Presentation - Data Mining

# Tillämpningsområden för tweets

Prediktera presidentval.

Hur presidentval reflekteras i tweets. Hur det folk skriver kan kopplas till resultat.

Upptäcka jordbävningar snabbt.

Vad sjukdomsupplysning kan behöva fokusera på vid sjukdomsutbrott.

Prediktera intäkter för en kommande film .

# Algoritmer som används

TF-IDF transformtions

Sentiment analysis

Word clouds

NB, KNN, Random forest, SVM (2)

SVM, Kalman filters (4)

K-means (5)

## Slide 1

Fix heading

## Slide 2 - What is Twitter?

General info about twitter

320 million monthly active users. 1 billion unique visits monthly to sites with embedded Tweets. 80 % active users on mobile. 3,900 employees around the world

35+ offices around the world. 79 % accounts outside the U.S. 35+ languages supported. 40 % employees in technical roles.

*All numbers approximate as of December 31, 2015.*

Source: https://about.twitter.com/company , 27/4 -16

Launched in July 2006.

https://en.wikipedia.org/wiki/Twitter, 27/4- 16

Want to know numbers of tweets per day, following link is not an official statement

http://uk.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2, 27/4- 16

1.3 billion registered users of Twitter

http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/, 27/4- 16

## Slide 3 -5 - Access Twitter data

Two public APIs that can be used for gathering twitter data, the REST API and the streaming API.

The REST API

The [REST APIs](https://dev.twitter.com/rest/public) provide programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more.

15 minute windows where how much data you send requests for is limited to a certain number of requests, depending on what you are requesting. For example are you allowed to do 180 or 450 twitter search requests depending on the form of authorization.(Think the normal case is app authorization, 450 twitter search requests per 15 minutes.

If these limitations are abused , the account will get blacklisted. Meaning that you will be unable to get a response from the Twitter API.

The Twitter Search API is part of Twitter’s REST API. It allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as [Twitter.com search](https://twitter.com/search). The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days.

Before getting involved, it’s important to know that the Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results. If you want to match for completeness you should consider using a [Streaming API](https://dev.twitter.com/streaming/overview) instead.

GET statuses/user\_timeline - Returns a collection of the most recent [Tweets](https://dev.twitter.com/overview/api/tweets) posted by the [user](https://dev.twitter.com/overview/api/users) indicated by the screen\_name or user\_id parameters.

The timeline returned is the equivalent of the one seen when you view a user’s profile on [twitter.com](http://twitter.com/).

This method can only return up to 3,200 of a user’s most recent Tweets. Native retweets of other statuses by the user is included in this total, regardless of whether include\_rts is set to false when requesting this resource.

https://dev.twitter.com/rest/public, 28/4 - 16

Streaming API

The Streaming APIs give developers low latency access to Twitter’s global stream of Tweet data.

If your intention is to conduct singular searches, read user profile information, or post Tweets, consider using the [REST APIs](https://dev.twitter.com/rest/public) instead.

If you want to match for completeness you should consider using a [Streaming API](https://dev.twitter.com/streaming/overview) instead.

Public streams: Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining.

User streams: Single-user streams, containing roughly all of the data corresponding with a single user’s view of Twitter.

Limitations: No public limitations, but to excessive connection requests may lead to some limitations of the possibility to connect to the API.

Search parameters:

The language of the tweets( will only return Tweets that have been detected as being written in the specified languages.).

User, or users, that you chose to follow. Will include tweets, retweets and replies from the chosen accounts.

Track. What words, hashtags, sentences or urls that you want to track.

Locations. A comma-separated list of longitude,latitude pairs specifying a set of bounding boxes to filter Tweets by. Only geolocated Tweets falling within the requested bounding boxes will be included—unlike the Search API, the user’s location field is not used to filter tweets.

https://dev.twitter.com/streaming/overview, 28/4-16

https://dev.twitter.com/streaming/overview/request-parameters, 28/4-16

## Slide 6 - The use of tweets

Just name the examples on the slides

## Slide 7 - US presidential elections in tweets

Analyze the public opinion on the main candidates, Barack Obama and Mitt Romney, the weeks before the presidential elections in 2012.

We extracted all the tweets that contained the candidate names, candidate addresses, election related hash tags and candidates hash tags (by crawling the Twitter using *Twitter API*). We collected the full text, the author, the written date and time of the tweets, and we ended up with 196,000 tweets that we stored in a database.

One day ratio of positive versus negative tweets about the candidate.

Only positive ratios for Obama, big peak at the of the elections. Romney mostly had negative ratios.

Word cloud analysis for analyzing which the most frequent terms are about the election. Words that have a frequency greater than 100. Discards words that not say anything, for example "and", "at", "I" and so on. In this case probably also the names of the candidates have been discarded as it not would be so interesting to see a word cloud with their names. It is more interesting to see what is written about the election apart from the most obvious things.

## Slide 8 - Next example

## Slide 9 - Final example

## Slide 10 - Data pre-processing

## Slide 11 - Methods for analyzing tweets