# Basic Information

|  |  |
| --- | --- |
| Project name | TBA |
| Abbreviation | *TBA* |
| Supervisor | *Doc. RNDr. Irena Holubová PhD.* |
| Consultants | *jméno <email>* |
| Annotation | *Extensible framework for analysis of data from social networks. It provides users with functionality to acquire large amount of data according to users’ credentials and analyse them further. Extensibility is demonstrated twofold. User is given opportunity to implement his own data acquisition and analysis modules.* |

## Motivation

For more than a decade already, there has been an enormous growth of social networks and their audiences. As people post about their life and experiences, comment on other people’s posts and discuss all sorts of topics, they generate a tremendous amount of data that are stored on these networks. It is impossible for users to get a concise overview of the general sentiment about a given topic.

This work focuses on a framework allowing users to connect to social network, download and analyze data related to some topic.

# Project description

The goal of this project is to create a framework for analysing public social networks content and computing sentiment over the data published on supported social network under given conditions (time period, topic…). The result of the analysis will be presented to the user in a structured format consisting of

* sentiment analysis
  + of a specific topic
  + in a given time frame
* significant keywords
* clustered opinions (positive, neutral, negative)

Main challenge of this project is to tackle limits of data acquisition – data are not available in the full extent. The data acquisition is limited (for a free user) by both, requests per time period (e.g. currently 15 minutes for Twitter, 1 hour for Facebook) and total amount of records retrieved. These limits make it impossible to support extensive online analysis.

A solution to this problem is to offer a different mode of data analysis called watchdog. In this mode, data are analysed as they are acquired regardless of its extent allowing user to have up to date analysis of the most recent data (especially useful for a long time user interested in his organization/business profile)

## Sentiment analysis

Three main tasks need to be solved by this module - topic modeling, features extraction and classification. Output of the topic modeling will be set of tags for every text (post/comment). Some of the well-known methods as Latent Dirichlet Allocation, Latent Semantic Analysis or lda2Vec (i.e. in the case word embeddings will be use even for other parts) will be used. Features extraction consists of different parts of language preprocessing including tokenization, lemmatization etc. Sentiment analysis will be than performed on this new features by some machine learning approach. Generally there are two possibilities - deep neural networks (probably pre-trained) and classical approach: support vector machines or logistic regression.

## Use cases

User interaction with this software is straight forward. User define task by selecting

* social site
* topic
* time in the past where the analysis should begin
* credentials

In further text, these data are referred to as *task definition*.

The task definition is then translated to a command chain beginning from data acquisition, transformation and storage, to analysis and selected form of presentation. Since the data acquisition takes a very long time, the user is given special link where a progress and temporary analysis results can be viewed.

***An example: market sentiment analysis***

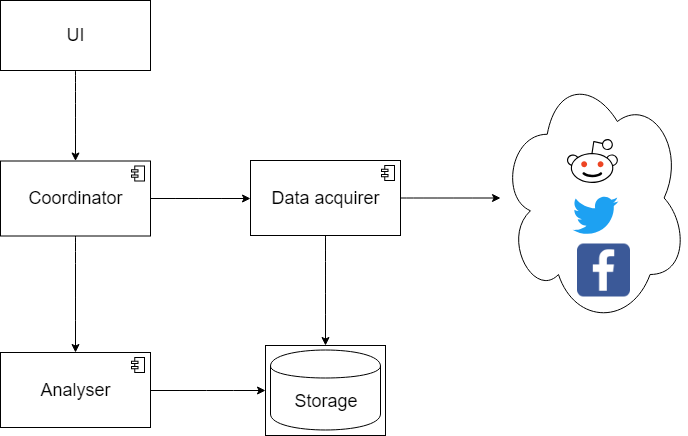
When referring to stock or financial market, we can analyse and model sentiment about a specific company stocks, or a currency. That way we can research behaviour of traders in reaction to certain stimuli. We can compare patterns in historical sentiment model with the graphs of actual market of stocks, as well as with events and news in the outside of the market. If we would be able to find correlation between these, we could possibly predict the development of market in future.

This can be especially useful with markets with high levels of volatility and manipulability (and therefore practically very manipulated). A first-hand example of such fast-paced and highly manipulated market are the cryptocurrencies, and their trading platforms.

# Platform, technology

This framework is a web application consisting of following modules

* Data acquisition - a module acquiring data from selected data sources.
* Analyser - a module analysing stored data according to selected analysis type.
* Storage - a wrapper encapsulation existing database solution
* Coordinator - a web application coordinating work of all previous modules from job definition to data acquisition and transformation to analysis to visualisation.
* UI – Frontend of the application



*Figure 1. Framework component model*

Further analysis will be performed for proper technology selection.

## Extensibility

The application is designed to obtain data from social network in small batches and then store them. The stored data can be then loaded and analysed. Nevertheless, additional functionality can be added using user-made modules.

For example, the *Data acquisition* module can be extended by adapters to various data sources: files, other social networks, news media… The only requirement is that data must follow given schema.

The *Analyser* module can be extended by a custom made analysis. This alternation must be accompanied with appropriate visualisation tools.

# Time estimation

## Team

|  |  |
| --- | --- |
| Name | Responsibilities |
| Irena Holubová | Supervisor, decision leader |
| Jan Pavlovsky | Machine learning engineer, software engineer – builds the platform with focus to machine learning integration |
| Petra Doubravová | Machine learning, linguistic specialist – develops sentiment analysis model |
| Jaroslav Knotek | Software engineer – builds the platform |
| Lukáš Kolek | Data engineer – design and implements data storage |

## Plan

|  |  |  |
| --- | --- | --- |
| Phase | Months estimated | description |
| Analysis | 1-2 | In the beginning, we need to state more specifically what should be the output of our work, assess our ability to carry it out, and carefully specify what if achievable in the time frame given. We also need to study existing materials about the topic in this phase. |
| Design | 1-2 | Following the specification, we as a team should make fundamental design and architectural decisions about the software. Based on that, we can divide work among team members. |
| Implementation | 3 | Individual team members will implement assigned modules of the system. We would like to adapt an iterative development approach, in which we would iteratively integrate and test individual system modules. |
| Testing | 2 | The system as a whole will be extensively test to ensure all functionality if working. Also there is space for various experiments with possible  software extensions. |
| Finalization | 1 | In the terminal phase, the framework will be prepared to be presented which requires acquisition of data reflecting various use-cases, ensuring connectivity. The team will also need to prepare for a defense of this project. |

# Project definition

|  |  |
| --- | --- |
| Diskrétní modely a algoritmy | |
|  | diskrétní matematika a algoritmy |
|  | geometrie a matematické struktury v informatice |
|  | optimalizace |
| Teoretická informatika | |
|  | Teoretická informatika |
| Softwarové a datové inženýrství | |
| x | softwarové inženýrství |
| x | vývoj software |
|  | webové inženýrství |
|  | databázové systémy |
| x | analýza a zpracování rozsáhlých dat |
| Softwarové systémy | |
|  | systémové programování |
|  | spolehlivé systémy |
|  | výkonné systémy |
| Matematická lingvistika | |
|  | počítačová a formální lingvistika |
|  | statistické metody a strojové učení v počítačové lingvistice |
| Umělá inteligence | |
|  | inteligentní agenti |
|  | strojové učení |
|  | robotika |
| Počítačová grafika a vývoj počítačových her | |
|  | počítačová grafika |
|  | vývoj počítačových her |

# Poznámky

*No social networks are promised yet*