Basic Information

Project name	Scent of Current Network Overview
Abbreviation	***SOCNETO***
Supervisor	*Doc. RNDr. Irena Holubová PhD.*
Consultants	*Mgr. Kateřina Veselovská, Ph.D.*
Annotation	*The aim of the project is to develop a framework for analysis of social network data. It provides users with functionality to analyse sentiment of large amounts of data and visualize the results. Data are acquired via social networks public APIs or custom 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 in these networks. It is virtually impossible for a user to get a concise overview about any given topic.

This project focuses on development of a framework allowing the users to download and analyze data related to a chosen topic from given social networks.

Use cases

The application serves as a tool for analysing vast social networks' content. In a typical use case, a user specifies a topic for analysis, selects data sources to be searched and submits the job. The system then runs all necessary tasks and when finished, supplies the user with results in the form of sentiment chart, significant keywords or post examples with the option to explore and search through them.

Sample use cases might be studying sentiment about a public topic (traffic, medicine etc.) after an important press conference, tracking the opinion evolution about a new product on the market, or comparing stock market values and the general public sentiment peaks of a company of interest.

This project does not serve to any specific user group, it tackles the problem of creating concise overview and designing a multi-purpose platform. It aims to give the user an ability to get a glimpse of prevailing public opinion concerning a given topic in a user-friendly form.

Challenges

The primary challenges of this project are data acquisition, sentiment analysis, data storage and transformation, user-friendliness and many problems corresponding to developing an extensible Big Data platform.

Data acquisition and limits

Data acquisition needs to cover variety of sources with various strict limits. These limits make it impossible to support an extensive online analysis. We will tackle this issue by implementing continuous analysis which will work as follows: In an, e.g., hour-long time, the platform will:

1. Collect as many post as limits allow for
2. Analyse the posts
3. Recalculate aggregated analyses
4. Store the results and allow users to use them

Large variety of sources will be covered by enabling users to add custom adapters implementing designed interface which requires a user to transform all data from any data source to a given unified structure.

Sentiment analysis

Three main tasks need to be solved by this module – topic modelling, feature extraction and classification. The output of the topic modeling will be a 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 used even for other parts) will be used. Feature extraction consists of various methods for language preprocessing including tokenization and lemmatization. Sentiment analysis will be then performed on these features with an appropriate machine learning approach. There are two possibilities we will work with – deep neural networks (probably pre-trained and fine-tuned on our data) or the classical approach utilising support vector machines and logistic regression. We assume that (possibly slightly modified) third-party libraries will be used to cover all the hard-lifting.

Data storage

Our vision is to develop a custom scalable storage for both downloaded and analysed data. Our solution will integrate existing tools. First, we will perform a research analysis in order to select which tools fit our storage needs the best. The required parts of the storage are:

- Scalable storage for posts – a NoSQL database

- Storage for internal data (users, jobs etc.) – a relational database
- Searching through posts – Elasticsearch

The storage may be extended with a graph database, which will help to store relations between posts or the users and make analysing of these relations easier.

User Interaction and visualisation

The user will interact with the system using a dedicated web application. In order to submit a job, manage it or list its result, the user will need to register and log in.

In order to submit a job, the user needs to select:

- topic(s) of interest
- social networks(s)
- own social-network credentials (optional)
- analysed time frame (optional)

When the task yields first results, the user is presented with aggregate data such as histogram of all posts in a given time period with visualised negative / positive sentiment ratio. The user can also see samples of posts along with their sentiment in order to check what exactly was written.

Since some tasks may take a non trivial time (e.g. because of long data acquisition), the user will also see progress of his not yet finished tasks.

The software will be built using existing framework DartAngular with Material Design.

Platform, technology

The framework infrastructure must support a flow of large amounts of data which are subsequently stored, analyzed and shown to the user. It will consist of the following modules:

- Data acquisition – a module acquiring data from social networks. Each social network is represented by its own adapter. Custom adapters can be added in future.
- Analyser – analyses sentiment of posts and comments
- Storage – stores social network data about posts, analysis and metadata in various types of databases (polyglot database)
- Coordinator – responsible for orchestrating cooperation among modules resulting in continuous data flow
- UI – frontend of the application

The referred modules will be separate web applications. That will allow for better separation of concerns. They will run on the university public infrastructure.

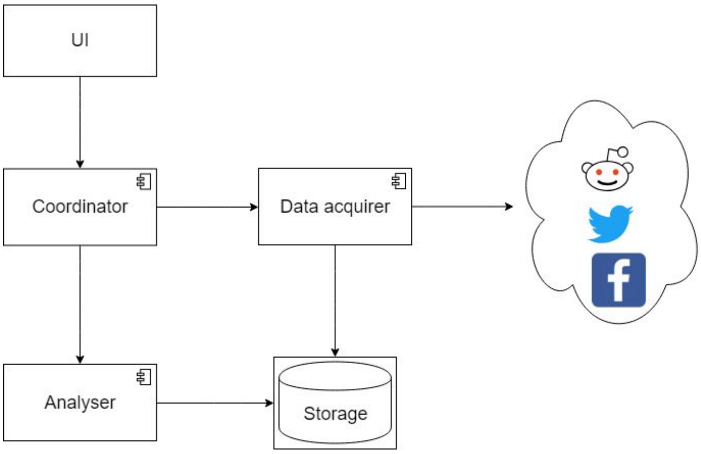

Figure 1. Framework component model

What we will implement and what integrate

The list of modules along with enumeration of their features which we will implement or use a third-party software is as follows:

- Data acquisition – The module itself will be created by us from scratch and allow a user to extend it by other acquisition modules respecting the given interface. We will support Twitter and Reddit out-of-the-box via third-party libraries.
- Analyser – We will utilize existing methods of sentiment analysis and train them by ourselves if necessary.
- Storage – This module will integrate multiple existing databases in order to fulfill our polyglot data requirements. Schema and data flow will be custom.
- Coordinator – Our currently most considered way to implement this is a message-broker with a publish-subscribe model. It will impact the whole system and might require deep analysis. An alternative is to use simple streaming channels between nodes.
- UI – We will use third-party framework DartAngular with Material Design

Team

Name	Responsibilities
Irena Holubová	Supervisor, decision leader
Jan Pavlovský	Machine learning engineer, software engineer – builds the platform with a focus on machine learning integration
Petra Doubravová	Machine learning, linguistic specialist – develops the sentiment analysis model
Jaroslav Knotek	Software engineer – builds the platform
Lukáš Kolek	Data engineer – designs and develops the data storage
Julius Flimmel	Web engineer – builds the web application and frontend

The Plan

Our plan has three primary milestones, that are related to tight deadlines of the project.

1. Proof of Concept
2. End2End solution
3. Full features implementation

The first milestone is to build a proof of concept that corresponds to a full-specification document which will be created along with the software. Key features will be analysed in order to accurately specify their impact. This phase takes 2 months exactly since it is the time when the full specification should be submitted.

The second milestone represents the time when we will have a working example without full documentation and without proper testing, although a draft documentation will be created during the whole development. This phase will take most of the time, approximately 6 months.

The last milestone is the first deadline of the software project submission. At this time we will have the software ready to be submitted and presented.

This project will follow more agile form of development since the project relies heavily on third-party components that we might find incompatible in the process of development.

The application will be developed incrementally following lean and agile methodologies, where each increment will encompass implementation, reasonable testing and also partial documentation.

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
x	webové inženýrství
x	databázové systémy
x	analýza a zpracování rozsáhlých dat
Softwarové systémy	
	systémové programování
	spolehlivé systémy
	výkonné systémy
Matematická lingvistika	
	počítačová a formální lingvistika
x	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