

BIG DATA ASSIGNMENT

JAN RYBÁK


MINI PROJECT

 **Spark**

+

 python

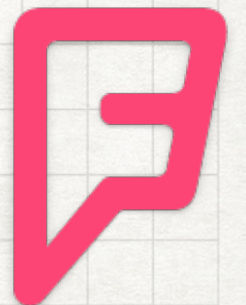
 **Scala**

 Java™

MINI PROJECT

Task #1:

Analysis of Foursquare dataset



Task #2:

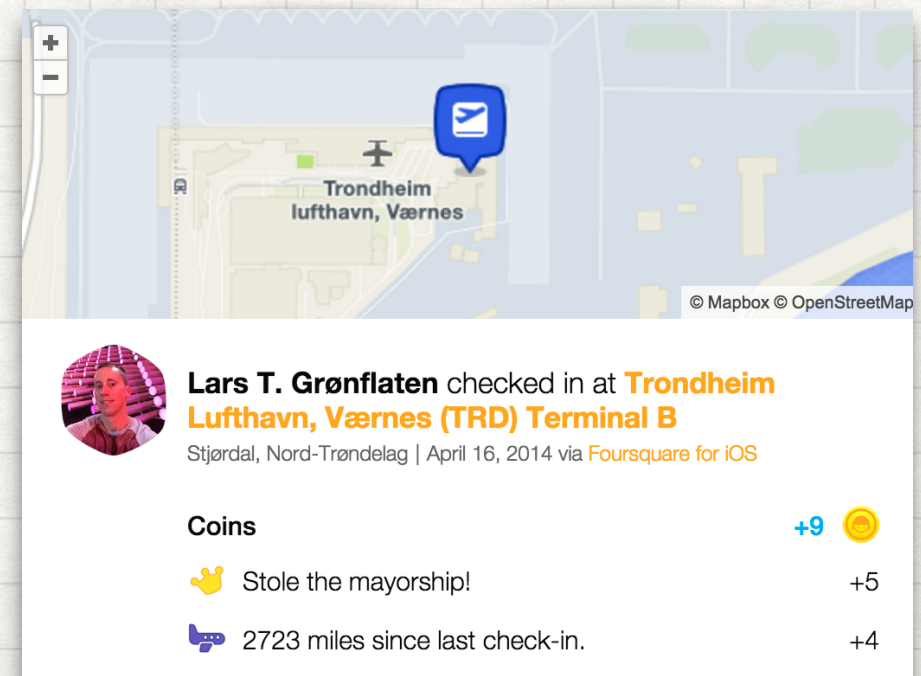
Sentiment analysis on Twitter



Task #1: Analysis of Foursquare dataset

Foursquare

- location-based social network
- recommendation of interesting places near you
- sharing of your actual location (check-in)
- 45M users
- data is private, but some can be retrieved via Twitter



Task #1: Analysis of Foursquare dataset


Foursquare data (19M records)

checkin_id	user_id	session_id	utc_time	timezone_offset	lat	lon	category	subcategory
25915736	1539	1539_AE_117	2013-02-06 17:46:30	-360	25.254118	55.303330	Food	Sushi Restaurant
22889371	1539	1539_AE_18	2012-11-11 12:29:30	-360	25.196387	55.280058	Outdoors & Recreation	Athletics & Sports
21089391	12021	12021_AE_1	2013-03-22 22:23:52	-180	25.219719	55.280236	Outdoors & Recreation	Scenic Lookout
19055908	12021	12021_AE_1	2013-03-22 22:25:40	-180	25.195058	55.278957	Food	French Restaurant
32211664	192383	192383_AE_21	2013-04-02 07:48:12	-180	25.161192	55.225976	Outdoors & Recreation	States & Municipalities

Brasilia	-15.792111	-47.897748	BR	Brazil	National and provincial capital
Goiania	-16.727004	-49.255001	BR	Brazil	Provincial capital
Campo Grande	-20.450997	-54.615996	BR	Brazil	Provincial capital
Puerto Presidente Stroessner	-25.526997	-54.622997	PY	Paraguay	Provincial capital
Talca	-35.423001	-71.659998	CL	Chile	Provincial capital

Task #1: Analysis of Foursquare dataset

Subtasks:

- Install *Spark* 
- Load the dataset.
- Calculate local time for each check-in.
- Assign a city and country to each check-in.
- Perform statistical analysis to answer the following questions:
 - How many unique users are in the dataset?
 - How many times did they check-in in total?
 - How many check-in sessions are there in the dataset?
 - How many countries are represented in the dataset?
 - How many cities are represented in the dataset?

Task #1: Analysis of Foursquare dataset

Subtasks:

- Calculate lengths of sessions (number of check-ins).
- For sessions with 4+ check-ins, calculate their distance (Km).
- Find 100 longest sessions that cover at least 50km and output them to a csv/tsv file.
- Create a free account on CartoDB.com, upload your output file and visualize the sessions.

Example output: <http://bit.ly/big-data-example>

Task #1: Analysis of Foursquare dataset

Delivery:

- Written report
 - Answers to the previously mentioned task.
 - Spark methods used to accomplish these tasks
 - Link to the visualization
 - Code of your solution
- Oral presentation
 - Short demonstration of your solution

Task #1: Analysis of Foursquare dataset

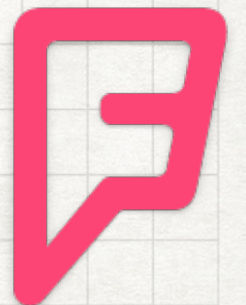
Tips:

- Sampling
- Experiment with Spark functions.
- Get some extra points for creative extensions!

MINI PROJECT

Task #1:

Analysis of Foursquare dataset



Task #2:

Sentiment analysis on Twitter



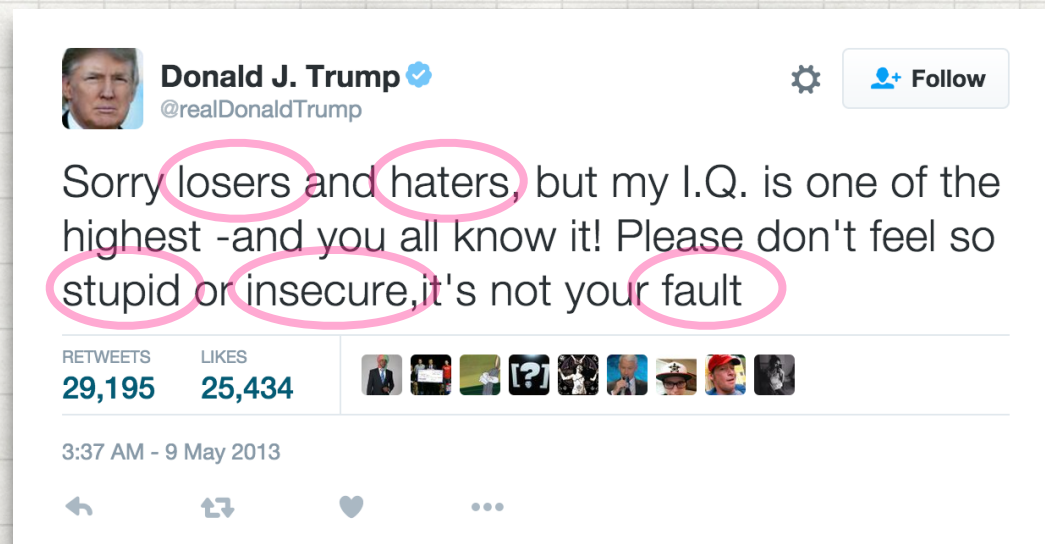
Task #2: Sentiment analysis on Twitter

Data

- localized tweets (toy dataset is around 11 M tweets)

Goal

- analyze polarity of tweets: positive / negative / neutral
- emphasis on performance



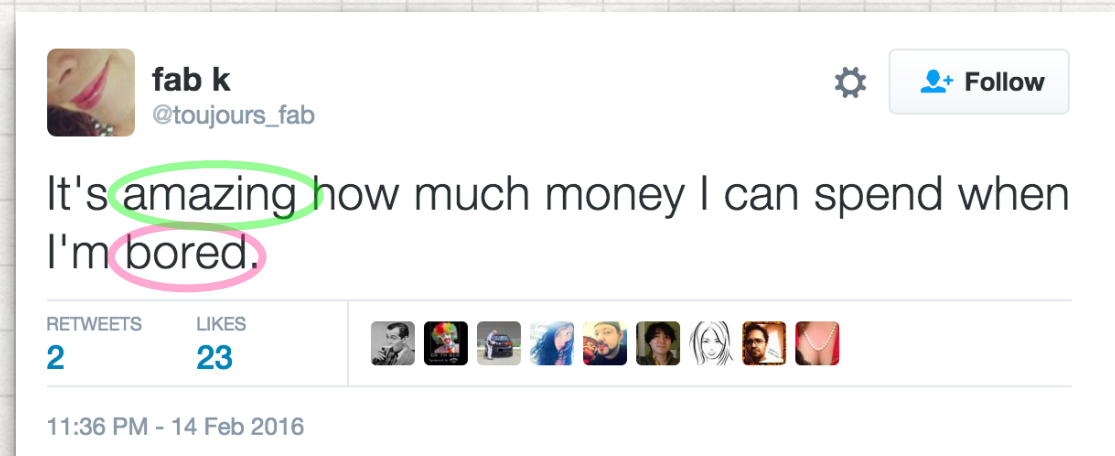
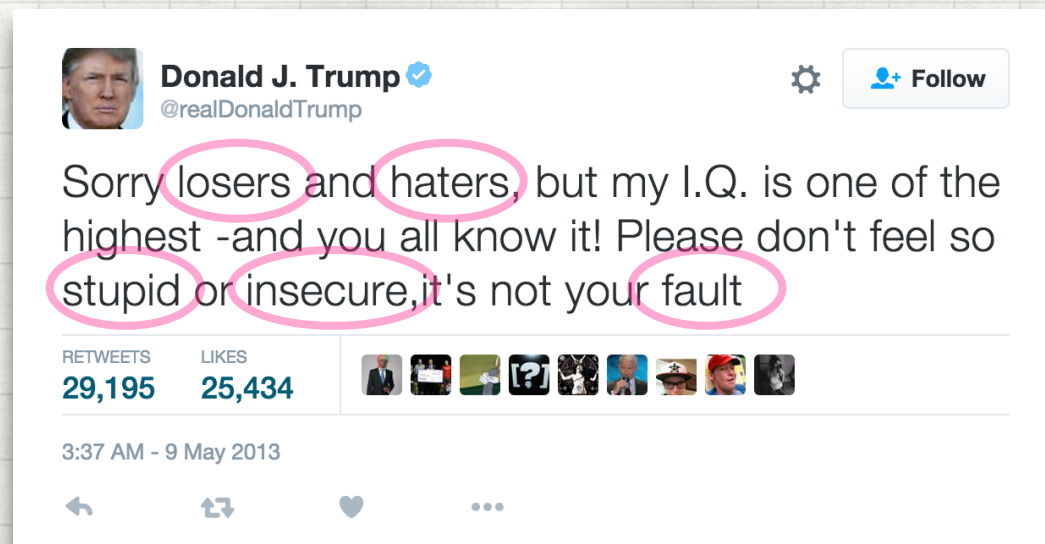
Task #2: Sentiment analysis on Twitter

Data

- localized tweets (toy dataset is around 11 M tweets)

Goal

- analyze polarity of tweets: positive / negative / neutral
- emphasis on performance



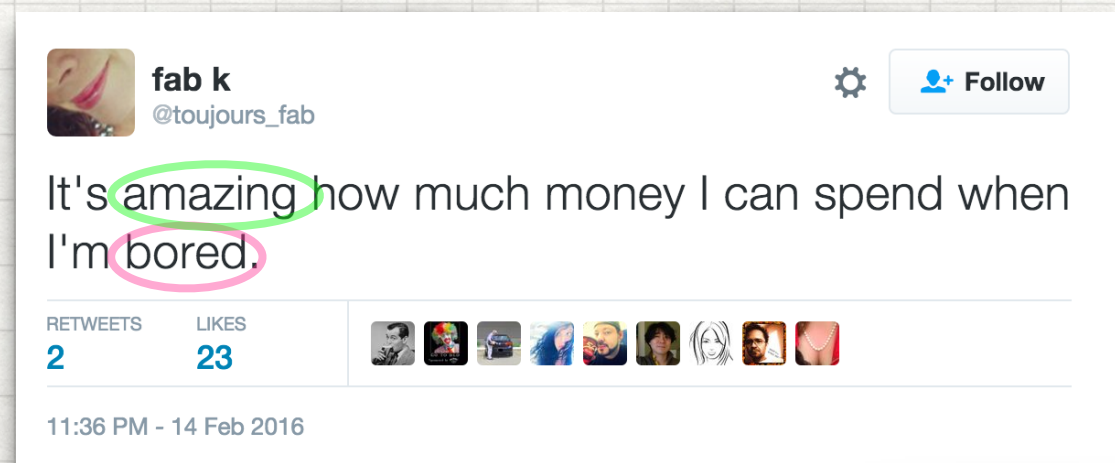
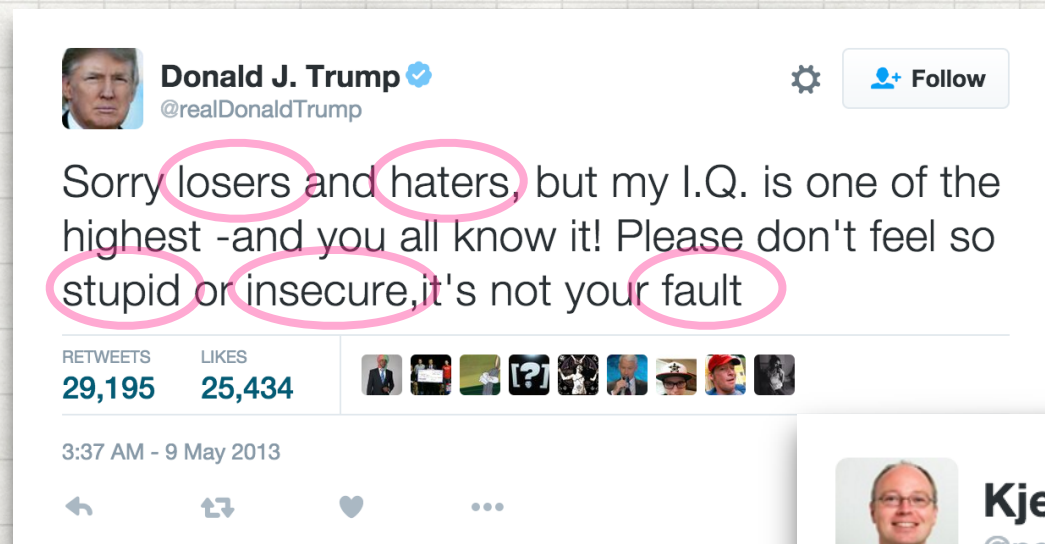
Task #2: Sentiment analysis on Twitter

Data

- localized tweets (toy dataset is around 11 M tweets)

Goal

- analyze polarity of tweets: positive / negative / neutral
- emphasis on performance



Task #2: Sentiment analysis on Twitter

How to find the polarity of a tweet?

- lexicon with positive and negative words
- tokenize each tweet into individual words and find all positive and negative expressions
 - More positive words => +1
 - More negative words => -1
 - Equal amount => 0
- Sentiment of a set of tweets is then sum of polarities of individual tweets.

Task #2: Sentiment analysis on Twitter

How to find the polarity of a tweet?

<tweet_1>	-1
<tweet_2>	1
<tweet_3>	-1
<tweet_4>	-1
<tweet_5>	0
- - - - -	
sentiment:	-2

Task #2: Sentiment analysis on Twitter

Subtasks:

- Load the dataset.
- Calculate local time for each check-in.
- Load sentiment lexicons (list of positive/negative words).
- Find polarity of each tweet.
- Find aggregated sentiment for:
 - tweets in English (`lang = 'en'`)
 - for all cities in the US (`place_type = 'city', country = 'us'`)
 - for each day of week

Example: <http://bit.ly/big-data-sentiment>

Task #2: Sentiment analysis on Twitter

Output:

Format:

```
city<tab>day<tab>sentiment
```

Example:

New York	Monday	-34409
New York	Tuesday	14777
New York	Wednesday	1777
...		
Chicago	Monday	4409
...		

Task #2: Sentiment analysis on Twitter

Delivery:

- Code
 - Will be tested on a larger dataset with the same format.
 - You can participate in a competition for the fastest solution!

Tips:

- Sampling, again.
- Normalize and lowercase the text of tweets.

```
u'aあä'.encode('ascii', 'ignore')
```


PRACTICAL ISSUES

Deadline

- Weeks 15, 16
- Doodle will be set up for your presentations.
- Solution for task #2 will be evaluated even before deadline

Groups

- 2 persons

Alternative assignment:

- Seminar: presentation of a paper