CrowdHealth
A crowd-based platform for mining health information
(MASTODONS since April 15, 2014)
S. Amer-Yahia, E. Dublé, N. Ibrahim
(CNRS/LIG)
Collaborators:
Paul-Annick Davoine, Shady El Bassuoni, Léopold Fezeu, Christina Aschan-Leygonie, Hélène Mathian, Luc Merchez, Themis Palpanas
Interns:
Andrea Kalaitzakis, Sumit Sidana, Beatrice Valeri
Goal

• Study the relationship between demographics (geo location, age, gender) and:
  – Nutrition
  – Health (Light or chronic ailment)
  – Lifestyle (exercise, wellbeing)

• Impact : Help health institutions in their health and nutrition recommendations in different regions

• Data: Twitter, Nutrinet, Crowd4U
One million tweet map
displaying the last one million tweets

Tweets with ‘coffee’ (nutrition-related)
One million tweet map

displaying the last one million tweets

Zooming in..
One million tweet map
displaying the last one million tweets

Tweets with stomach pain (health-related)
Why Twitter?
M.J. Paul, M. Dredze, You are what you tweet: Analyzing Twitter for Public Health, International Conference on Weblogs and Social Media (ICWSM), 2011
Crowdhealth roadmap

- **Data collection (crowdsourced):** collect data from Twitter, Crowd4U, and Nutrinet

- **Data preparation:**
  - Tweets:
    - filter by user and construct a corpus for each user over several months
    - collect and classify tweets in 4 dimensions (nutrition, health, lifestyle) + that user’s demographics
  - Crowd4U:
    - collect and verify restaurant ratings in Grenoble

- **Data exploration:**
  - Extract and compare correlations between Twitter and Nutrinet
Twitter data collection

• Work by Andreas Kalaitz (master’s thesis)
  – 200 Million tweets collected (in Europe): stream of 1% available for free
  – Preliminary analysis and topic classification

• Tweets are obtained with associated metadata
  – User profile data (gender, user-specified location string, etc.)
  – Sometimes GPS coordinates (2% of tweets, if user activates it)

• Rate test (at 1 p.m. in Grenoble on July 5th 2014):

<table>
<thead>
<tr>
<th></th>
<th>tweets/second</th>
<th>tweets/year (rough est.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling stream (1% of all</td>
<td>46</td>
<td>1.5 billions</td>
</tr>
<tr>
<td>tweets)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geo-located tweets (World)</td>
<td>98*</td>
<td>3 billions</td>
</tr>
<tr>
<td>Geo-located tweets (Europe)</td>
<td>2</td>
<td>60 millions</td>
</tr>
</tbody>
</table>
The 200M tweets collection

• Collected using the stream API for several months
• How many tweets per user?
  – For a given user, we got around 5% of tweets only on average (because of the rate limitation)
  – 100 most active users: more than 50 tweets/user
  – 27000 least active users: less than 5 tweets/user
• Topic classification (using LDA):
  – 0.1% on health
  – 0.28% on nutrition
  – 0.22% on lifestyle
  – 1.2M users have a tweet on one of the dimensions
  – Only 5479 (out of 103M users) users have tweets on all 3 dimensions
• Data is sparse: we need users with tweets related to at least 2 dimensions
Data preparation cont’d

• Sumit Sidana
  – PhD student at IIT Kanpur India
  – Agenda:
    • Exploit Crowd4U to classify tweets along all dimensions
    • Extract correlations: location, nutrition and health
    • Verify extracted correlations
Data preparation using Crowd4U

- Classify geo-located tweets in 3 dimensions (nutrition, health, lifestyle) using Crowd4U

![Crowd4U microtask]

- Use an SVM classifier
Data exploration
(2 approaches)

• Extract correlations from one source and verify them on another

• Cross-source validation
  • Formalize a model to express hypotheses:
    vegetarians who live in Normandy are less prone than others to cardiovascular diseases
  • Adapt the model in [1] to express nutrition/health/geo hypotheses
  • Implement and verify source correlations on target data
    • using the method in [2] to extract correlations
    • using statistical measures in [1]

• On Crowd4U

[1] G. Liu, M. Feng, Y. Wang, L. Wong, S-K Ng, T. Mah, E. Lee, Towards exploratory hypothesis testing and analysis, ICDE’11 Proceedings of the 2011 IEEE 27th International Conference on Data Engineering
Restaurant ratings on Crowd4U

• Beatrice Valeri
  – PhD student at the Univ. of Trento
  – Internship (June+July+October 2014)
• Shady El Bassuoni
  – Prof. at the American Univ. in Beirut
• Developed algorithms to correlate ratings with demographics and other restaurants and verify consistency using a core set of trusted users
Crowdhealth roadmap revisited

• Goal of CrowdHealth
  – Study relationship between geo location, nutrition, health (light ailments)

• Crowdsourcing used for:
  – Data collection (Twitter, Crowd4U) + Nutrinet
  – Data preparation: annotate tweets in Crowd4U
  – Data exploration: test extracted correlations

• Need for interactive visualization in data exploration
Current Visualization
Mufida Miratul-Khusna (intern)

Nutrition in France

<table>
<thead>
<tr>
<th>Food</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>27%</td>
</tr>
<tr>
<td>Tea</td>
<td>6.6%</td>
</tr>
<tr>
<td>Chocolate</td>
<td>51%</td>
</tr>
<tr>
<td>Other</td>
<td>15.4%</td>
</tr>
</tbody>
</table>

Coffee 232 (51%)

Health in France

<table>
<thead>
<tr>
<th>Condition</th>
<th>Total Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headache</td>
<td>86</td>
</tr>
<tr>
<td>Stomach ache</td>
<td>90</td>
</tr>
<tr>
<td>Insomnia</td>
<td>72</td>
</tr>
<tr>
<td>Anxiety</td>
<td>60</td>
</tr>
<tr>
<td>Depression</td>
<td>50</td>
</tr>
<tr>
<td>Asthma</td>
<td>40</td>
</tr>
</tbody>
</table>

Total Population: 268