Training the Cloud with the Crowd

Twitter + CrowdFlower + Google Prediction API

Kevin Cocco  SproutLoop.com
Weather Sentiment Prediction: Drag Marker Resize Radius Search within 96.26 miles of Arcadia, CA, USA

Positive 47% 😊 53% Negative

<table>
<thead>
<tr>
<th>GMT</th>
<th>Person</th>
<th>Tweet Text</th>
<th>Predicted Label</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:35,</td>
<td></td>
<td>Nice sunny day for all the ugly and pretty ladies <a href="http://t.co/SAp2XPCR">http://t.co/SAp2XPCR</a></td>
<td>positive</td>
<td>0.99802</td>
</tr>
<tr>
<td>Apr 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:34,</td>
<td></td>
<td>Good thing I didn't wear what I originally picked out because I would have</td>
<td>negative</td>
<td>0.547456</td>
</tr>
<tr>
<td>Apr 18</td>
<td></td>
<td>been FREEZING lol.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:34,</td>
<td></td>
<td>New blog post featuring some amazing @FreePeople dresses to incorporate</td>
<td>not_weather_related</td>
<td>0.989584</td>
</tr>
<tr>
<td>Apr 18</td>
<td></td>
<td>into your warm weather wardrobe! #freepeople</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://t.co/WF378WyH">http://t.co/WF378WyH</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:33,</td>
<td></td>
<td>@CollinMorrice22 Warm, but a tad on the windy side...roll on May</td>
<td>negative</td>
<td>0.551684</td>
</tr>
<tr>
<td>Apr 18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:33,</td>
<td></td>
<td>Okay, it's only MAYBE 50 degrees outside and yet someone turned on the air!</td>
<td>negative</td>
<td>0.999651</td>
</tr>
<tr>
<td>Apr 18</td>
<td></td>
<td>FREEZING!!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

goo.gl/qqWFL - http://www.sproutloop.com/prediction_demo
Creating & Using a Predictive Model

Twitter
• API 101

CrowdFlower
• Training data crowdsourcing

Google Prediction API
• About
• Upload - Train - Predict

Results & Discoveries
http://search.twitter.com/search.json?q=weather%20OR%20hot%20OR%20raining&rpp=20&lang=en&geocode=36.114,-115.172,10mi

• http://search.twitter.com/search.json?
  q=weather%20OR%20hot%20OR%20raining
  rpp=20
  lang=en
  geocode=36.114,-115.172,10mi
• **Streaming** API
• Twitter re-syndicators: DataSift.com & GNIP.com
• More... dev.twitter.com
DialogueEarth.org
• 120k+ Weather Related Tweets
• 5 CrowdFlower Judgements per Tweet
• Avg. $0.02 per Judgement / $0.10 per Unit(Tweet)
• http://crowdflower.com/docs/api

That sunny weather just pulled a whodini...straight disappeared!
What emotion does author express specifically about the weather? (required)
- Positive
- Negative
- I can't tell
- Neutral / author is just sharing information
- Tweet not related to weather condition
See table of emotions in instructions.
Predictive Modeling

LMNOP NLP/ML Model
Let's walk through some math:

\[
\mathbb{E}[\langle \alpha, f_j \rangle \Lambda_n \mid \mathcal{Y}_n] = \sum_{u=1}^{N} \sum_{v=1}^{M} \sigma_u^{-1} s_v^{-1} \exp \left\{ \frac{1}{2} Y_n^2 - \frac{1}{2} \frac{(Y_n - \mu_u)^2}{\sigma_u^2} \right\} \\
\times \exp \left\{ \frac{1}{2} C_n^2 - \frac{1}{2} \frac{(C_n - m_v)^2}{s_v^2} \right\} \\
\times \frac{\mathbb{E}[\Lambda_{n-1} \langle \alpha, f_j \rangle \langle Z_n^1, e_u^1 \rangle \langle Z_n^2, e_v^2 \rangle \mid \mathcal{Y}_{n-1}]}{\mathbb{E}[\Lambda_{n-1} \mid \mathcal{Y}_n]} \\
\times \frac{\mathbb{E}[\Lambda_{n-1} \mid \mathcal{Y}_n]}{E[\Lambda_{n-1} \mid \mathcal{Y}_n]} \\
E = mc^2
\]
Google Prediction API

- Process: **Upload** -> **Train** -> **Predict**
- Web Service, RESTful, OAuth
- Many modeling techniques, two types:
  - Regression: Estimating numeric value
  - Categorical: Choose a category of unstructured text
- Paid Service (free for first 6 months)
  - Base fee: $10 per/month - Includes 10k predictions
  - Predictions $0.50/1,000 - beyond initial 10k
  - Training $0.002/MB
- [https://developers.google.com/prediction/](https://developers.google.com/prediction/)
Google Prediction API

• **Upload** - Training Data
  - API, GSUtil, or Google Cloud Storage Manager

• Formatting Training Data
  - .CSV (label, feature1, feature2, ...)
  - "positive", "i love this weather @cancun"

• Google Refine
  - "Power tool for working with messy data"
  - [http://code.google.com/p/google-refine/](http://code.google.com/p/google-refine/)
Train - Build Model
  o API Console https://code.google.com/apis/console/
  o API Explorer https://code.google.com/apis/explorer/

Predict - Query model
  o RESTful API & libraries: Python, Java, PHP,...
  o JSON Response to Prediction "I love this weather"

```json
"outputLabel": "positive",
"outputMulti": [
  {
    "label": "negative",
    "score": 0.000202
  },
  {
    "label": "neutral",
    "score": 0.000122
  },
  {
    "label": "positive",
    "score": 0.995215
  }
]...
```
Crowd vs Cloud
Comparison of labels

Number of Tweets

- positive
- neutral
- negative
- not_weather
- cannot_tell

Crowdflower Labels
Google Prediction Labels
## Confusion Matrix

### Accuracy - Precision - Recall

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Crowdflower</th>
<th>Predicted Labels</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>negative</td>
<td>positive</td>
<td>not weather</td>
<td>neutral</td>
<td>cannot tell</td>
</tr>
<tr>
<td>negative</td>
<td>1824</td>
<td>196</td>
<td>251</td>
<td>300</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>positive</td>
<td>232</td>
<td>1732</td>
<td>200</td>
<td>215</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>not weather</td>
<td>294</td>
<td>216</td>
<td>2967</td>
<td>303</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>neutral</td>
<td>413</td>
<td>200</td>
<td>460</td>
<td>2263</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>cannot tell</td>
<td>94</td>
<td>61</td>
<td>64</td>
<td>86</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Accuracy** 0.71

<table>
<thead>
<tr>
<th>Sentiment</th>
<th></th>
<th>Precision</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>negative</td>
<td>positive</td>
<td>not weather</td>
<td>neutral</td>
<td>cannot tell</td>
</tr>
<tr>
<td>Precision</td>
<td>0.64</td>
<td>0.72</td>
<td>0.75</td>
<td>0.71</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.71</td>
<td>0.73</td>
<td>0.78</td>
<td>0.68</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>
Predictive Model Training Size

- Accuracy varies based on type of data and nuance
- More training data = better accuracy
Discoveries

• Tweet sentiment analysis is confusion for humans
  o All 5 C.F. workers agreed on tweets sentiment 44%
• When humans agree on tweet, model does well
  o 100% CF agreement = predictive accuracy 91%
• Google Prediction API Poor for:
  o Batch predictions
  o Model tuning
• Google Prediction API Good for:
  o Real time
  o Scaling - PaaS
  o Easy integration
Thank you!

Kevin Cocco
@kcocco | kcocco@sproutloop.com

SproutLoop
Simplified Building and Using Predictive Models

slides: goo.gl/Ws4Gh