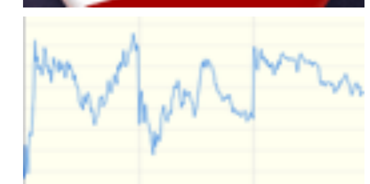
Recall Estimation for Rare Topic Retrieval from Large Corpuses

Praveen Bommannavar (Twitter), Alek Kolcz (Twitter),
Anand Rajaraman (Stanford)

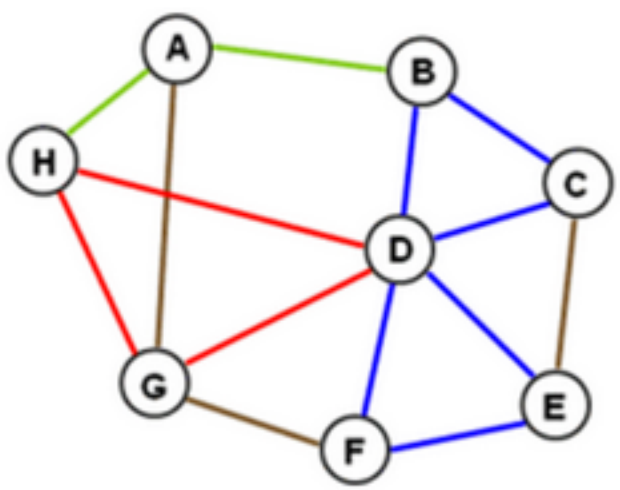
Mining Large Corpuses

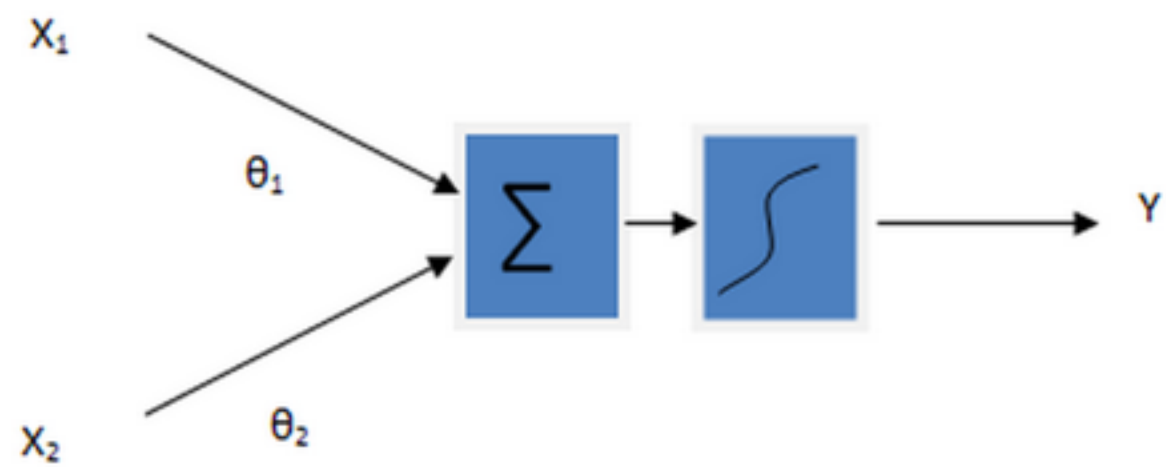
- Core offering of social media analytics companies
  - Analyze sentiment around products/brands
  - Estimate popularity of politicians
  - Uncover financial trends



Mining Large Corpuses

- Keyword filters, random walks, trained classifiers…





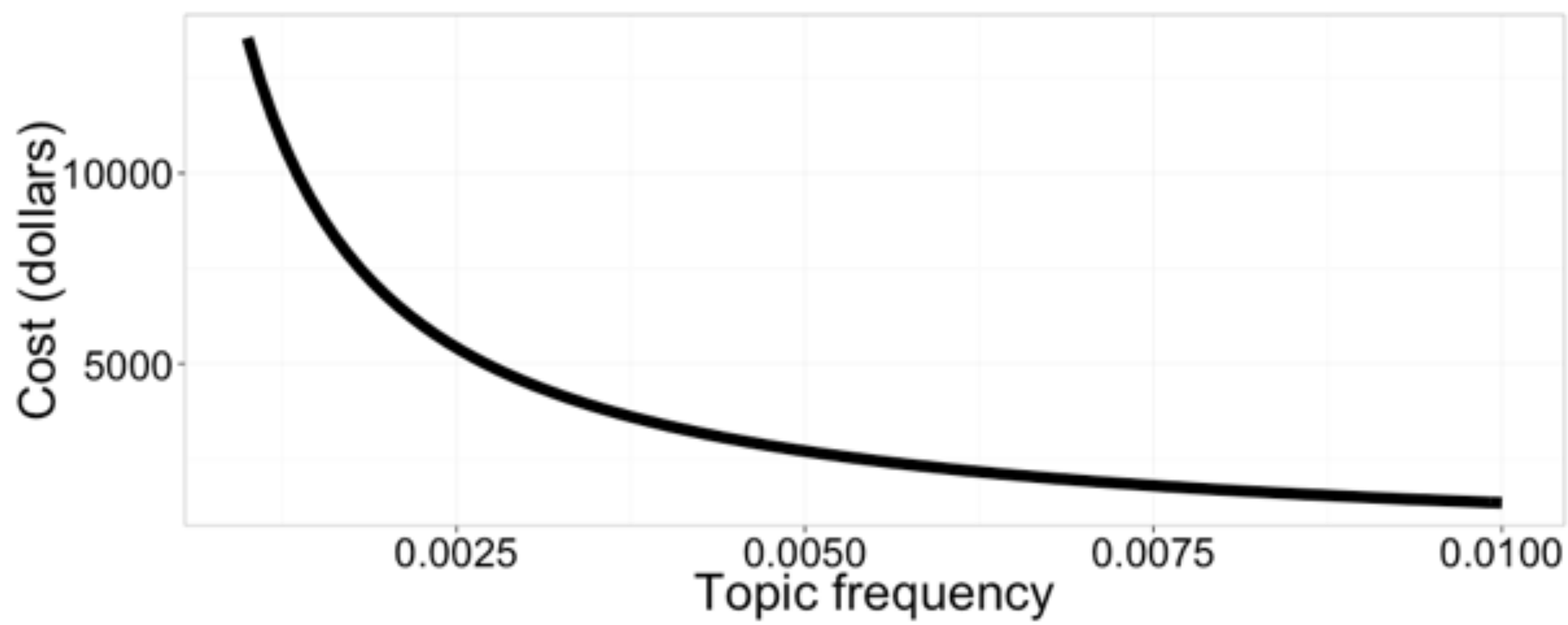
- With *any* approach: want **precision** and **recall**
    - Others: AUC, FPR, DCG, etc.

Metrics for Rare Topics

- Precision: sample positively classified docs
  - 384 samples for 95% confidence interval of size 0.1
  - Pay approximately $0.05 per evaluation => $19
- Recall: sample **all** docs to find enough true positives
  - Can be very expensive if topics are rare

Metrics for Rare Topics

- Skewed topic distribution => expensive recall est.



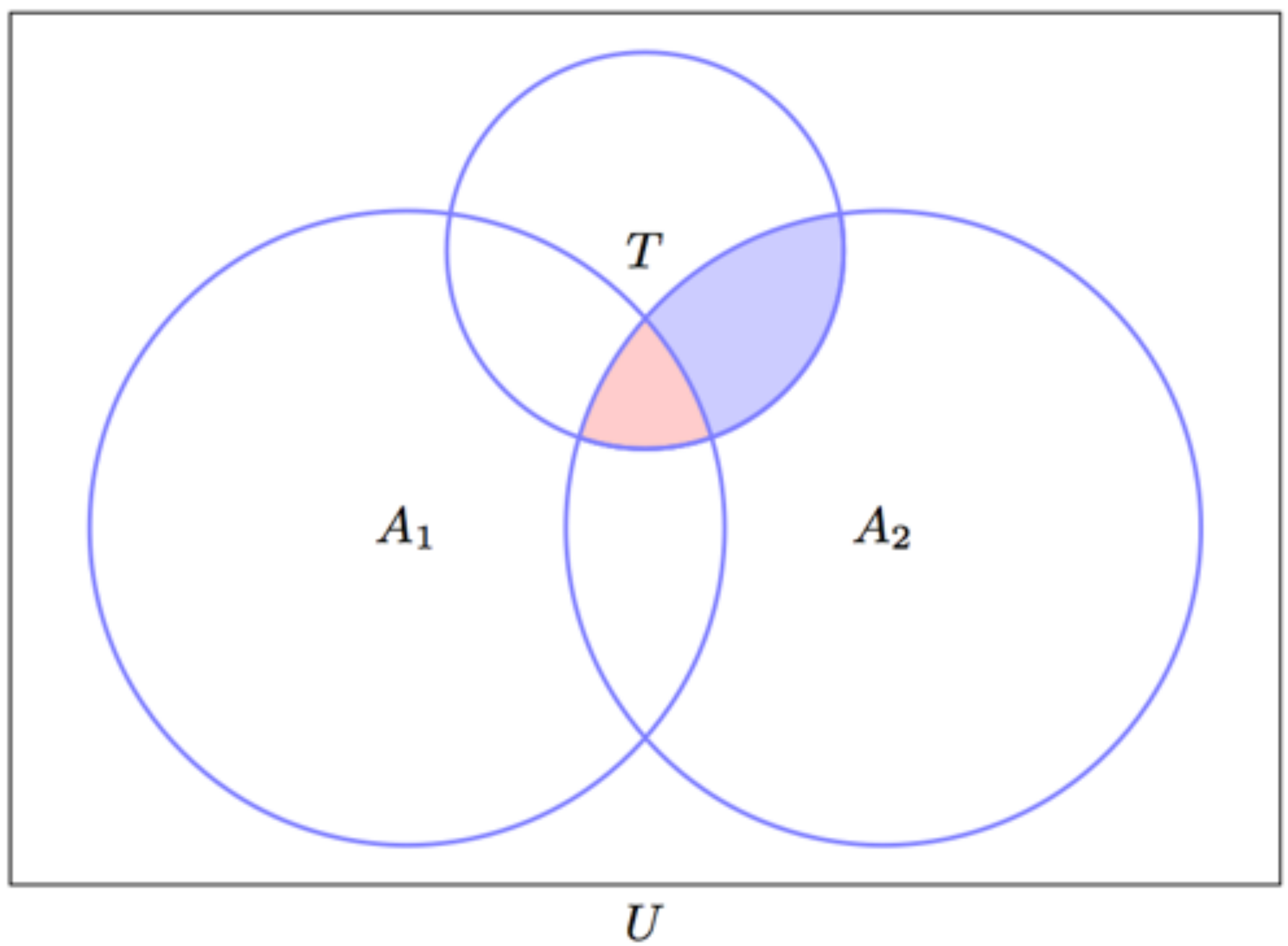
- Contribution: estimating recall on the cheap for rare topics

Related Work

- Calibration approaches for precision [Bennett & Carvalho]
- Confidence intervals for recall (frequent classes) [Webber]
- Counting positives despite inaccurate classification (frequent classes) [Forman]
- We emphasize **cost** and **rare** classes

Intuition

- Use pairs of *sufficiently independent* classifiers



Conditional Independence

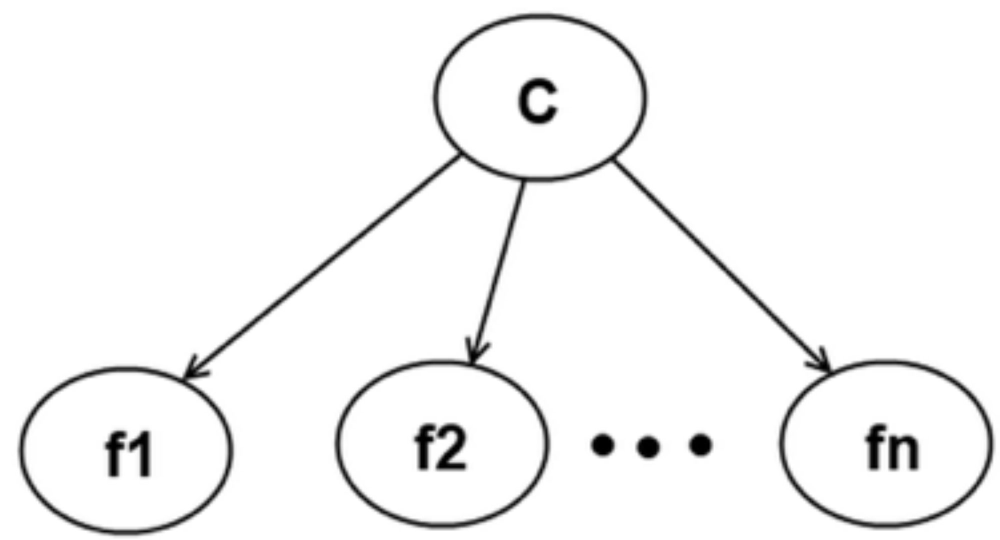
- T = set of on topic documents
- Classifiers C1, C2 return document sets A1, A2

Assumption 1. *(Conditional Independence 1) For the set of on-topic documents T, C_1 and C_2 are independent classifiers. That is,*

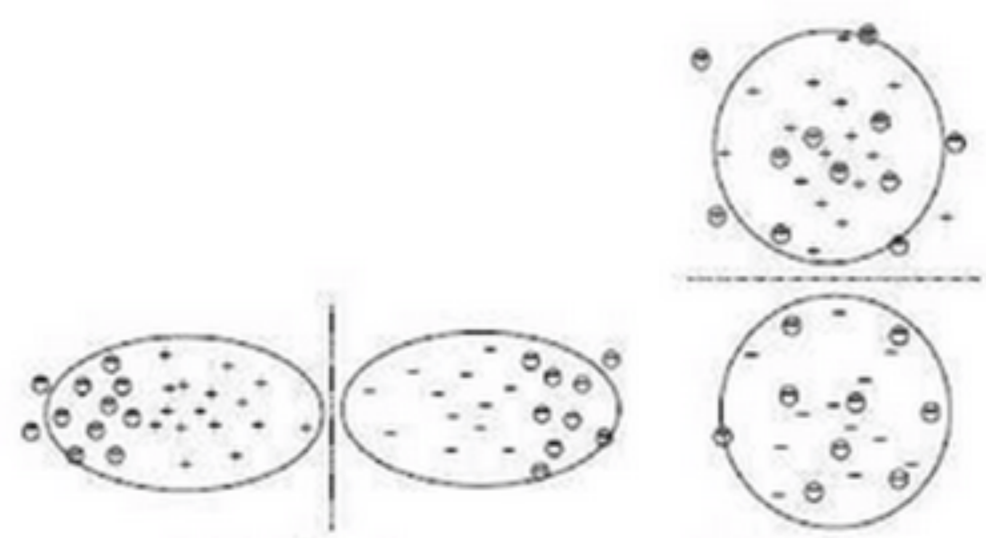
$$\delta_1 := \frac{P[A_1 \cap A_2 | T]}{P[A_1|T]P[A_2|T]} \approx 1.$$

Conditional Independence

- Naive Bayes



- Co-training



(a) x^1 view (b) x^2 view

Measuring Recall

- We can estimate recall using only precision!
- precision = P[T | A] recall = P [A | T]

$$r_1 = P[A_1|T] = \frac{P[A_1|T]P[A_2|T]}{P[A_2|T]} \approx \frac{P[A_1 \cap A_2|T]}{P[A_2|T]}$$

$$\begin{aligned} r_1 \approx \frac{P[A_1 \cap A_2|T]P[T]}{P[A_2|T]P[T]} &= \frac{P[A_1 \cap A_2 \cap T]}{P[A_2 \cap T]} \\ &= \frac{P[T|A_1 \cap A_2]P[A_1 \cap A_2]}{P[T|A_2]P[A_2]} \\ &= \frac{p_{12}|A_{12}|}{p_2|A_2|} \end{aligned}$$

Measuring Recall cont.

- What if we don't have joint classifier precision p12?
- With a couple more assumptions, we're still in luck:

Assumption 2. *(Conditional Independence 2) For the set of off-topic documents T^c, C_1 and C_2 are independent classifiers. That is,*

$$\delta_2 := \frac{P[A_1 \cap A_2 | T^c]}{P[A_1|T^c]P[A_2|T^c]} \approx 1.$$

Assumption 3. *(Sparsity) The number of on-topic documents T is small, as compared with the total universe of documents U. That is, $P[T] << 1$*

$$r_1 \approx \frac{|A_{12}|}{p_2|A_2|}\left[1 - \frac{(1-p_1)(1-p_2)|A_1||A_2|}{|U||A_{12}|}\right]$$

Constructing classifier pairs

- Great! Where do we get these classifier pairs from?
- Documents tend to be redundant; same info is expressed in different ways
    - Anchor text, headers, linked URLs, etc.
- Social media contains special structure

Dataset 1: sampled Tweets

- ~1M English language Tweets from Aug 6, 2012
  - topics: {apple, mars, obama, olympics, none}
  - Approx $20k budget to fully label

Table 2: Examples of seed terms and high Jaccard similarity neighbors (Twitter statuses).

[*apple*, #*apple*]	[#ios6,#ipad3,#iphone,hack,macintosh, iphones, #siri, ios, macbook, icloud, ipad, samsung, #ipodtouch, 4s, itunes, cydia, cider, #gadget, #tech, #tablet, app, connector, #mac, ...]
[*mars*, #*mars*]	[rover, nasa, #curiosity, #curiousity, image, mission, surface, #curiosityrover, bruno, milky, budget, @marscuriosity, gale, crater, orbiting, successfully, lands, landing, breathtaking ...]
[*obama*, #*obama*]	[@barackobama, barack, bush, #mitt2012, #obama2012, obamas, #dems, #gop, #military, romney, #idontsupportobama, potus, #president, administration, pres, #politics, voting ...]
[*olympics*, #*olympics*]	[medalist, gold, london, gb, kirani, gymnast, kate, sprinter, winning, won, #boxing, soccer, watch, watching, #usa, #teamgb, #canada, #london, javelin, nbc, match, 2012, 400m ...]

Dataset 1 recall estimates

- All recall estimates are within 0.10 absolute error and within 15% relative error
- O($1000) to O($10)

Table 3: Experimental results: tweet keyword filters. Both recall estimation schemes are within 0.10 absolute error and 15% relative error of the true recall for all topics.

Topic	$\|A\|$	$\|A_{seed}\|$	$\|A_{kw}\|$	$\|A_{joint}\|$	$\hat{p}_{seed}$	$\hat{p}_{kw}$	$\hat{p}_{joint}$	$\hat{r}^{(1)}_{seed}$	$\hat{r}^{(2)}_{seed}$	r_{seed}	$\hat{r}^{(1)}_{kw}$	$\hat{r}^{(2)}_{kw}$	r_{kw}
`Apple`	3038	676	10217	420	0.655	0.247	0.774	0.129	0.166	**0.146**	0.734	0.943	**0.830**
`Mars`	2372	1783	7703	1433	0.904	0.264	0.938	0.661	0.704	**0.680**	0.834	0.889	**0.857**
`Obama`	1253	851	7400	513	0.984	0.116	0.994	0.596	0.599	**0.668**	0.609	0.613	**0.683**
`Olympics`	23126	4595	45705	2688	0.986	0.330	0.989	0.176	0.178	**0.196**	0.587	0.593	**0.653**

Dataset 2: Twitter Stories

- 10.5M Discover stories from March 10, 2013: Tweets with hyperlinked URLs

- C1: tweet LR classifier, C2: web page LR classifier

- {ads and marketing, education, real estate and food, none}

Dataset 2: recall estimates

- Evaluation via random sampling (prevalent enough topics)
- All recall estimates within 0.10 absolute error and most are within 15% relative error

Table 4: Experimental results: story text and webpage logistic regression classifiers. Both recall estimation schemes are within 0.10 absolute error of the true recall for all topics and most topics are within 15% relative error.

Topic	$\lvert A_{tw}\rvert$	$\lvert A_{web}\rvert$	$\lvert A_{joint}\rvert$	$\hat{p}_{tw}$	$\hat{p}_{web}$	$\hat{p}_{joint}$	$\hat{r}^{(1)}_{tw}$	$\hat{r}^{(2)}_{tw}$	r_{tw}	$\hat{r}^{(1)}_{web}$	$\hat{r}^{(2)}_{web}$	r_{web}
Ads/Marketing	42073	76771	4369	0.825	0.698	0.900	0.073	0.077	**0.075**	0.113	0.120	**0.145**
Education	93292	76535	21426	0.827	0.868	0.873	0.282	0.319	**0.206**	0.242	0.275	**0.214**
Real Estate	42841	31978	12411	0.836	0.918	0.989	0.418	0.420	**0.413**	0.343	0.346	**0.380**
Food	42376	218507	20493	0.875	0.842	0.898	0.100	0.110	**0.122**	0.496	0.546	**0.522**

Dataset 3: ODP Entries

- 110K ODP entries - similar structure to Discover

> www.criticalsoftware.com - Develops and markets software products for business and mission critical information systems, and provide consulting and engineering services for enterprises.

- C1: description LR classifier, C2: web page LR classifier

- 12 topics

Dataset 3: recall estimates

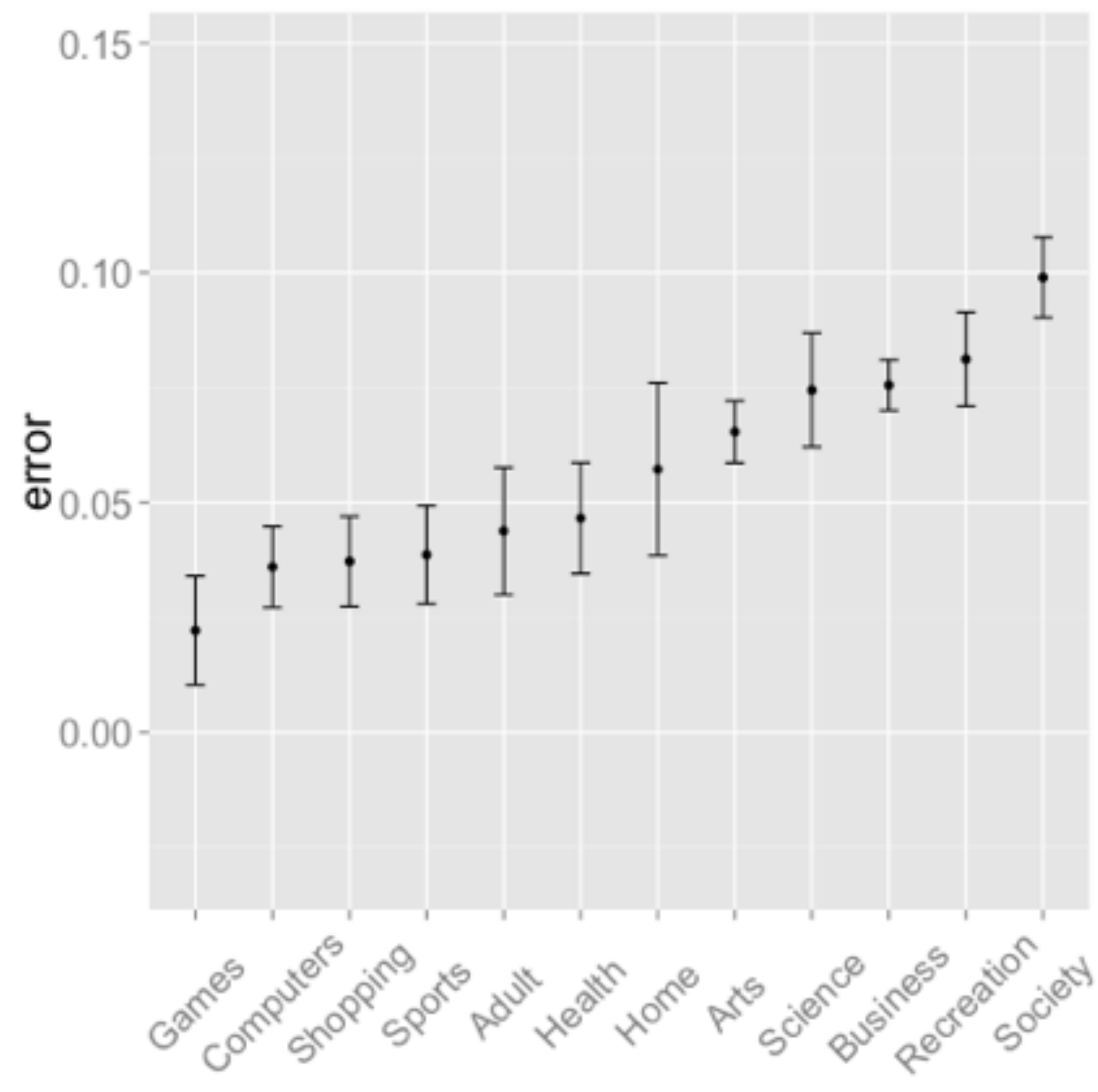
- Using joint precision directly is OK but Assumptions 2 and 3 break down

Table 5: Experimental results: prevalence and recall estimation in ODP records. Using joint precision directly gives high fidelity recall estimates for most topics, but attempting to approximate it results in poor recall estimates.

Topic	$\|A\|$	$\|A_{desc}\|$	$\|A_{web}\|$	$\|A_{joint}\|$	$\hat{p}_{desc}$	$\hat{p}_{web}$	$\hat{p}_{joint}$	$\hat{r}^{(1)}_{desc}$	$\hat{r}^{(2)}_{desc}$	r_{desc}	$\hat{r}^{(1)}_{web}$	$\hat{r}^{(2)}_{web}$	r_{web}
Adult	1470	10080	2815	757	0.089	0.430	0.710	0.624	0.878	**0.593**	0.839	1.180	**0.814**
Arts	13811	13719	12469	5523	0.524	0.766	0.865	0.581	0.671	**0.510**	0.771	0.891	**0.692**
Business	32304	18766	23888	10849	0.748	0.870	0.938	0.520	0.554	**0.443**	0.770	0.820	**0.646**
Computers	11235	11802	11756	4111	0.431	0.706	0.804	0.498	0.620	**0.453**	0.814	1.012	**0.746**
Games	2146	4245	3723	764	0.223	0.441	0.754	0.465	0.616	**0.439**	0.807	1.070	**0.766**
Health	5986	6180	6881	2991	0.576	0.667	0.872	0.649	0.744	**0.602**	0.837	0.960	**0.766**
Home	1546	7565	3616	643	0.118	0.299	0.574	0.594	1.035	**0.543**	0.717	1.249	**0.705**
Recreation	10846	9022	10712	4488	0.626	0.713	0.899	0.586	0.652	**0.505**	0.792	0.881	**0.703**
Science	5540	6005	7710	1706	0.380	0.462	0.658	0.474	0.720	**0.417**	0.739	1.123	**0.640**
Shopping	12386	13534	14610	3865	0.335	0.642	0.773	0.419	0.542	**0.375**	0.868	1.122	**0.757**
Society	12925	8397	11289	4071	0.627	0.720	0.922	0.501	0.543	**0.408**	0.774	0.839	**0.622**
Sports	6049	6929	6775	3197	0.550	0.708	0.910	0.664	0.729	**0.638**	0.835	0.918	**0.778**

Dataset 3: robustness

- Estimates obtained using Assumption 1 are robust
- Random 70-30 splits



Summary

- Have expressed recall estimates in terms of precision
- Precision is cheap to measure
- Conditionally independent classifiers can be constructed via redundancies in document structure
- **Possible future work:** Use multiple pairs of classifiers to stabilize recall estimates

Human Evaluation

- Not exactly a turn-key system
- What could go wrong?
    - Worker impatience, fatigue & boredom, domain/lingual proficiency, laziness/scammers, definitional issues, regional differences, etc..
- What does “on-topic” even mean anyway?

Human Evaluation

- Some remedies (not comprehensive)
  - Gold questions & agreement with other workers
  - Example answers to difficult/borderline questions (not just the easy ones)
  - Break down complex tasks into simpler ones (can’t expect workers to memorize a taxonomy)
  - **Communication**

Human Evaluation

- Sometimes workers don’t answer questions well, but many possible reasons. Don’t simply block!
- They rate you too…

Human Evaluation

- Ran a survey about biggest pain points:
  - Communication is at the top of the list
- After some soul searching:

Human Evaluation

- Email overload
  - "My dog jumped on my lap and hit my keyboard while I was working on this HIT. I'm sorry. If the answer my dog gave is wrong, I will understand the rejection. (The dog will get no treats for a week ...)"

- Other stray comments..
  - “Reading all these tweets has shattered the last little bit of hope I had for humanity. Holy hell people are stupid”