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 $\Rightarrow$ 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 $C_1, C_2$ return document sets $A_1, A_2$

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
Measuring Recall

• We can estimate recall using only precision!

• precision = $P[T | A]$          recall = $P[A | T]$

\[
\begin{align*}
  r_1 &= P[A_1 | T] = \frac{P[A_1 | T] P[A_2 | T]}{P[A_2 | T]} \approx \frac{P[A_1 \cap A_2 | T]}{P[A_2 | T]} \\
  r_1 &\approx \frac{P[A_1 \cap A_2 | T] P[T]}{P[A_2 | T] P[T]} = \frac{P[A_1 \cap A_2 \cap T]}{P[A_2 \cap T]} \\
  &= \frac{P[T | A_1 \cap A_2] P[A_1 \cap A_2]}{P[T | A_2] P[A_2]} \\
  &= \frac{p_{12} | A_{12}}{p_2 | A_2} 
\end{align*}
\]
Measuring Recall cont.

• What if we don’t have joint classifier precision $p_{12}$?

• With a couple more assumptions, we’re still in luck:

\[
\text{Assumption 2. (Conditional Independence 2) For the set of off-topic documents } T^c, \text{ } C_1 \text{ and } C_2 \text{ are independent classifiers. That is,}
\]
\[
\delta_2 := \frac{P[A_1 \cap A_2 | T^c]}{P[A_1 | T^c]P[A_2 | T^c]} \approx 1.
\]

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

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

| [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, #curiosity, 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

| Topic       | $|A_{tw}|$ | $|A_{web}|$ | $|A_{joint}|$ | $\hat{p}_{tw}$ | $\hat{p}_{web}$ | $\hat{p}_{joint}$ | $\hat{r}_{tw}^{(1)}$ | $\hat{r}_{tw}^{(2)}$ | $r_{tw}$ | $\hat{r}_{web}^{(1)}$ | $\hat{r}_{web}^{(2)}$ | $r_{web}$ |
|-------------|--------|---------|-------------|--------------|---------------|----------------|----------------------|----------------------|--------|----------------------|----------------------|--------|
| Ads/Marketing | 42073  | 76711   | 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** |

Table 4: Experimental results: story text and webpage logistic regression 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.
Dataset 3: ODP Entries

• 110K ODP entries - similar structure to Discover

• 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.

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

<table>
<thead>
<tr>
<th>AMT Requester</th>
<th>Rating [info]</th>
<th>Description</th>
<th>Date</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Praveen Bommannavar</td>
<td>FAIR: 1/5</td>
<td>Rejected my first 3 test hits within 5 minutes. He hasn’t responded back yet.</td>
<td>Sep 03 2012</td>
<td>flag</td>
</tr>
<tr>
<td>A1WBH67VFAHTUE</td>
<td>FAST: 1/5</td>
<td></td>
<td></td>
<td>comment</td>
</tr>
<tr>
<td>HIT Group »</td>
<td>PAY: 1/5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>COMM: 1/5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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”