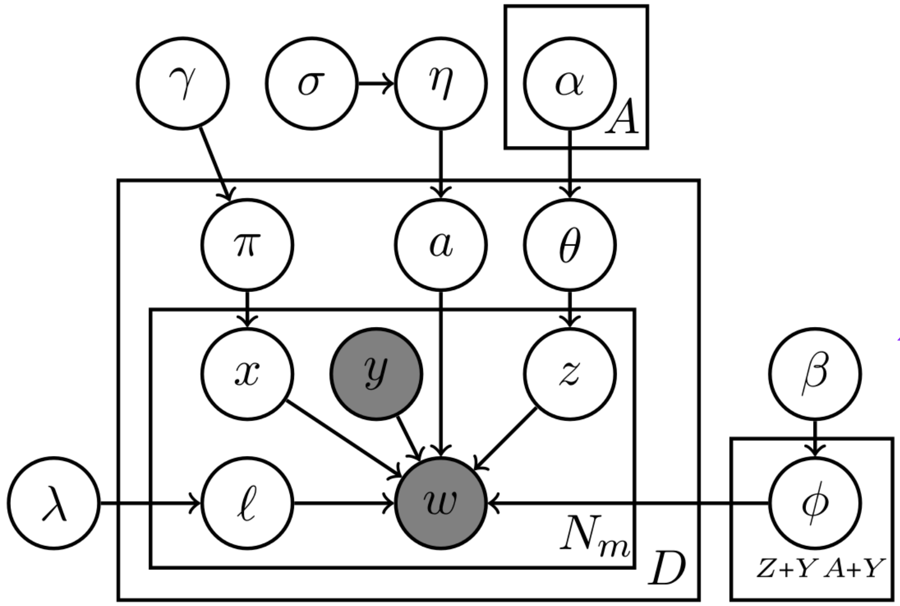
- single-membership unigram LM (Naive Bayes), matching each tweet against WebMD articles' worddist, to identify the ailment. Is this kind of what the model is doing? How different is it?
- the model adds more variability: you can talk about things other than the ailment.
- since i think the supervision from webmd seems important, i wish i had a sense how well this would do. maybe it would have lousy lexical coverage?



- Set the background switching binomial λ
- Draw an ailment distribution $\eta \sim \mathrm{Dir}(\sigma)$
- Draw word multinomials $\phi \sim \mathrm{Dir}(\beta)$ for the topic, ailment, and background distributions
- For each message $1 \leq m \leq D$:
  - Draw a switching distribution $\pi \sim \mathrm{Beta}(\gamma_0, \gamma_1)$
  - Draw an ailment $a_m \sim \eta$
  - Draw a topic distribution $\theta \sim \mathrm{Dir}(\alpha_a)$
  - For each token $1 \leq n \leq N_m$:
    - Draw aspect $y_n \in \{0, 1, 2\}$ (observed)
    - Draw background switcher $\ell_n \in \{0, 1\} \sim \lambda$
    - If $\ell_n == 0$:
      - Draw $w_n \sim \phi_{B,y_n}$ (background noise)
    - Else:
      - Draw $x_n \in \{0, 1\} \sim \pi$
      - If $x_n == 0$: (draw word from topic z)
        - Draw topic $z_n \sim \theta$
        - Draw $w_n \sim \phi_{T,z_n}$
      - Else: (draw word from ailment a aspect y)
        - Draw $w_n \sim \phi_{A,a_m y_n}$

Oh my

These things are typically less complex than they look in this format

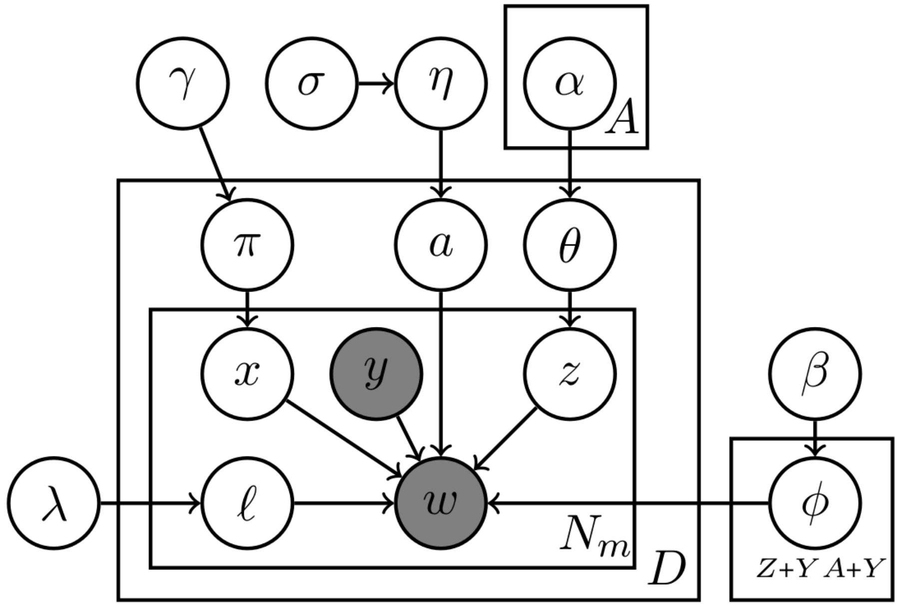


- Set the background switching binomial λ
- Draw an ailment distribution $\eta \sim \mathrm{Dir}(\sigma)$
- Draw word multinomials $\phi \sim \mathrm{Dir}(\beta)$ for the topic, ailment, and background distributions
- For each message $1 \leq m \leq D$:
  - Draw a switching distribution $\pi \sim \mathrm{Beta}(\gamma_0, \gamma_1)$
  - Draw an ailment $a_m \sim \eta$
  - Draw a topic distribution $\theta \sim \mathrm{Dir}(\alpha_a)$
  - For each token $1 \leq n \leq N_m$:
    - Draw aspect $y_n \in \{0, 1, 2\}$ (observed)
    - Draw background switcher $\ell_n \in \{0, 1\} \sim \lambda$
    - If $\ell_n == 0$:
      - Draw $w_n \sim \phi_{B,y_n}$ (background noise)
    - Else:
      - Draw $x_n \in \{0, 1\} \sim \pi$
      - If $x_n == 0$: (draw word from topic z)
        - Draw topic $z_n \sim \theta$
        - Draw $w_n \sim \phi_{T,z_n}$
      - Else: (draw word from ailment a aspect y)
        - Draw $w_n \sim \phi_{A,a_m y_n}$

This should be shaded: lexicons are partially observed from WebMD, as Dirichlet priors!!!!

m is left out!

Also the a -> theta dependence is missing



$$P(w_{dn}=v|a_d=i,y_{dn}=j,\theta,\phi,\lambda,\pi)=$$

$$(1-\lambda)\phi_{B,jv}+$$ Background model

$$\lambda[(1-\pi)(\sum_k \theta_{dk}\phi_{T,kv})+$$ Topic model

$$\pi\phi_{A,ijv}]$$ Ailment model

	message d, ailment i, token n within message
a_d	20-class: ailment for this message
w_{dn}	OBSERVED word for nth token of doc d
y_{dn}	OBSERVED 3-class aspect: symptom, treatment, or other
z_{dn}	Topic for this token
$\phi_{B,j}$	background dist for class $y=j$
m_i	OBSERVED, the WebMD worddist for ailment i
s_i	FIXED? scalar controlling webmd prior strength
$\beta_i = s_i m_i$	asymm Dir prior for ailment's worddist
$\phi_{A,ij} \sim Dir(\beta_i)$	ailment's worddist
α_i	ailment's topicdist prior (concentration fixed??)
$\theta_d \sim Dir(\alpha_{a_d})$	topic dist for doc
λ	FIXED to 0.2, rate of non-background words
π	rate of ailment vs. topic words

(If this is really high, this part approaches supervised NB)

biased, asymmetric dirichlets

- library(gtools)
- barplot(rdirichlet(1, c(1, 1, 1)), ylim=c(0,1))
- barplot(rdirichlet(1, c(.1, .1, .1)), ylim=c(0,1))
- barplot(rdirichlet(1, c(10,1,1)), ylim=c(0,1))

Modeling notes/questions

- collapsed gibbs sampling: easier than it looks! seriously. CGS's simplicity is a major reason to use dirichlet-multinom models.
- hyperparam inference
- large scale tricks: parallelization, subsamples

- i didn't at first understand definition of aspects y. seems important. but it sounds like they're from the 20,000 keyphrases drawn from different health websites. these keyphrases are partitioned into symptoms vs treatments, i guess.
- how much do posterior phis deviate from webmd prior? (how much does the supervision do?) If not much, this isn't "discovery". If it's a reasonable amount lot, maybe we should think of it as "lexicon enrichment", since the ontology is essentially fixed?
- in general, how much do you get out of the latent variable modeling?
  - COMPARE: single-membership unigram LM (Naive Bayes), matching each tweet against WebMD articles' worddist, to identify the ailment. Is this kind of what the model is doing? How different is it?
  - Is this a paper about unsup learning, or a paper about smart message classification plus smart use of lexical knowledge resources? smart use of lex knowledge is pretty great, so that's ok too!

Non-Ailment Topics						
TV & Movies	**Games & Sports**	**School**	**Conversation**	**Family**	**Transportation**	**Music**
watch	killing	ugh	ill	mom	home	voice
watching	play	class	ok	shes	car	hear
tv	game	school	haha	dad	drive	feelin
killing	playing	read	ha	says	walk	lil
movie	win	test	fine	hes	bus	night
seen	boys	doing	yeah	sister	driving	bit
movies	games	finish	thanks	tell	trip	music
mr	fight	reading	hey	mum	ride	listening
watched	lost	teacher	thats	brother	leave	listen
hi	team	write	xd	thinks	house	sound

	Ailments					
	Influenza-like Illness	**Insomnia & Sleep Issues**	**Diet & Exercise**	**Cancer & Serious Illness**	**Injuries & Pain**	**Dental Health**
General Words	better	night	body	cancer	hurts	dentist
	hope	bed	pounds	help	knee	appointment
	ill	body	gym	pray	ankle	doctors
	soon	ill	weight	awareness	hurt	tooth
	feel	tired	lost	diagnosed	neck	teeth
	feeling	work	workout	prayers	ouch	appt
	day	day	lose	died	leg	wisdom
	flu	hours	days	family	arm	eye
	thanks	asleep	legs	friend	fell	going
	xx	morning	week	shes	left	went
Symptoms	sick	sleep	sore	cancer	pain	infection
	sore	headache	throat	breast	sore	pain
	throat	fall	pain	lung	head	mouth
	fever	insomnia	aching	prostate	foot	ear
	cough	sleeping	stomach	sad	feet	sinus
Treatments	hospital	sleeping	exercise	surgery	massage	surgery
	surgery	pills	diet	hospital	brace	braces
	antibiotics	caffeine	dieting	treatment	physical	antibiotics
	fluids	pill	exercises	heart	therapy	eye
	paracetamol	tylenol	protein	transplant	crutches	hospital

Figure 2. Top words associated with ailments and topics. The highest probability words for a sample of ailments and non-ailment topics. The top ten general words are shown for ailments along with the top five symptom and top five treatment words. The top ten words are shown for topics. The names of the ailments and topics are manually assigned by humans upon inspection of the associated words.
doi:10.1371/journal.pone.0103408.g002

Results

- i'm really confused: does ATAM discover ailments or are they predefined to 20 with webmd priors?
- in general i'm not understanding the semantics of the model ... what parts of it are intended to do what, with how much supervision?
- topic coherence (learned word cluster) human evaluation, vs LDA: 11/18 good?

Results

- (Note: granularity affects correlation results!! e.g. my icwsm 2010 paper)

- Flu: correlate messages to CDC ILI at weekly granularity, all-USA
- Allergies: correlate messages to Gallup survey, at weekly granularity, all-USA
- Geographic trends: diet/exercise correlation against BRFSS (behav. risk factors, phone survey)

- Keywords do better than topic model's inferences?
  - Conclusion notes: topic model helps with keyword identification (my experience too)
  - My Q: are keywords subsumed by WebMD worddists? Or higher precision? Or...?
  - Keywords' efficacy likely depends on supervised filter pipeline

BRFSS

- http://apps.nccd.cdc.gov/brfss/

Table 4. Pearson correlations between various Twitter models and keywords and CDC BRFSS data for various serious illness risk factors.

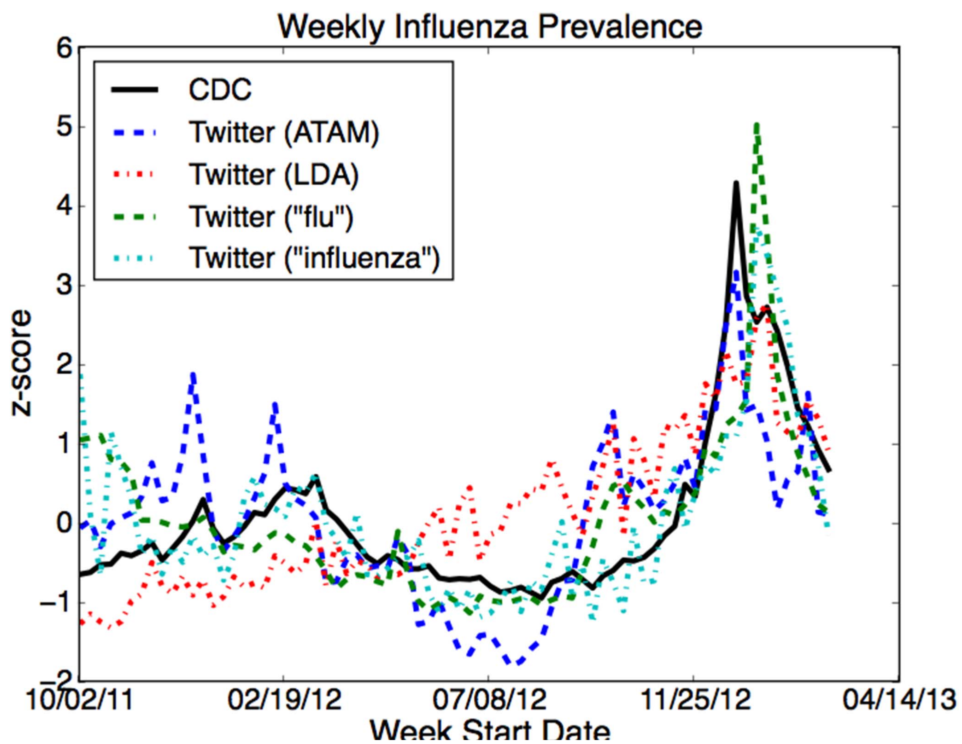
	Cancer	Tobacco	Heart Disease	Heart Attack
ATAM	.030	.069	.043	.080
LDA	−.045	−.005	−.069	−.023
"cancer"	−.037	−.180	−.232	−.181
"surgery"	−.049	.188	.021	.060

doi:10.1371/journal.pone.0103408.t004

Table 3. Pearson correlations between various Twitter models and keywords and CDC BRFSS data for various diet and exercise risk factors.

	Activity	Exercise	Obesity	Diabetes	Cholesterol
ATAM	.606	.534	−.631	−.583	−.194
LDA	.518	.521	−.532	−.560	−.146
"diet"	.546	.547	−.567	−.579	−.214
"exercise"	.517	.539	−.505	−.611	−.170

doi:10.1371/journal.pone.0103408.t003



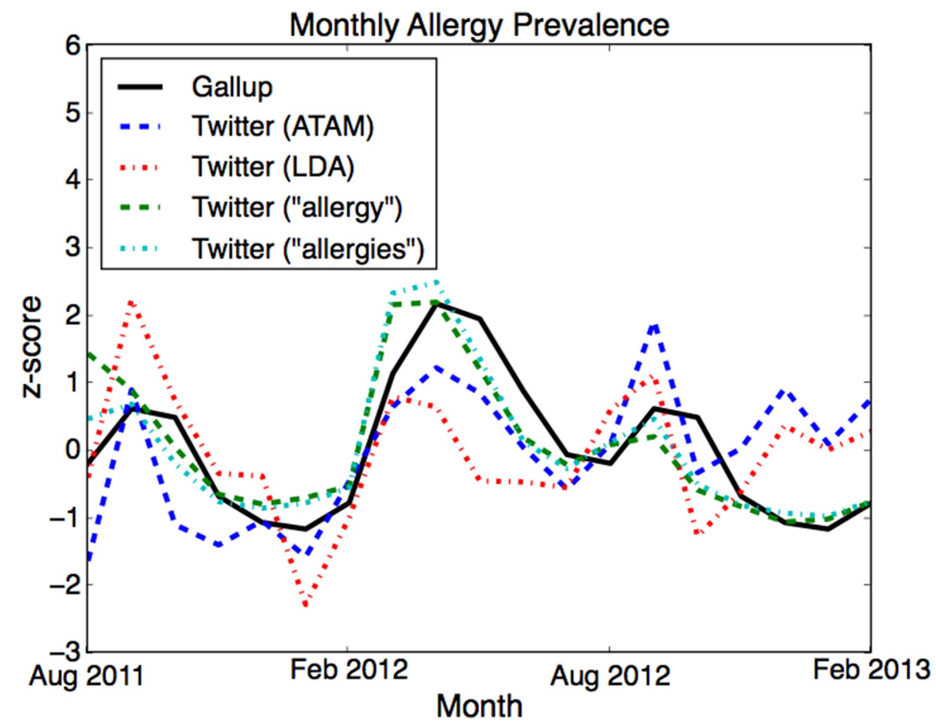


Table 1. Pearson correlations between various Twitter models and keywords and CDC influenza-like illness (ILI) surveillance data for three time periods.

	2011–12	2012–13	2011–13
ATAM	.613	.643	.689
LDA (1)	.670	.198	.455
LDA (2)	−0.421	.698	.637
"flu"	.259	.652	.717
"influenza"	.509	.767	.782

The two LDA rows correspond to two different LDA topics.

Table 2. Pearson correlations between various Twitter models and keywords and Gallup allergy survey data for two time periods.

	08/11–04/12	08/11–02/13
ATAM	.810	.479
LDA	.705	.366
"allergy"	.873	.823
"allergies"	.922	.877

The earlier period is the original data, while the data after April 2012 is from the previous year (05/2011–02/2012).
doi:10.1371/journal.pone.0103408.t002

Flu tracking

- Later paper (Lamb, Paul, Dredze NAACL 2013)

Data	System	2009	2011
Google	Flu Trends	0.9929	0.8829
Twitter	ATAM	0.9698	0.5131
	Keywords	0.9771	0.6597
	All Flu	0.9833	0.7247
	Infection	**0.9897**	**0.7987**
	Infection+Self	0.9752	0.6662

Table 4: Correlations against CDC ILI data: Aug 2009-Aug 2010, Dec 2011 to Aug 2012.

Supervised to predict CDC ILI trends from search query frequencies

ATAM to label flu-related tweets

Sup learning

- Supervised classifier for flu tweets, with
  - Flu related vs. not
  - Concerned Awareness vs. Infection
  - Self vs. Other