DISCOVERING POPULAR EVENTS FROM TWEETS

Calin RAILEAN

Alexandra MORARU

ailab.ijs.si

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 milion tweets
  - tweet text, hash tags, time of posting the tweet, geographical coordinates etc.

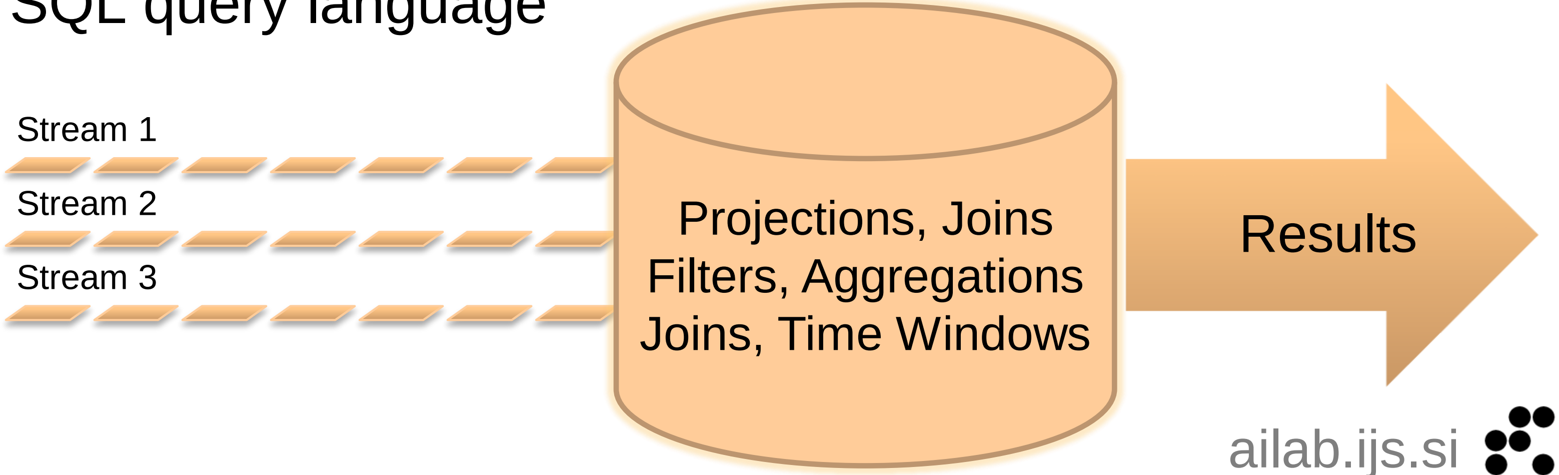
- London, March 6th to April 11th 2013

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

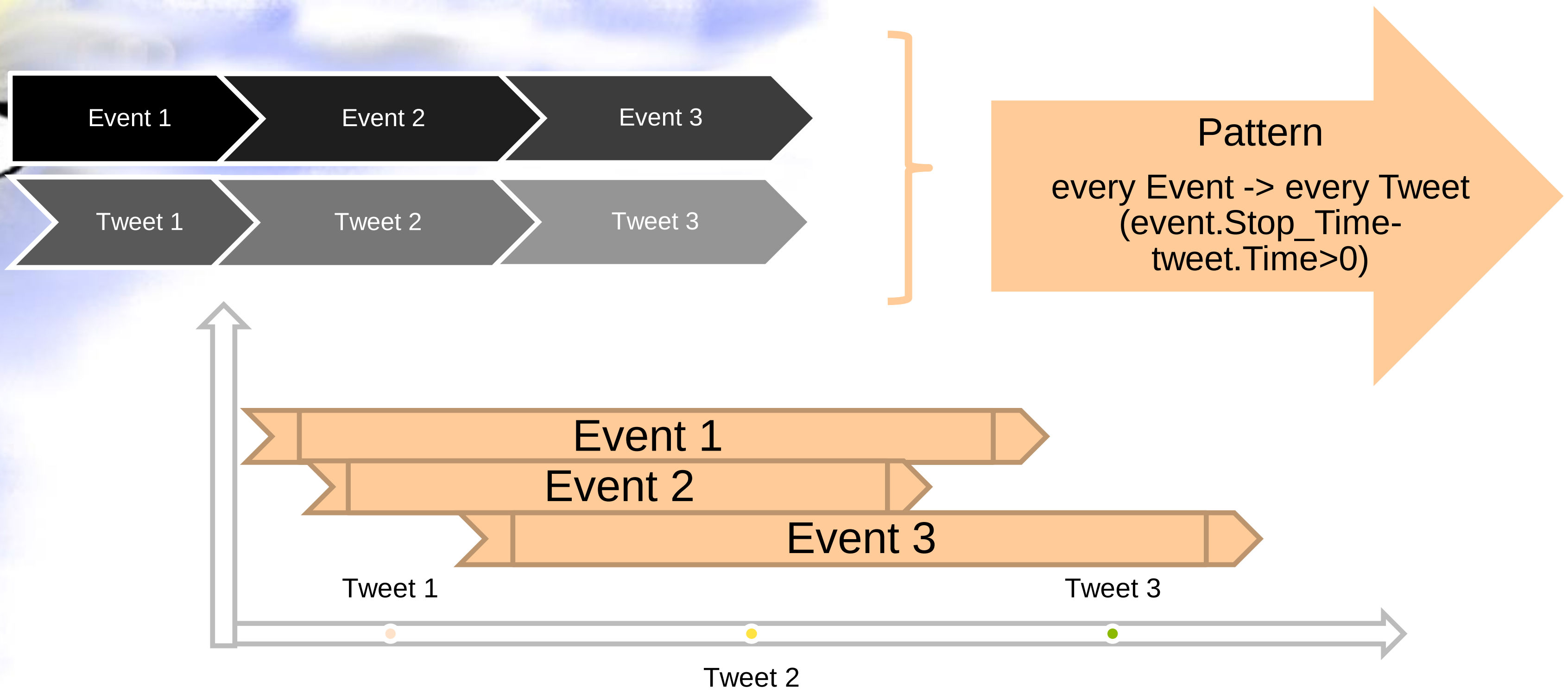
NEsper

- 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)

Time 1

- Event 1 -> Tweet 1
- Event 2 -> Tweet 1

Time 2

- Event 1 -> Tweet 2
- Event 2 -> Tweet 2
- Event 3 -> Tweet 2

Time 3

- Event 3 -> Tweet 3

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

Results

- 15455 tweets correlated with 572 music events having the AC higher than 0.25

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

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 coefficent), 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

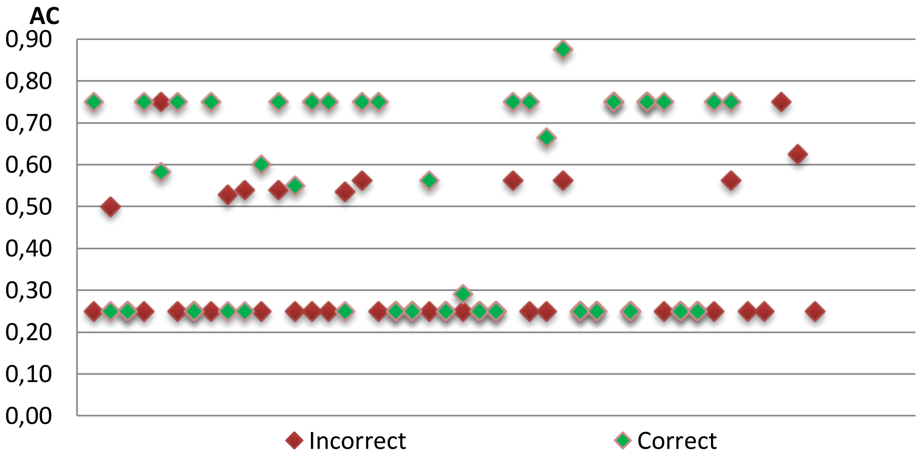
- Cohen's kappa coefficient is a statistical measure of inter-annotator agreement for qualitative items.

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

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

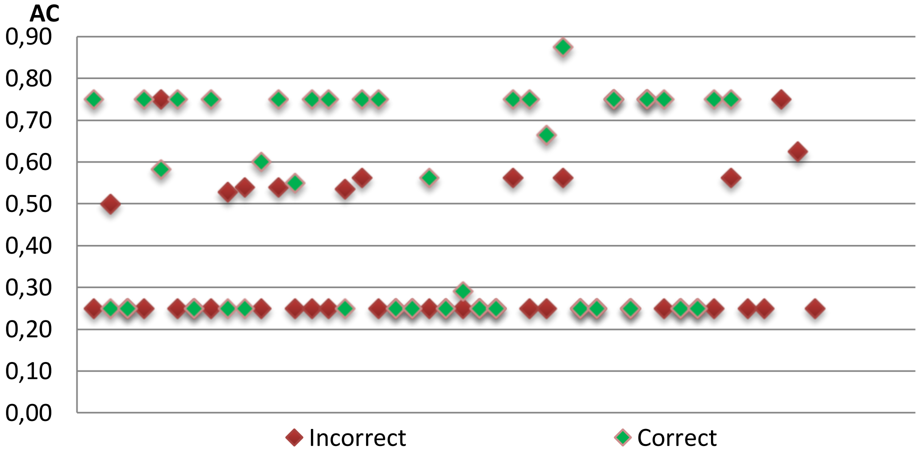
- 0,661 – substantial level of agreement

Evaluation

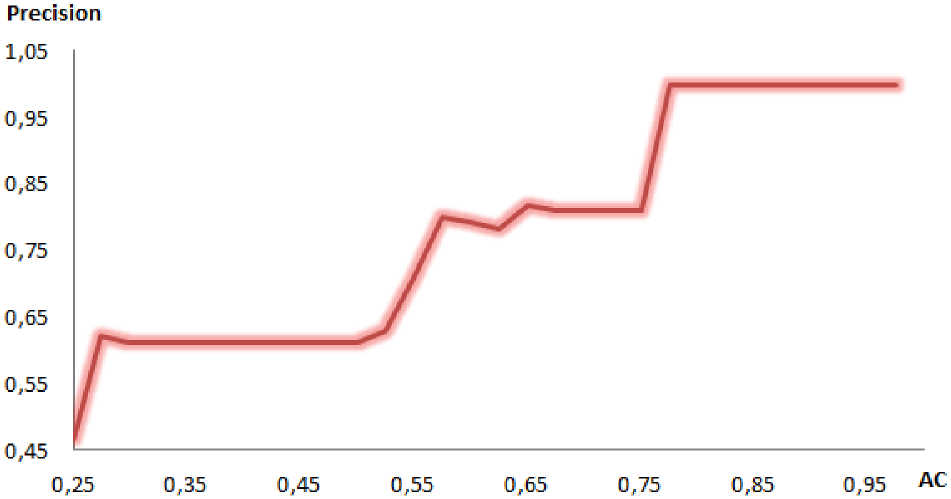


- ***Values of AC for the associations of tweets and events evaluated***

Evaluation

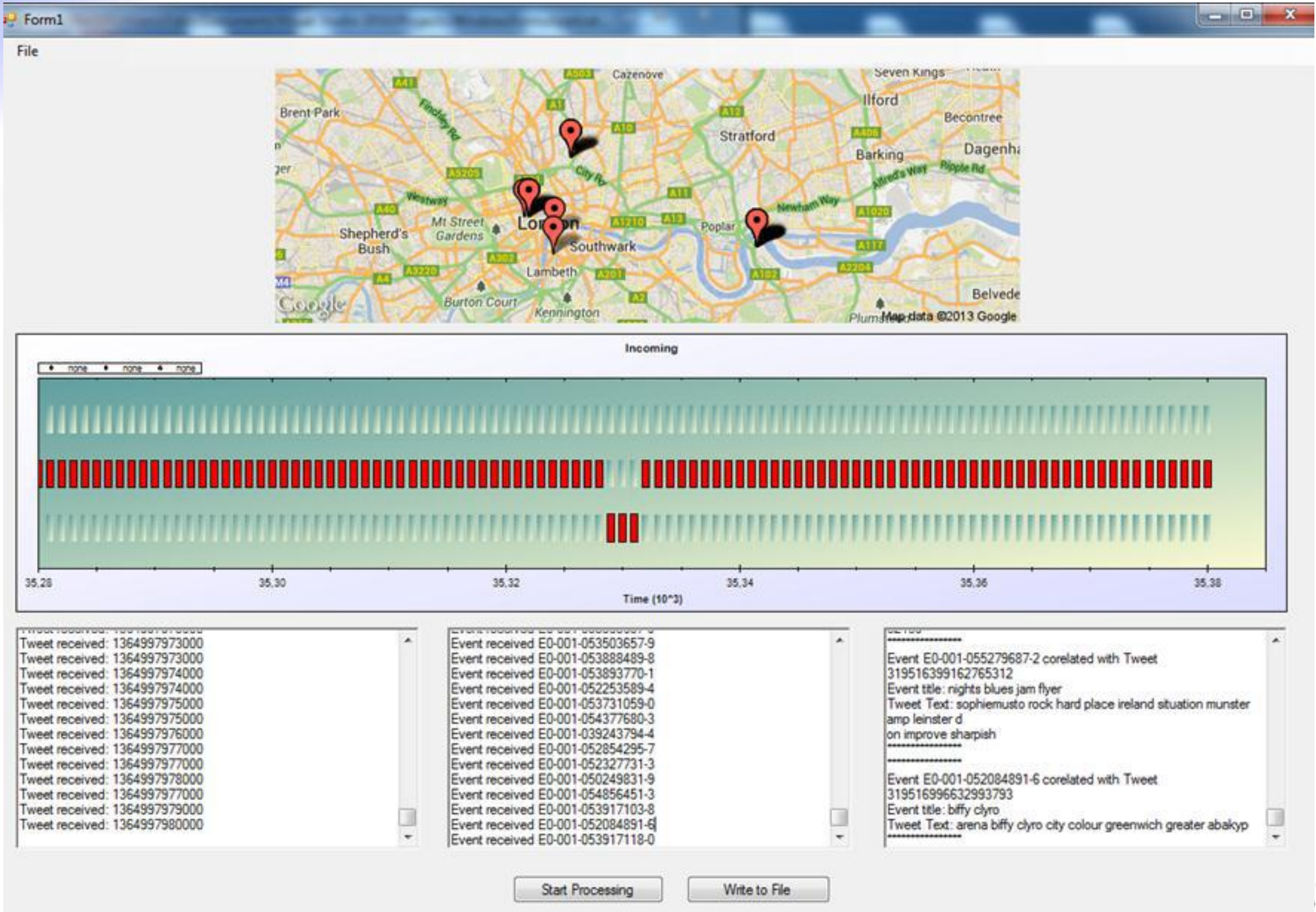


- ***Values of AC for the associations of tweets and events evaluated***



- ***Precision performance for different values of AC***

Application



```
Tweet received: 1364997973000
Tweet received: 1364997973000
Tweet received: 1364997974000
Tweet received: 1364997974000
Tweet received: 1364997975000
Tweet received: 1364997975000
Tweet received: 1364997976000
Tweet received: 1364997977000
Tweet received: 1364997977000
Tweet received: 1364997978000
Tweet received: 1364997977000
Tweet received: 1364997979000
Tweet received: 1364997980000
```

```
Event received E0-001-053503657-9
Event received E0-001-053888489-8
Event received E0-001-053893770-1
Event received E0-001-052253589-4
Event received E0-001-053731059-0
Event received E0-001-054377680-3
Event received E0-001-039243794-4
Event received E0-001-052854295-7
Event received E0-001-052327731-3
Event received E0-001-050249831-9
Event received E0-001-054856451-3
Event received E0-001-053917103-8
Event received E0-001-052084891-6
Event received E0-001-053917118-0
```

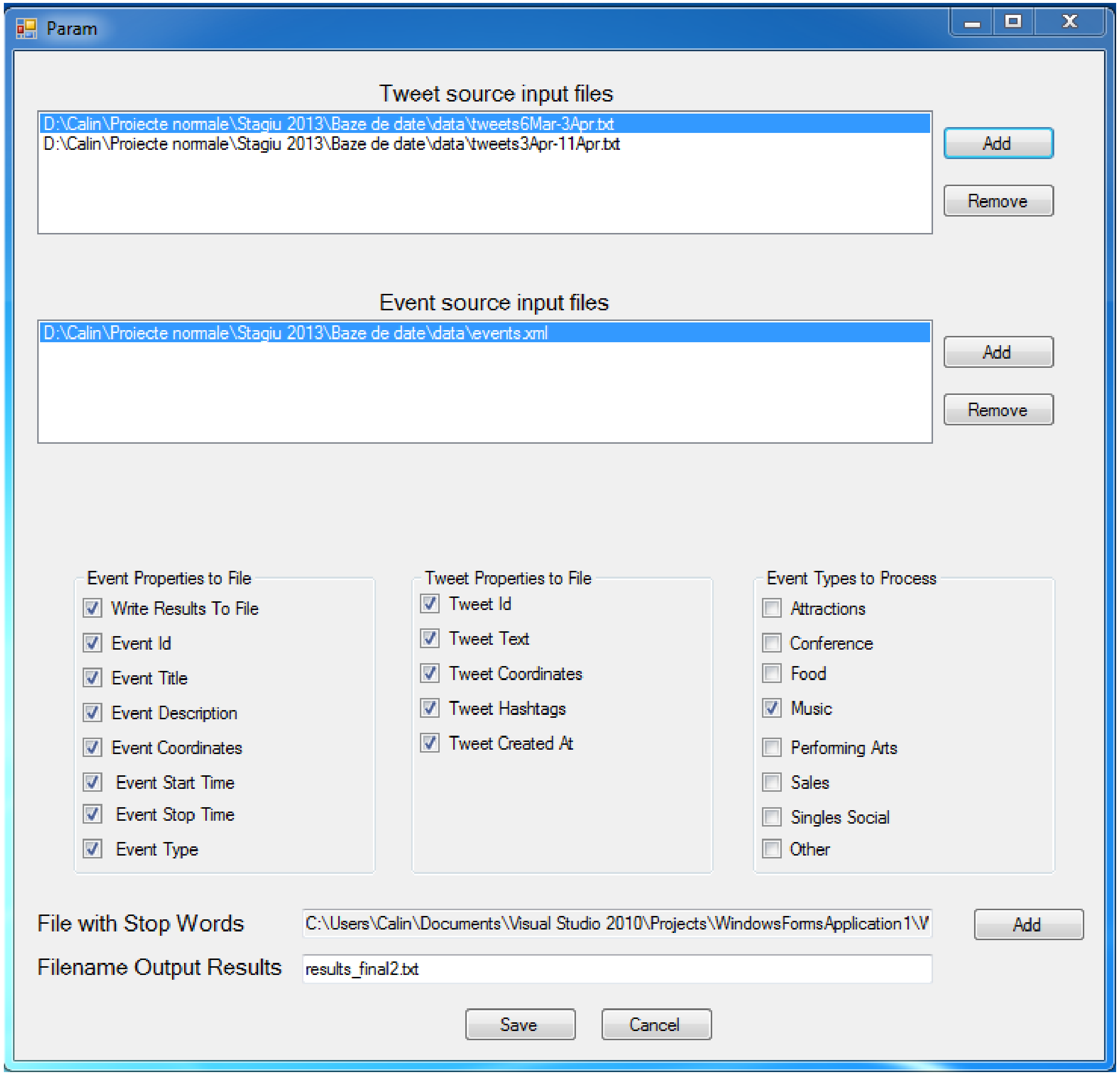
```
****************

Event E0-001-055279687-2 corelated with Tweet
319516399162765312
Event title: nights blues jam flyer
Tweet Text: sophiemusto rock hard place ireland situation munster
amp leinster d
on improve sharpish
****************

****************

Event E0-001-052084891-6 corelated with Tweet
319516996632993793
Event title: biffy clyro
Tweet Text: arena biffy clyro city colour greenwich greater abakyp
****************
```

Application



Param

Tweet source input files

D:\Calin\Proiecte normale\Stagiu 2013\Baze de date\data\tweets6Mar-3Apr.txt
D:\Calin\Proiecte normale\Stagiu 2013\Baze de date\data\tweets3Apr-11Apr.txt

Event source input files

D:\Calin\Proiecte normale\Stagiu 2013\Baze de date\data\events.xml

Event Properties to File

- ☑ Write Results To File
- ☑ Event Id
- ☑ Event Title
- ☑ Event Description
- ☑ Event Coordinates
- ☑ Event Start Time
- ☑ Event Stop Time
- ☑ Event Type

Tweet Properties to File

- ☑ Tweet Id
- ☑ Tweet Text
- ☑ Tweet Coordinates
- ☑ Tweet Hashtags
- ☑ Tweet Created At

Event Types to Process

- □ Attractions
- □ Conference
- □ Food
- ☑ Music
- □ Performing Arts
- □ Sales
- □ Singles Social
- □ Other

File with Stop Words C:\Users\Calin\Documents\Visual Studio 2010\Projects\WindowsFormsApplication1\W

Filename Output Results results_final2.txt

ailab.ijs.si

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?