DISCOVERING POPULAR EVENTS FROM TWEETS

Calin RAILEAN
Alexandra MORARU
Outline

- Introduction
- Dataset Description
- NEsper tool
- Association of tweets and events
- Results & Evaluation
- Conclusions
INTRODUCTION

- Social events happening in a city can influence and affect a large number of the citizens

- Different metrics to measure the popularity of such events can be useful

- Social media channels report about such events
Goal

- Determining the popularity of social events (i.e. music concerts) based on their presence in social media (i.e. tweets).
Goal

- Determining the popularity of social events (i.e. music concerts) based on their presence in social media (i.e. tweets).

- The larger the number of tweets associated to an event, the more popular the event is.
Dataset Description

- 10033 social events
  - Eventful.com
  - event title, start and stop time of event, type of event, location, performers name, short bio description etc.

- Over 4 million tweets
  - tweet text, hash tags, time of posting the tweet, geographical coordinates etc.

- London, March 6th to April 11th 2013
Dataset Preprocessing

- Parsing JSON format for tweets and XML for events
- Tweet and Event class (C#)
- Missing value: stop time of events
  - We calculate stop time as median value for each type of event
NEsper

- Event Stream Processing (ESP)
  - Processing streaming data related to events that are happening
- Complex Event Processing (CEP)
  - Event processing that combines data from multiple sources
- Event Processing Language (EPL)
  - Contains queries that has been designed for similarity with the SQL query language
Preparing Input for NEsper

- Assign to NEsper types of objects it will receive: tweets(Tweet class) and social events(Event class)
- Create a pattern in EPL syntax by using the unary operator “every” and the operator followed-by “->”
Pattern
every Event -> every Tweet
(event.Stop_Time-tweet.Time>0)
**Association Event->Tweet**

- We use an *association coefficient* (AC) defined by the next formula:
  
  $$AC = 0.5 * P + 0.25 * W + 0.125 * L + 0.125 * B,$$

- $P = 1$ if tweet text contains the event’s performer name
- $W = $ the ratio between the tweet’s words matching the event’s title and the total number of words in the event’s title
- $L = 1$ if location name is found in the tweet’s text
- $B = 1$ if the tweet’s text contains short-bio description of the performer
15455 tweets correlated with 572 music events having the AC higher than 0.25

<table>
<thead>
<tr>
<th>Event Title</th>
<th>Tweet</th>
<th>AC</th>
<th>Popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olly Murs</td>
<td>I'm at O2 Arena - @the_o2 for Olly Murs, Tich and Loveable Rogues (Greenwich, Greater London) w/ 8 others <a href="http://t.co/lDTRyUTCpb">http://t.co/lDTRyUTCpb</a></td>
<td>0.75</td>
<td>23</td>
</tr>
<tr>
<td>Halestorm</td>
<td>I'm at @eballroomcamden for Halestorm, In This Moment and Sacred Mother Tongue</td>
<td>0.75</td>
<td>9</td>
</tr>
<tr>
<td>Beyonce</td>
<td>#NowWatching @Beyonce #LifeisButADream</td>
<td>0.75</td>
<td>9</td>
</tr>
<tr>
<td>The Script</td>
<td>O2 arena the script <a href="http://t.co/nItxTURl88">http://t.co/nItxTURl88</a></td>
<td>0.75</td>
<td>54</td>
</tr>
<tr>
<td>Bastille</td>
<td>Seeing &quot;Bastille&quot;. I am cool and with it. (@ O2 Shepherd's Bush Empire - @o2sbe w/ 7 others) <a href="http://t.co/BweCOsv4s5">http://t.co/BweCOsv4s5</a></td>
<td>0.75</td>
<td>106</td>
</tr>
<tr>
<td>Thursday Night</td>
<td>&quot;#bigreunion concert on a Thursday night. Loving it!</td>
<td>0.25</td>
<td>98</td>
</tr>
<tr>
<td>Union</td>
<td>tweeting union council agenda avidly from @UKMStudentLive</td>
<td>0.25</td>
<td>16</td>
</tr>
<tr>
<td>Over The Moon</td>
<td>Because of this, ive lost my faith in humanity! I'm done, <em>disappears to moon</em></td>
<td>0.25</td>
<td>21</td>
</tr>
<tr>
<td>Everything on Red! - Columbia - Sabre Tooth Monk</td>
<td>No red card will ever too that</td>
<td>0.535</td>
<td>118</td>
</tr>
</tbody>
</table>
Evaluation methodology

- Manually evaluated a random set of 100 associations of events and tweets with AC higher than 0.25
- Two human annotators have analyzed the tweet and the event title and evaluated them as correct or incorrect
- Calculate inter-annotator agreement for 100 associations (Cohen coefficient), 2 human annotators.
Evaluation

- Cohen's kappa coefficient is a statistical measure of inter-annotator agreement for qualitative items.

\[ k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \]

- where \( Pr(a) \) is the relative observed agreement among annotators, and \( Pr(e) \) is the hypothetical probability of chance agreement.
Evaluation

- Cohen's kappa coefficient is a statistical measure of inter-annotator agreement for qualitative items.

\[ k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \]

- where \( Pr(a) \) is the relative observed agreement among annotators, and \( Pr(e) \) is the hypothetical probability of chance agreement.

- 0.661 – substantial level of agreement
Evaluation

Values of AC for the associations of tweets and events evaluated
Evaluation

Values of AC for the associations of tweets and events evaluated.

Precision performance for different values of AC.
Application
Application
Conclusions

- We have proposed and evaluated a method for discovering popular events based on tweets.
- The results show a positive outcome, validating the proposed solution.
  - The precision can be increased by setting a higher threshold for the AC coefficient.
Conclusions

- We have proposed and evaluated a method for discovering popular events based on tweets.
- The results show a positive outcome, validating the proposed solution
  - The precision can be increased by setting a higher threshold for the AC coefficient
- Possible improvements
  - Including geo-location parameters in the AC equation,
  - improving the preprocessing of data (extending the stop-word list or by including NLP techniques)
Thank you for your attention! Questions?