Incorporating social media for personalization of ISSUU content

Dimitar Lesov
s101557

Kogens Lyngby 2013

IMM-M.Sc.-2013-31

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
IMM-M.Sc.-2012-31

Abstract

ISSUU[i] is a leading digital publishing platform delivering exceptional reading experiences of magazines, catalogs and newspapers. It is the fastest growing digital publishing platform in the world with around 60 million monthly readers and over 10 million publications available, but also a very popular destination site where people are engaging with the web's best publications and where publishers build their audience.

In order to progress and improve the relations with its users and customers ISSUU started to incorporate integrations with social media sites like Facebook and Twitter. Due to the early stages of the incorporation and focusing on the different objectives ISSUU hasn't used the full potential of the Social Media integration and hasn't benefited from that in order to improve its services.

The importance of the social networking today is outstanding, and most of the companies have already invested a lot of time and money into their social efforts. It is undeniable that the society had shifted into a new era of social. Social Media now takes up the majority of people's media consumption. So, not only do social media allow companies to reach consumers more easily and effectively, it also may be the qualifying factor as to whether or not any business gets done.

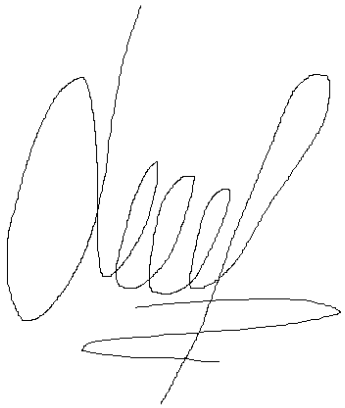
Preface

This thesis was prepared at the Department of Applied Mathematics and Computer Science, at the Technical University of Denmark and with the mutual cooperation of the private company ISSUU in fulfillment of the requirements for acquiring the MSc degree of Digital Media Engineering.

This thesis deals with Social Media data analysis. The main focus of this project is to design and create tools for Social Media data analysis suited for ISSUU, utilizing Social Media content and personalization.

30 ECTS credits worth project was started on the 1st of October 2012 and finished on the 29th of March 2013. Supervisor of the project is Michael Kai Petersen.

Lyngby, 01-October-2012



Dimitar Lesov

Table of Contents

List of Figures

List of Tables

List of abbreviations

LDA	Latent Dirichlet allocation
FQL	Facebook Query Language
API	Application Programming Interface
CSS	Cascading Style Sheets
HTML	Hyper Text Markup Language

Glossary

Axure	Axure is a wireframing, rapid prototyping, and specification software tool aimed at web and desktop applications. It offers capabilities typically found in diagramming tools like drag and drop placement, resizing, and formatting of widgets. In addition, it has features for annotating widgets and defining interactions such as linking, conditional linking, simulating tab controls, show/hide element etc.
Connected user	A ISSUU user that has connected his
Collaborative filtering	Collaborative filtering is a method of making automatic predictions about the interests of a user by collecting preferences or taste information from many users.
Content Network	A network generated by the End user's pre-selected and cleaned Social Data.
End user	A person who is using ISSUU services
Heroku	Heroku is leading open language cloud application platform that supports Ruby, Java, Python, Clojure, Scala and Node.js.
ISSUU User	The company ISSUU viewed as a user of the developed by the Master Thesis Project product.
ISSUU Data	Any data available for the End user in the ISSUU platform, including the End user publication readings.
LDA	LDA is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
Network Graph	A visual way of representing network data and can often provide insights into

	relationships which are hard to identify in tabular data.
NetworkX	NetworkX is a Python library for studying graphs and networks.
Promoted Publication	An ISSUU publication that is promoted to the End users via the recommendation system. Every time an End user reads Promoted Publications ISSUU receives revenue from the Publisher User.
Product Owner	The person responsible for maintaining the Product Backlog by representing the interests of the stakeholders, and ensuring the value of the work the Developer does.
Product Backlog	A prioritized list of high-level requirements.
Publisher User	An ISSUU user that publishes documents on the ISSUU platform.
Python	Python is a general-purpose, high-level programming language whose design philosophy emphasizes code readability.
SCRUM	Scrum is an iterative and incremental agile software development framework for managing software projects and product or application development.
Sigmajs	Sigmajs is an open-source lightweight JavaScript library to draw graphs, using the HTML canvas element. It has been especially designed to display interactively static graphs and display dynamically graphs that are generated on the fly.
Social Web	The Social Web is currently used to describe how people socialize or interact with each other throughout the World Wide Web.
Social Data	Any data that the End user has provided through the Social Media.
Sprint Backlog	A prioritized list of tasks to be completed during the sprint.
Sprint	A time period (typically 1–4 weeks) in which development occurs on a set of backlog items that the team has committed to. Also commonly referred to as a Time-box or iteration.
(User) Story	A feature that is added to the backlog is commonly referred to as a story and has a specific suggested structure. The structure of a story is: "As a <user type> I want to <do some action> so that <desired result>" This is done so that the development team can identify the user, action and required result in a request and is a simple way of writing requests that anyone can understand.
User Reading Data	The reading patterns of the ISSUU End user.
Velocity	The total effort a team is capable of in a sprint. The number is derived by adding all the story points from the last sprint's stories/features. This is a guideline for the team and assists them in understanding how many stories they can do in a sprint.

1. Introduction

The adoption of Social Media is expanding at an exceedingly rapid rate and revolutionizing the way we interact, communicate and do business. Despite being at an early stage regarding business use, the concept of Social Media has managed to attract attention. Originally intended to enable internet users to communicate and publish personalized content, there is business potential that these platforms have. The richness of personal data variety supported by Social Media in conjunction with the large user base allows them to be used as a unique tool for gathering personalized information about the customers, as well as attracting new customers.

1.1. Motivation

This section is about the choice of Master Thesis subject and the motivation behind it.

Upon finishing his Master programme the author was intending to create his Master Thesis project in the field of the Social Media Data Analysis. In the recent years that field has become an area of interest for the scientific society with its potential to improve understanding of people's interactions, current interests and their change over time.

The development of the project's idea and its work process was inspired by the ISSUU Company, which is a leading digital publishing platform that provides exceptional reading experiences of magazines, catalogs and newspapers.

ISSUU has been looking for a way to incorporate better its Social Media integrations in order to improve its personalization services and bring more value to its customers.

1.2. Identifying the problem

In this section the author is going to make an overview of what is the current status of ISSUU services in order to clearly formulate the addressed problem.

In the core of ISSUU services there are three main activities (refer to Figure 1) that are responsible for the End user experience and providing personalization to the customers.

In order to understand the problem at hand one has to understand the main activities and how they could be improved with incorporating social media for better personalization of the ISSUU content.

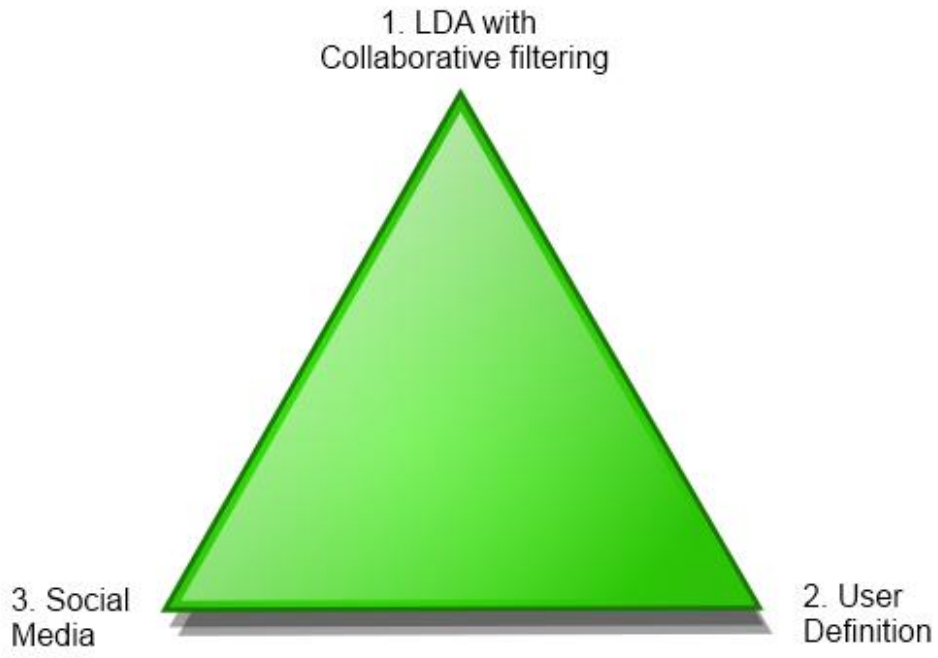


Figure 1: ISSUU Service Components

1.2.1. LDA with Collaborative filtering

To analyze the content of the published magazines in over 180 languages and to recommend similar reads to the customer, ISSUU is using LDA[ii] with Collaborative filtering[iii]. That allows ISSUU to make explanations on sets of observations by analyzing unobserved groups explaining why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. In addition to that ISSUU apply techniques involving collaborative filtering among multiple agents, viewpoints, data sources, etc. to filter the information and reveal patterns and make automatic predictions about the interests of an End user by collecting preferences or taste information from many End users (collaborating).

Providing recommendations for Promoted Publications is one of the main revenue flows of the company. The Publisher User creates Promoted Publications and ISSUU gets revenue every time an End user reads the Promoted Publication. That's why it is in the ISSUU best interest to recommend publications to the End user that are in his field of interests.

This is one of their key service components and they have strongly invested in the developing of that functionality and there is very little that the author could do to improve that component,

therefore the focus of the Master Thesis Project will be improving the other two service components and the connection between them.

1.2.2. User Definition

In the ISSUU platform, a user is defined when he creates an account. The user has two options:

- To create an account by typing in name (first and last), email, gender and age. After the user creates his account ISSUU starts analyzing what he reads and make recommendations based on that. The shortcoming of that is that there is no social media information and integration for the user and the profile information is limited to four items, which most of the time are not correctly filled in.
- To create an account by Facebook Login[iv]. By doing that the user's Facebook account is connected with user's ISSUU account. Currently that integration is not used for any other benefits than fast account creation and log-in and that presents a field of focus and improvement for the author, since the Facebook integration could be used to improve the user's definition in various ways and the user information available from the social media is more reliable.

When defying a user, ISSUU uses its LDA with Collaborative filtering mechanisms to recommend publications for that user. The user definition is not used in the recommendation mechanisms because the social media integrations haven't been used in its fullest potential. Incorporating the reliable social media user information into the user definition and potentially using it for the recommendation system is one of the Master Thesis goals.

1.2.3. Social Media

ISSUU has recently understood the importance of social media and started devising social media integrations. Currently available as social media integrations are the following:

- Facebook Connect and Facebook Login which provides an easy account creation and fast log-in.

- The End user has the option to like a publication by Like Button[v], which posts the publication on his Facebook Timeline[vi].

- The option to share the publication via Facebook, which posts the publication on his Facebook Timeline.

- The option to share the publication by Twitter, which tweets the publication in the user's Tweeter profile.

- The option to share the publication by Google+, which posts the publication in the user's Google+ profile.

Those integrations are a good start, but they only scratch the surface of what Social Media could be used for. Improving the Social Media aspects of ISSUU service components is one of the primary goals of this Master Thesis Project.

1.3. Thesis Structure

The report is divided into several chapters presenting the process of making system for Social Data gathering, analysis and visualization.

The motivation of the thesis is presented at the beginning. It describes why the author is interested in the thesis subject. A clear problem formulation is presented – a basis for the thesis goals.

The next chapter is about the used methods and related work introducing the topic.

Chapter three is focused on the user needs. It consists of prioritized high level UML use cases which are used as basis for requirements.

In Chapter four the Design of the prototype application is presented. It describes user stories, improvement of User Definition and methods for data collection, analysis and visualization.

Implementation process is described in Chapter five. Firstly, the social data gathering and analyzing tool running on a web cloud hosting service had been implemented to test early results of the analysis. Several kinds of tests were performed including validity, case specific, etc.

Finally, the Discussion chapter focuses on what has been attained with the project and relates to the user needs and the initial thesis goals.

2. Requirements

In this section a requirements analysis is performed identifying the user needs and defining the thesis goals.

2.1. User Needs

The user needs are the foundation of creating the project requirements.

A mutual collaboration with ISSUU was in place during the initial phases of the project.

By looking through what is already available and what could be done in order to improve the provided services the author was able to identify the User needs

- Provide more Value to its customers through improved Personalization.

This is a key user need – the availability of a rich data defining the user, provided by the social media, is the basis on which the data analysis is going to relay. The variety in parameters when performing collaborative filtering is going to ensure better results in discovering trends in the End user reading patterns.

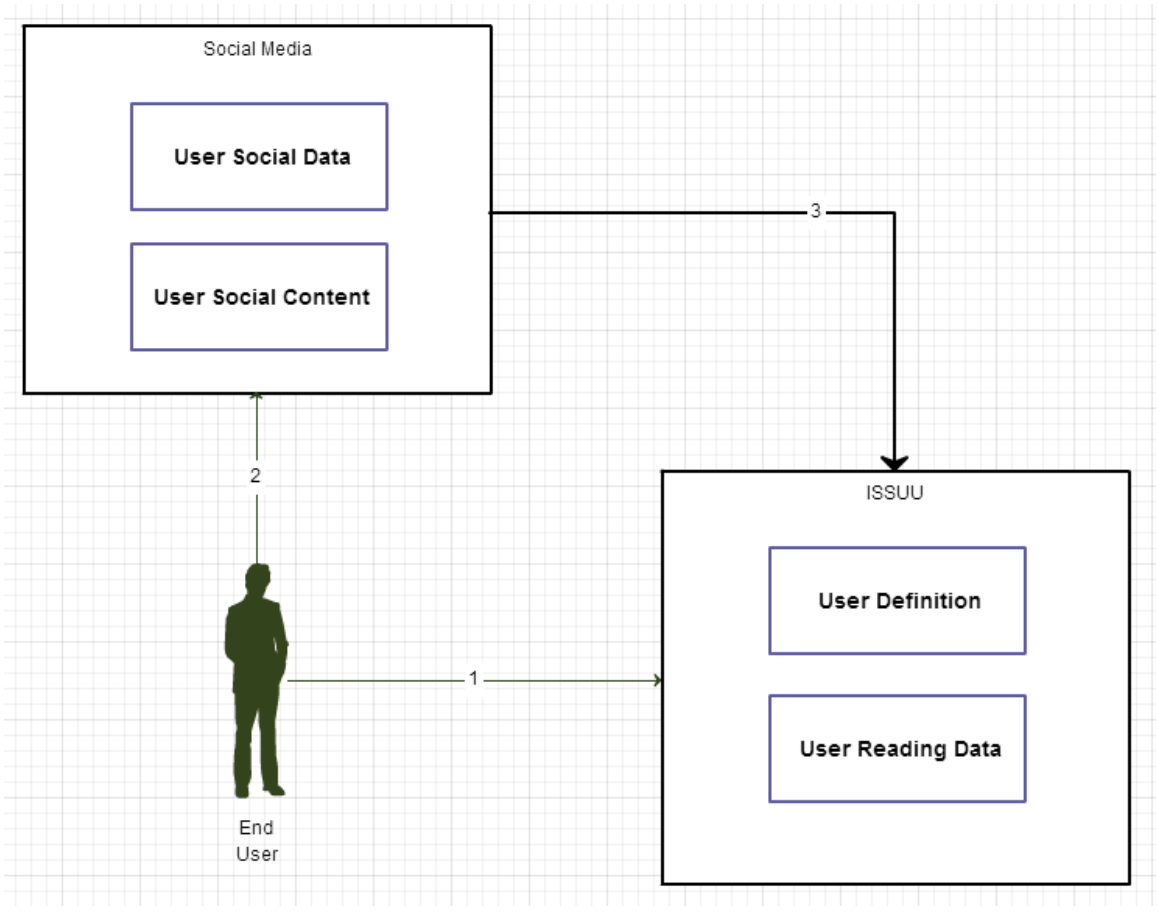


Figure 2: Import Personal Data

1 – End user creates ISSUU account
2 – End user has Social Media account
3 – Import personal data from the Social Media to ISSUU in order to improve personalization services.

By gaining more knowledge of the End user preferences it is possible to improve the provided recommendation service and that is going to increase the user satisfaction and the revenue flow.

As an End user it would be great to have a tool to make the personalization without any additional steps, silently, without spending time providing information about yourself, information which is already available on the social web.

- Perform Data Analysis over its Users' Social Data in order to improve ISSUU services.

In order to "translate" the Social Media Data into the ISSUU definition of user data, that data needs to be analyzed with the relevant methods. That ensures the possibility of comparison and/or consolidation of the two datasets. With that the author ensures that a wrong conclusion, based on the differently analyzed datasets, is not drawn.

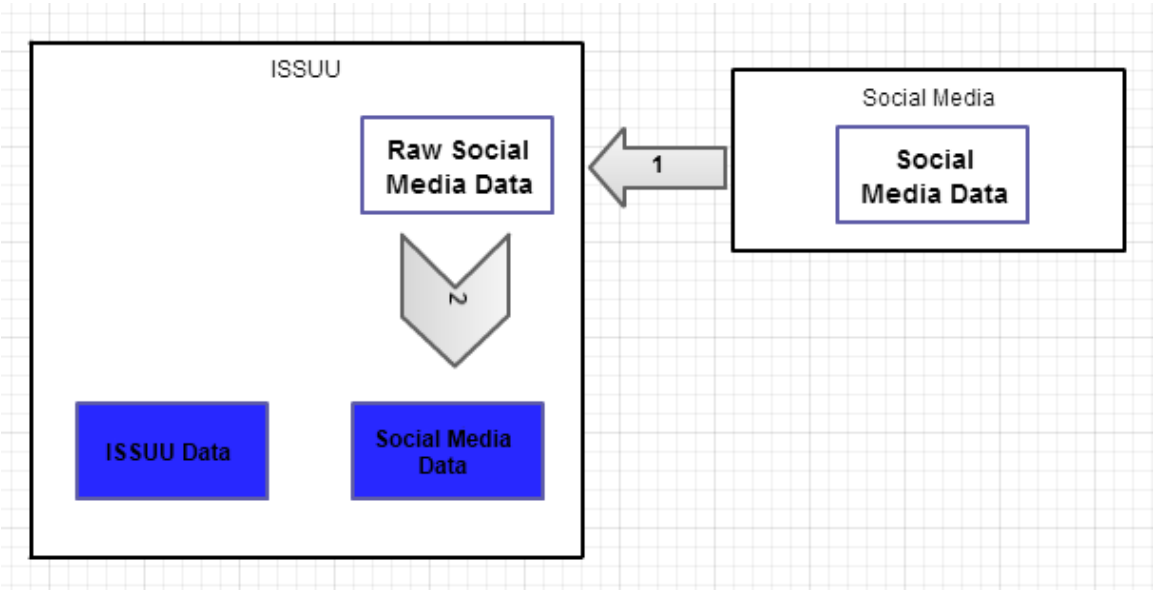


Figure 3: Analyze the Data

1 – Acquire the Social Media Data and import it to the ISSUU system.
2 – Analyze the raw Social Media Data and convert it to the same format as the ISSUU Data.

- Provide tools for comparison/consolidation between End user's Social Data and ISSUU Data to improve recommendation system and increase Value for the customers.

ISSUU views its End users as "You are what you read". This is a very good approach for a digital publishing platform, but it lacks variety in the data types and has only one angle from which it views its customers. To ensure a broader perspective ISSUU don't need to adopt different approach to its users, just needs to take one already existing – the Social Media. The rich social media content provides another angle to the end user interests and the reliable personal information helps locate trends in both ISSUU and Social Media data.

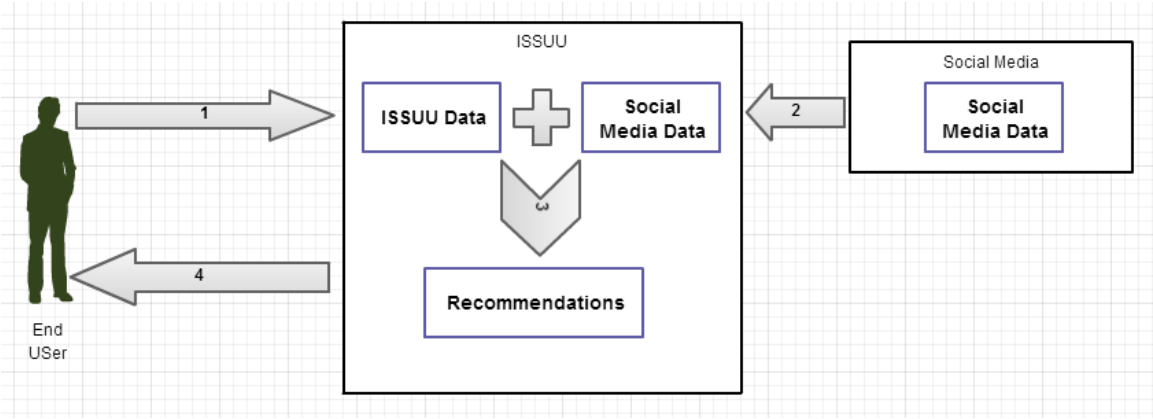


Figure 4: Data Consolidation

1 – The End user is providing his personal information for the ISSUU User Definition and his reading patterns to form the ISSUU Data.
2 – The system takes the End user personal data and posted content from social media to form the Social Media Data.
3 – Compare and/or consolidate the two datasets to create recommendations for the End user.
4 – Make recommendations for reading publications in the fields of interest of the End user.

- Generate more Social Value for End users through creating social media application.

The social Media Application is going to be the bridge that connects the Social Media and ISSUU. It is going to perform important role in satisfying the user needs and achieving the project goals.

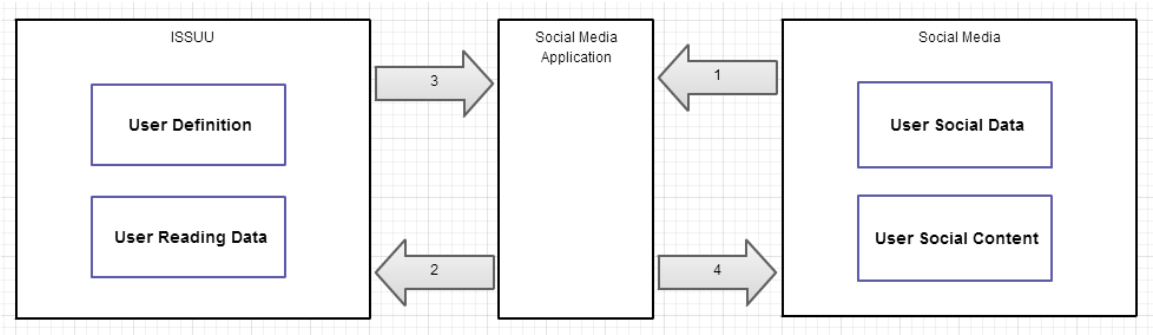


Figure 5: Social Media Application Interaction

1 – Social Media Data is gathered by the Social Media Application.
2 – The converted and analyzed data is transferred to the ISSUU system for safekeeping and further analysis.
3 – ISSUU Data is transferred to the Social Media Application.
4 – The output of the Social Media Application is shared through the Social Media.

As an End user it would be great to have such tool to make the connection between ISSUU and the Social Media. There the End user will be able to see not only what is he reading in ISSUU in an asynchronous graphical way, but he could do the same for the Social Media content made available. Dong this will give more perspective on the one's interests. Moreover sharing his Social Media Application results with his friends/followers could attract attention and increase the interest to the End user.

- Attract more customers through the new Social Media functionalities visualization.

The application is going to introduce new Social Media functionalities to the End user and increase the popularity of the publishing platform thought the Social Media. The end user will be given the opportunity to edit part of his data, visualize it, share it or compare it.

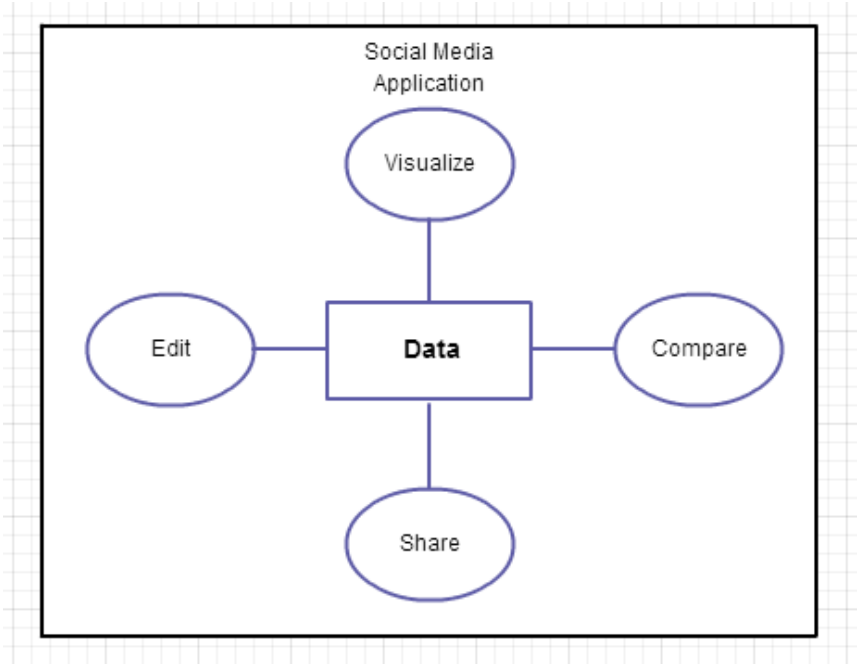


Figure 6: Social Media Application Use Case

The visualization, in form of network graph, is going to be the most compelling visual feature created. Its dynamic interaction with the End user will provoke interest and grab the attention. As an End user it would be great to have such tool that is able to visualize content, making it available for asynchronous graphical comprehension. That will make the End user feel as if he

has his own small data laboratory. Sharing the experience is going to make it even more interesting and involve the End user friends/followers.

2.2. Thesis Goals

The main goal of the project is to design and implement prototype application which is able to acquire ISSUU End User Social Media Data, perform Data Analysis over it, prepare it as basis for personalization of ISSUU content and finally visualize it in form of a network graph.

One of the earliest goals is to create an app prototype, which acquires the End User Social Data. It should be available on the web and allow the performance of first tests of implemented algorithms.

The prototype application should be available through the Social Web and through ISSUU.

The prototype application should work on a web browser and it should be able to cooperate with server responsible for gathering and analyzing the data.

Here the next goal appears – creating of server side Social Data gathering and analyzing mechanism. The gathered Social Data should be used for expanding the ISSUU User Definition and the analyzed Social Media content should be used for personalization of the recommended ISSUU content.

The prototype application then should be able to visualize the analyzed data as network graph.

The last goal is to test and evaluate results of the Social Data analysis and visualization.

All of the thesis goals could be summarized in three general objectives:

1) Increase personalization with Social Media data.
2) Create Data Analysis tools for the acquired Social Data.
3) Visualize the analyzed data.

3. Methods

In this section is discussed what is the network science approach on describing somebody based on relations and what it means to characterize different groups/types inside a network.

3.1. Network Theory

The network science method described in this thesis is based on the work of Albert-László Barabási, described in his book Network Science[vii]. As Barabási formulates:

".. behind each complex system, there is an intricate network that encodes the interactions between the system's components"

Applying that to the given scenario with the various and yet interconnected Social Media content, which is one of the focus points of this project, means that by analyzing the Content Network is it possible to unfold the component interactions and understand the principles behind the person's flow of interests.

The way the content components are connected to each other reveals interesting reflections (se Figure 7): Are there distinct subgroups? From which aspect of your personal history do they come from? Are there any components that connect across the groups? Is there any change in the network dynamics?

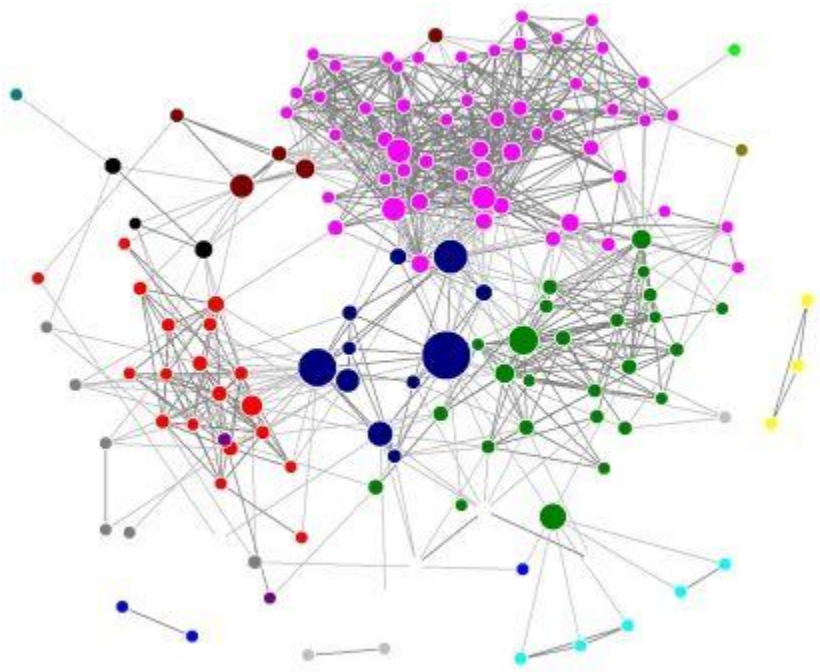


Figure 7: Network components interaction

To be able to describe the behavior of a system consisting of hundreds to billions of interacting components (see Figure 8), what is firstly needed is a map of the system's wiring diagram. In a social media content system, this would require knowing the list of different data types, the specifics of the given types and the lists of the individual components and how they connect.

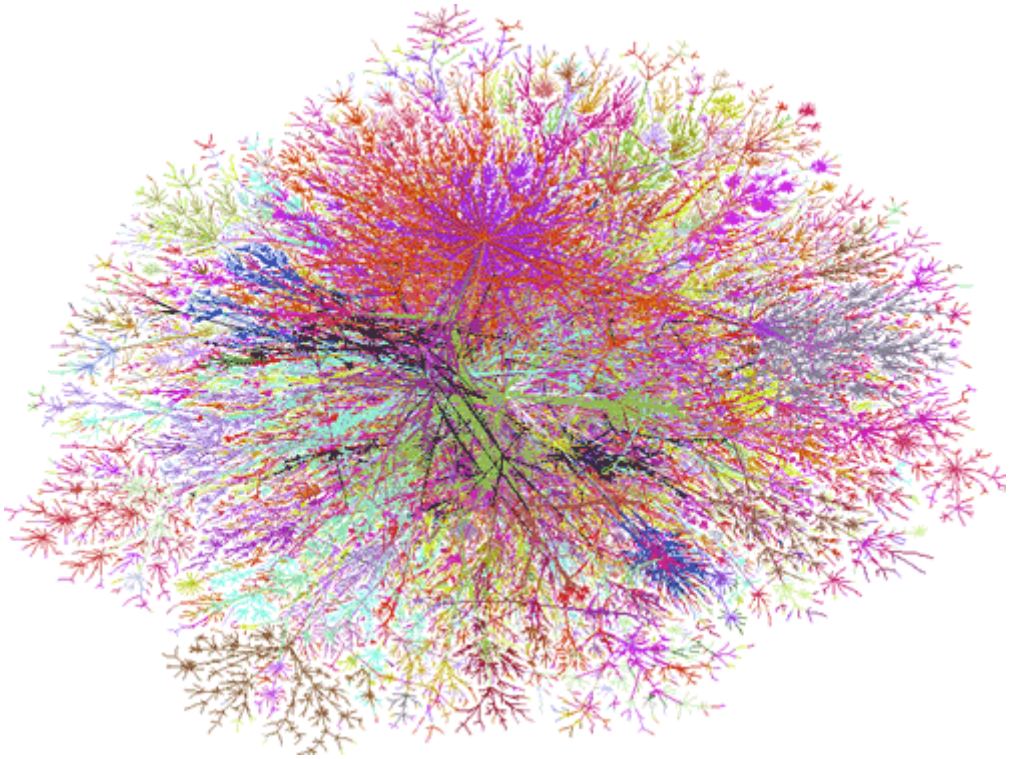


Figure 8: Example of complex network

Moreover, the networks existing around us are not only complex, but they are utterly different – there is big diversity in size, nature, scope, history, and evolution of the networks.

But despite its complexity and variety there is universal similarity in the networks characteristics. The architecture of the various domains of science, nature and technology are similar to each other. That allows the use of a common set of mathematical tools to explore these systems.

As the network science aims to unravel the universality of the networks as a whole, the author uses its tools to uncover specific network properties and will aim to understand the content network origins, encoding the laws that shape the network evolution, as well as its consequences in understanding the content network behavior.

4. Design

In this chapter a plan for the solution is developed. It includes low-level component selection and design as well as the architectural view.

4.1. Overview

As discussed above the ISSUU services consists of three main components – LDA with Collaborative filtering analysis, User Definition and Social Media. To improve the services it is important not only to improve the different components, but also the connection between them (see Figure 9). ISSUU has heavily developed the LDA with Collaborative filtering (component 1) and the connection between (connection 1) that component and the User Definition (component 2), but the other two components – User Definition (component 2) and Social Media (component 3) have had less effort put into them. By improving those components (2 and 3) and strengthening the connections (2 and 3) between the LDA with Collaborative filtering and Social Media, and between Social Media and User Definition the author is going to fulfill a big part of the User Needs.

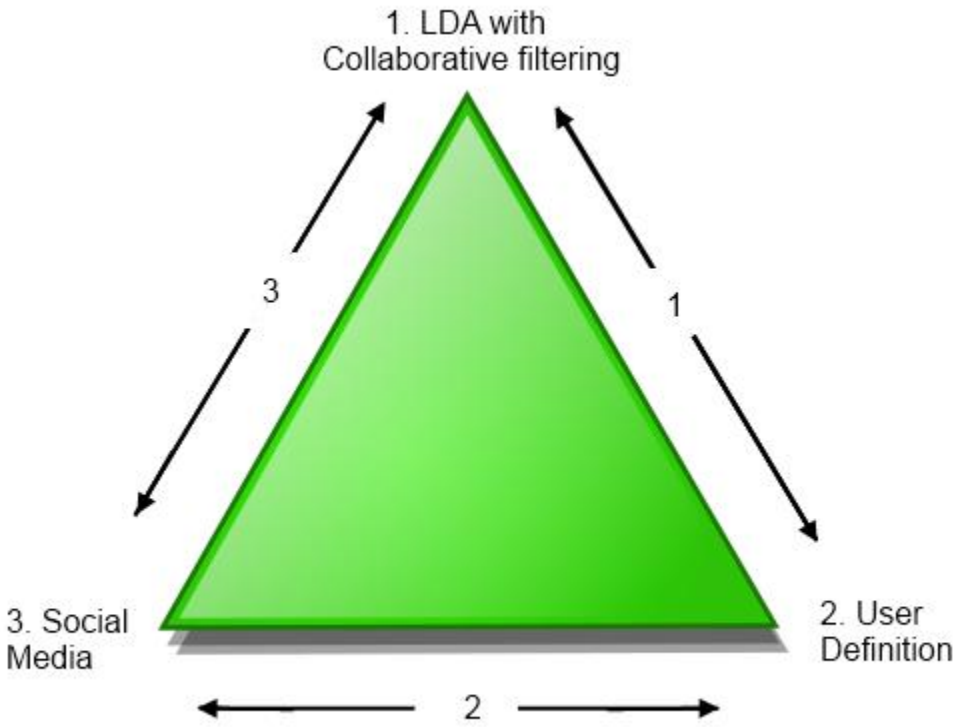


Figure 9: Components Interaction

4.2. Select Social Media

ISSUU has already created integrations with the Social Medias that they want to target and creating integrations for new Social Media is out of the scope for the Master Thesis Project. For the above reasons the author selected Social Media among the ones already integrated with ISSUU. To select the most appropriate Social Media a table has been created with positive and negative sides for each Social Media already integrated with ISSUU (see Table 1).

Facebook		Twitter		Google +	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Facebook Login already existing on the ISSUU platform	The diverse user Social Data posted on Facebook needs careful data selection/filtering	Well documented API with a lot of functionalities	Used to share short text messages – inefficiency of information	Well documented API	Lower User base compared to Facebook and Tweeter
Like button existing on the ISSUU platform		Big User base	No existing data on the ISSUU publications shares	Share over Google + already existing on the ISSUU platform	Lower User activity rate compared to Fb and Twitter
Share over Facebook already existing on the ISSUU platform		Fewer types of data to analyze compared to Facebook and Google +	Users mostly from USA		
Over 40% of the End users have their Facebook connected to ISSUU		Share over Twitter already existing on the ISSUU platform			
Existing information on publications shares by Facebook					
Well documented API with a lot of functionalities					
Big User base					

Table 1: Social Media comparison

As visible from the table above Facebook provides more positives than Twitter and Google + and it is a logical choice for the Social Media selection.

By using Facebook as Social Media data source the author is able to cover more thoroughly the goals of the Master Thesis Project.

4.3. User Stories

By creating User Stories it is easier to identify the user, action and required result in a request. It is a simple way of requesting features that anyone can understand. The stories that are related to each other are grouped in Epics. There are three Epics for the Master Thesis Project:

4.3.1. User Definition

This Epic groups all User Stories that are from ISSUU user point of view and concern the User Definition improvement.

User Definition Epic
As an ISSUU user I want to get public user information about the End user from Facebook so that the User Definition has Age.
As an ISSUU user I want to get public user information about the End user from Facebook so that the User Definition has Gender.
As an ISSUU user I want to get public user information about the End user from Facebook so that the User Definition has Location.
As an ISSUU user I want to get public user information about the End user from Facebook so that the User Definition has known Languages.
As an ISSUU user I want to get public user information about the End user from Facebook so that the User Definition has the End user's Education information.
As an ISSUU user I want to get public user information about the End user from Facebook so that the User Definition has the End user's Work information.

Table 2: User Definition Epic

4.3.2. Social Media Data Analysis

This Epic groups all User Stories that are from ISSUU user point of view and concern the Social Media Data Analysis.

Social Media Data Analysis Epic
As an ISSUU user I want to get the End user posted content from Facebook so that can be filtered and analyzed.
As an ISSUU user I want to get the End user posted content from Facebook so that the End user's interests shared in Facebook can be identified.
As an ISSUU user I want to get the End user interests shared in Facebook so that can be compared to the End user's ISSUU publications interests.
As an ISSUU user I want to get the End user ISSUU publications shared/liked in Facebook so that can be compared to the End user's ISSUU publications interests.
As an ISSUU user I want to get the End user ISSUU publications shared/liked in Facebook so that can be compared to the End user's interests shared in Facebook.

Table 3: Social Media Data Analysis Epic

4.3.3. Facebook App

This Epic groups all User Stories that are from End user point of view and concern the creation of a Facebook App.

Facebook App Epic
As an End user I want to retrieve my posted content from Facebook so that can be visualized as content network.
As an End user I want to retrieve my friends' posted content from Facebook so that can be visualized as content network.
As an End user I want to automatically find out my Facebook content network topics so that they can be visualized in the content network.
As an End user I want to automatically find out my friends' Facebook content network topics so that they can be visualized in the content network.
As an End user I want to automatically find out my Facebook content network node importance so that can be visualized in the content network.
As an End user I want to automatically find out my friends' Facebook content network node importance so that can be visualized in the content network.

Table 4: Facebook App Epic

The above User Stories are going to be divided into Work Items, which in turn are going to be divided into Tasks, simplifying the development process and creating a roadmap for the Master Thesis Project.

4.4. User Definition

In order to improve the User Definition in the ISSUU platform the connection between the Social Media Facebook ant the platform needs to be used for gathering more information about the End user.

4.4.1. Improving the User Definition with Social Data

It is important to improve the User Definition for a couple of reasons:

- There are more criteria added for observing trends while performing data analysis.
- When an End user is registering(through Facebook Connect) and entering for a first time into ISSUU, there are no read publications by that End user and in that case the Recommendation system (based on the LDA with Collaborative filtering on the read publications) is not able to make personalized recommendation. But if an already existing information form Facebook is used, then the Recommendation system has personal information about the End user and is able to suggest a publication in the field of interest.
- The User Definition can take a small role in the forming of the recommendations, based on the available Social information for the End user.

By carefully selecting and grouping personal information about the End user the User Definition is going to be improved. Information like Gender and Age could indicate interests and give more criteria for later data analysis (e.g. collaborative filtering). Location and known Languages are

also going to enrich the criteria for the data analysis phase, but also give indication about the linguistic reading preferences of the End user. To add to the User Definition information about Work and Education could be beneficial in data analysis to observe trends connected to person's education and field of work. Personal likes could be used as pointers to what the person's general interests are.

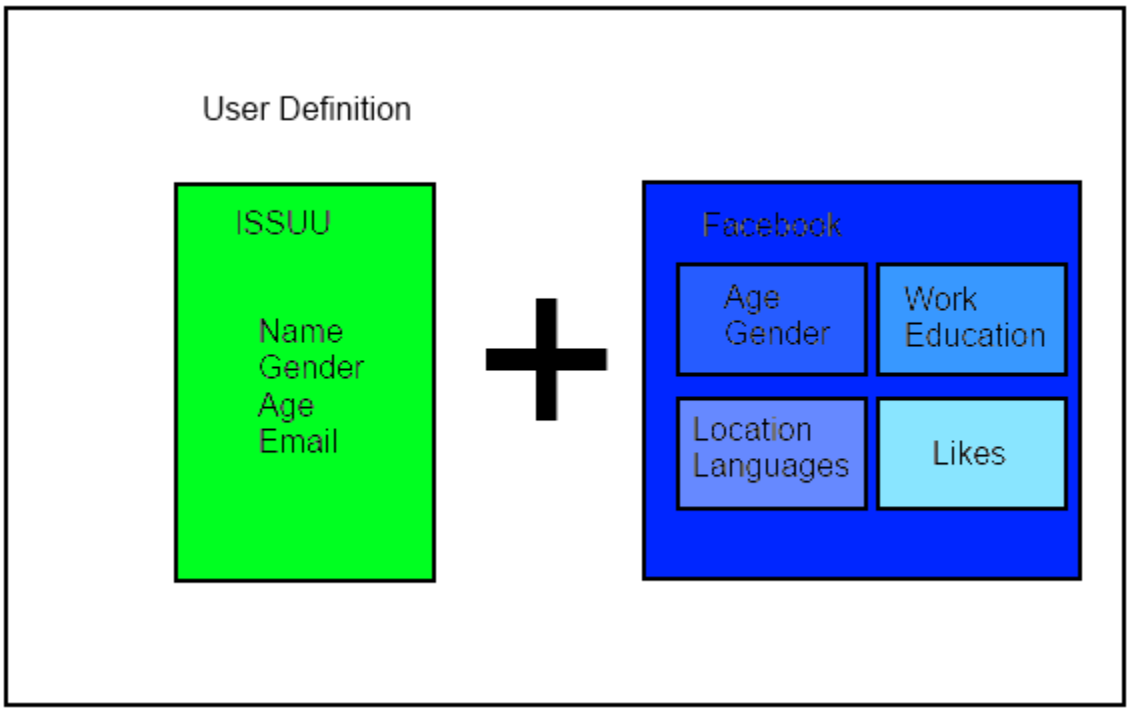


Figure 10: User Definition

Adding the mentioned above Facebook personal user data is going to happen upon End user registration and is going to be periodically updated over time diapasons of 1-6 months depending on system limitations and best practices.

Sensitive personal information as Religious views, Political views and Relationship Status are better left out of the User Definition because classifying End users on those criteria could be offensive for some individuals. That way the improvement in the ISSUU services could get a negative reaction from the public and repel customers, and that is against the goals of the Master Thesis Project.

Other way of enhancing the User Definition is by extending the ISSUU perspective on the users. According to ISSUU the End User is what he reads. That is not always accurate enough – there are fields in which the End user has interest, but are not in his reading patterns. Information on the End user fields of interest could be gained through the already existing data on the Social

Media. Combining the two different ways to describe a user gives more perspective and helps to find the exact area of the End user interests.

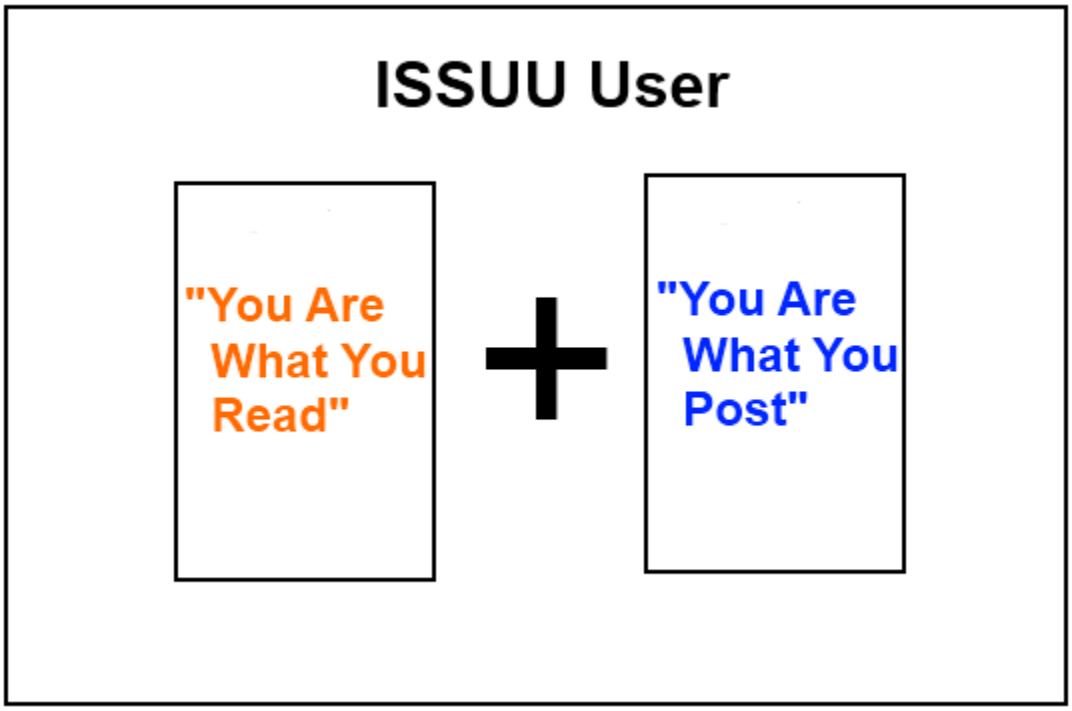


Figure 11: Defining the End user

The desired result is to intuitively recommend publications based not only on the End user reading patterns, but also on his Social Media content. Naturally, not all of the posted by the End user social media content is going to be of interest and that's why it is essential to perform data analysis over the gathered data, in order to identify the important topics and key components of the content network.

4.4.2. Contradiction in the End User Data

In some cases there could be contradiction between the stated in ISSUU by the End user personal information and the one obtained from Facebook. The best approach is to select the information from the Social Media, because that information trends to be the correct one – people are less likely to falsely state their gender in the Social Media than in the quick two-step registration for ISSUU.

4.4.3. Obtaining the Facebook End user Data

Obtaining the End user's Facebook personal data is going to be implemented through Facebook's Graph API. It is a very well documented and features–reach API that provides easy access to the user basic public information. The User Definition is going to be improved for each End user who has connected his Facebook account to his ISSUU account.

4.4.4. Integrating with ISSUU

ISSUU is keeping the User Definition information in a database. By extending the User database table and transferring the End user personal information from Facebook to the ISSUU platform the author will ensure that the information is available all the time for a fast use.

4.5. Social Media Data Analysis

In this section a design for the data analysis will be discussed and potential gains coming from it. The selected method of obtaining the data is through the Facebook's Graph API [viii]- it is the primary way that data is retrieved or posted to Facebook.

4.5.1. Data Selection

Selecting the data that is going to be analyzed is one of the most important aspects of the data analysis. There are a several types of data (Objects[ix]) that is possible to take from Facebook:

Achievement(Instance)	The achievement(Instance) object represents the achievement achieved by a user for a particular app.
Album	The User, Page and Application objects have an albums connection of type album.
Application	An application registered on Facebook Platform as represented in the Graph API. Applications a user administers can be retrieved via the /accounts connection on the User object.
Checkin	A Checkin represents a single visit to a location. The User and Page objects have checkin connections.
Comment	A comment on a Graph API object
Domain	A web site domain within the Graph API.
Errors	Requests made to Facebook API can result in a number of different error responses, however there are only a handful of basic recovery tactics.
Event	Specifies information about an event, including the location, event name, and which invitees plan to attend. The User and Page objects have an events connection.
FriendList	A Facebook friend list. This object represents the list itself and not the members of the list. The User object has a friendlists connection
Group	A Facebook Group. The User, Page and Application objects have groups connections.
Insights	An object containing statistics about applications, pages, and domains. The Page, Application and Domain objects have an insights connection of this type. Post objects also have this connection where the Post was made by a Page. Available metrics include application and page hits, adds, removes, and likes.
Link	A link shared on a user's wall. The User, Application, and Page objects have a links connection.

Message	An individual message in the new Facebook messaging system. Every message has a message ID that represents an object in the Graph.
Note	A Facebook Note. The User object has a notes connection.
Offer	An Offer represents an offer that is published by a page. Only Page objects have offers connections.
Order	You can use the order object to interact with orders created by the application using Facebook Payments to view and update orders as needed.
Page	Pages are for businesses, organizations and brands to share their stories and connect with people. Like timelines, you can customize Pages by adding apps, posting stories, hosting events and more.
Photo	An individual photo as represented in the Graph API.
Pictures	You can render the current profile picture for any object that has a picture associated with it.
Post	An individual entry in a profile's feed as represented in the Graph API. The User, Page, Application and Group objects have feed connections containing post objects that represent their walls. In addition the User and Page objects have a connection named posts containing Posts made by the User and the Page respectively.
Privacy Parameter	The privacy parameter controls the audience on Facebook that can see a post made by an app on behalf of a user. Its value is a JSON object that can be applied to a single Open Graph action, status update, photo, video, album, or other kinds of content posted by apps using Facebook APIs.
Publishing	A publishing performed by a user, as represented in the Graph API.
Question	A question asked by a user, as represented in the Graph API.
QuestionOption	An option allowed as an answer to a Question, as represented in the Graph API.
Realtime Updates	The Graph API supports Realtime Updates to enable your app to subscribe to changes in data.
Review	An object representing a review for an application. The Application object has a reviews connection containing review objects.
Status message	A status message on a user's wall as represented in the Graph API.
Thread	A message thread in the new Facebook messaging system as represented in the Graph API. The User object has a threads connections.
User	A user profile as represented in the Graph API.
Video	An individual Video in the Graph API. This will return videos that the user has uploaded or has been tagged in.

Table 5: Graph API Objects

All of the listed Objects are composed of Fields and Connections. Many of those above Fields require special permissions from the user to be obtained. Therefore the selection of data to be analyzed must be done with precaution not to require too much access permissions from the End user. Requiring too much access could be viewed as personal information interference and the request for permission could be denied.

The End user public basic information does not require access permission to obtain and is going to be used for improving the ISSUU User Definition. That information is in the **User** Object.

To obtain the posts of the End user, **Post** Object permissions are needed. That object contains all posts on the End user Facebook wall.

The below table shows which fields are going to be used for retrieving data from Facebook and the permissions needed.

Field Name	Description	Permission	From Object
id	The user's Facebook ID	access_token	User
name	The user's full name	access_token	User
first_name	The user's first name	access_token	User
middle_name	The user's middle name	access_token	User
last_name	The user's last name	access_token	User
gender	The user's gender: female or male	access_token	User
locale	The user's locale	access_token	User
age_range	The user's age range; only returned if specifically requested via the fields URL parameter	access_token	User
languages	The user's languages	user_likes	User
books	The books listed on the user's profile.	user_likes	User
games	Games the user has added to the Arts and Entertainment section of their profile.	user_likes	User
likes	All the pages this user has liked.	user_likes	User
movies	The movies listed on the user's profile.	user_likes	User
music	The music listed on the user's profile.	user_likes	User
television	The television listed on the user's profile.	user_likes	User
education	A list of the user's education history	user_education_history	User
location	The user's current city	user_location	User
work	A list of the user's work history	user_work_history	User
id	The post ID	access_token	Post
from	Information about the user who posted the message	access_token	Post
to	Profiles mentioned or targeted in this post	access_token	Post
message	The message	access_token	Post
caption	The caption of the link (appears beneath the	access_token	Post

	link name)		
description	A description of the link (appears beneath the link caption)	access_token	Post
place	Location associated with a Post, if any	read_stream	
comments	Comments for this post	read_stream	Post
created_time	The time the post was initially published	read_stream	Post
updated_time	The time of the last comment on this post	read_stream	Post

Table 6: Fields and permissions required

The fields pointed in the above table are sufficient to meet the requirements for the Master Thesis Project, providing all the needed Social Media information for the End user.

As visible from Table 6 there are only five access permissions required. The range of the data retrieved from Facebook could be tuned on during Implementation/Integration phase if it is too much/not enough.

4.5.2. Data Cleaning

Data cleaning is an important procedure during which the data are inspected, and erroneous data is corrected. Data cleaning can be done during the stage of data entry. If this is done, it is important that no subjective decisions are made. A guiding principle is: during subsequent manipulations of the data, information should always be cumulatively retrievable. In other words, it should always be possible to undo any data set alterations. Therefore, it is important not to throw information away at any stage in the data cleaning phase. All information should be saved and all alterations to the data set should carefully and clearly document.

After obtaining the Facebook data the cleaning process that needs to be applied in order to omit the parts that give no information on the End user interests. For the purpose a NetworkX cleaning techniques are intended.

The goal of the Data Cleaning process is to leave only meaningful text content about the End user that could be analyzed.

4.5.3. Initial Data Analysis

The most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that are aimed at answering the research question. The quality of the data should be checked as early as possible.

- Checks on data cleaning: have decisions influenced the distribution of the variables? The distribution of the variables before data cleaning is compared to the distribution of the variables after data cleaning to see whether data cleaning has had unwanted effects on the data.
- Analysis of missing observations: are there many missing values? The missing observations in the data are analyzed to see whether more than 25% of the values are missing. If so a correction must be made.

- Analysis of extreme observations: outlying observations in the data are analyzed to see if they seem to disturb the distribution.

4.5.4. Main Data Analysis

The main data analysis needs to be done with the already applied LDA with Collaborative filtering methodology that ISSUU is using. They are already very proficient in applying the LDA with Collaborative filtering analysis and have developed methodology to do the analysis. That gives precise idea about the topics of the End user's posts and gives the possibilities of comparing the Social Data with ISSUU Data of the End user.

It does not make sense to use another data analysis methodology for the data comparison/conjunction, because that is going to cause loss of the ability to compare Social Data with ISSUU Data – it will be like comparing oranges with apples.

On the other hand, the benefits that the network theory data analysis offers are undeniable. As a complex system the Content Network has underlying structure that could be understood with the tools of the Network Science. Performing the second type of Data Analysis is going to benefit both the End user and ISSUU. For the end user it will supply an intriguing network available for visualization and for ISSUU it will provide another useful tool for data analysis.

By performing both types of the Data Analysis different project goals are addressed and that gives the analysis phase an integrate and complete and form.

4.5.5. Save the Social Data

In order to have all the social data (both raw and analyzed) available at all times, the data is going to be saved in the ISSUU database with the respective timestamp.

4.5.5.1. Raw Data

The database tables holding the raw data are going to have simple structure – a table called "facebook_raw" referenced to the already existing user table – see Figure 12:

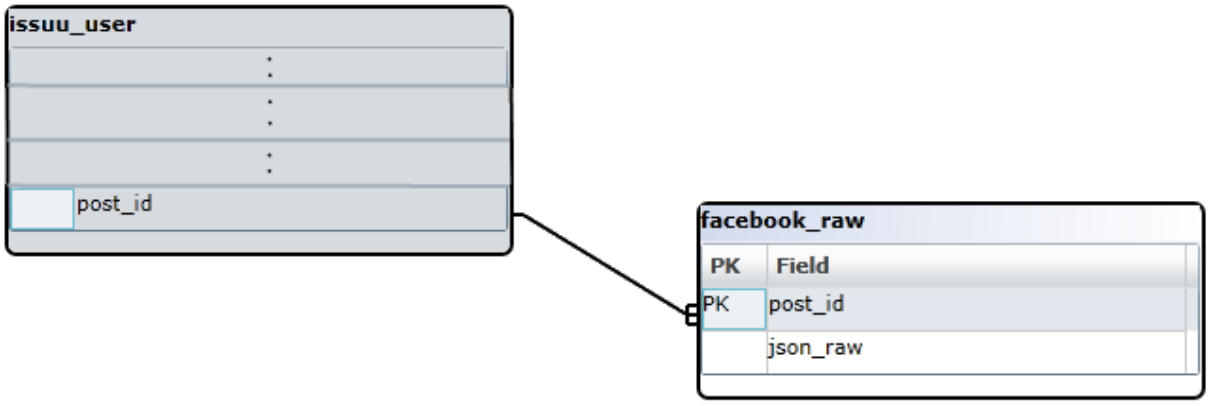


Figure 12: Facebook Raw Data table

4.5.5.2. Analyzed Data

ISSUU already has devised database structures to save LDA analyzed data and the project is going to take advantage of those structures.

4.5.6. Compare the Data

After applying the LDA analysis the three types of data could be compared:

- End user interests analyzed from his Facebook posts
- End user Facebook likes of ISSUU publications
- End user all ISSUU publications readings

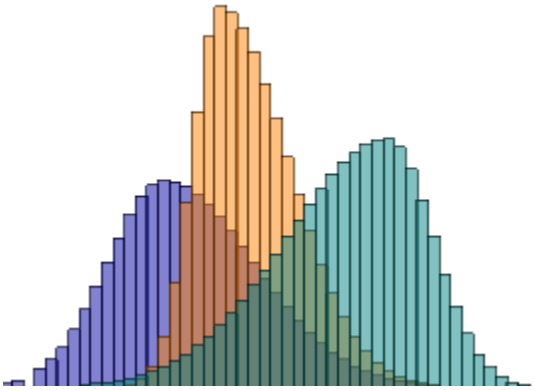


Figure 13: End user Interests

As expectation the comparison between those three data sets is going to ensure a better understanding of the End user interests and that way provide improvement in the ISSUU recommendation system.

4.5.7. Analyze Data Timeline

For each of the three data sets a timeline is going to be created – each entry saved with a timestamp. That could show:

- Change of interests/readings/likes of the End user over time.
- How active is the End user in during different time periods.
- Is there any leading trends of interest change – e.g. first the end user starts to write about some topic on Facebook and then started to read ISSUU publications on the topic.

4.6. Facebook App

The best way to fulfill the User Needs for increasing the Social Media value for ISSUU is to create an application that provides new functionalities to the ISSUU End users. The new application has to fulfill two purposes:

- Increase value for existing ISSUU customers.
- Attract new customers.

Providing additional functionalities is going to increase the value for the existing customers and creating attractive visuals (shared through Facebook) is going to draw in new customers.

The new functionalities that the Facebook application is going to provide are in the area of the social data analysis – more precisely in visualizing the analyzed data. Since the LDA analysis provides a multi-dimensional structure which is very difficult to visualize and most of the time even more difficult to understand, the author decided to simplify the visualization for the sake of the End user. The data gathered from the End user's Facebook posted content is going to be represented as network graph according to the Network Theory methodology described in Chapter 2. Simplified - the nodes of the network are the words and the edges are the connections between those words.

The goals of the designed Facebook app are:

- Analyze the Social Data according to the Network Theory methodology
- Visualize the analyzed data in a dynamic graphical way
- Highlight the different topics
- Highlight the importance of the different nodes

Since the developed Facebook app is only a prototype, the only supported language on that stage is going to be the English language. In the further development of the ISSUU Facebook App more languages should be added to the supported list.

4.6.1. Content network visualization

To visualize the content that the End user has been reading rom ISSUU and posting on Facebook in an interesting and attractive way, the visualization needs to be:

- Interactive: Dynamically respond to the End user actions
- Interesting: Attract the attention of the End user with something meaningful
- Informative: Give a new information about the displayed data

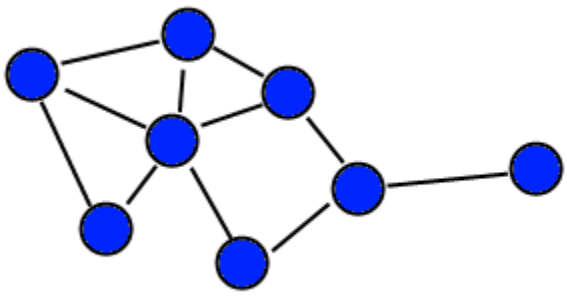


Figure 14: Content Network visualization

To accomplish the above visualization goals the following functionalities has to be implemented:

- When an End user hover the mouse over the content network, that area should be magnified and the name of the nodes should show.
- The applied layout should compute the nodes positions and move the nodes in front of the End user, showing hidden structure in the initially viewed as “hairball” content network.
- Run algorithms over the content network showing different topics in the text and how they are connected.
- Run algorithms over the content network showing the degree distribution of the different nodes.

Using dynamically visualized networks shared over Facebook is going to attract more customer interest in ISSUU and increase the value for the established ISSUU users.

4.6.1.1. Degree Distribution

In the study of networks a degree of a node in a network is the number of connections it has to other nodes and the Degree distribution [x] is the probability distribution of these degrees over the whole network. When applied to the End user content network it should look like the network on Figure 15.

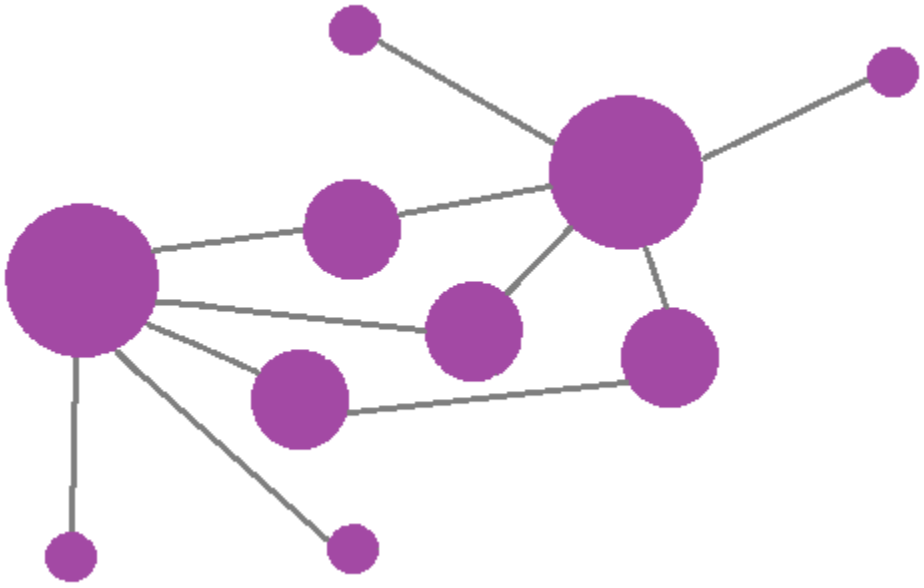


Figure 15: Network Degree Distribution

4.6.1.2. Modularity

Modularity[xi] is a measure of the structure of networks. It was designed to measure the strength of division of a network into clusters. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

The author is going to use Modularity to detect different topics in the End user content network. When applied, the content network should look like the network on Figure 16.

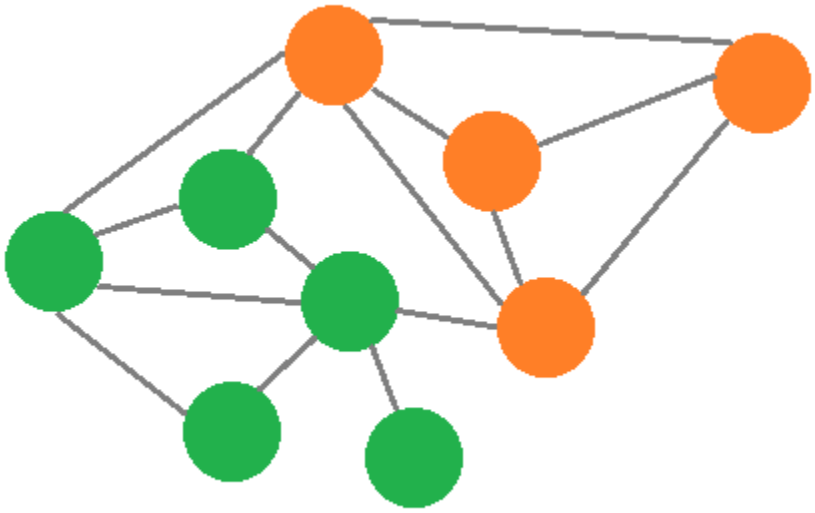


Figure 16: Network Modularity

4.6.1.3. The Resulting network

After applying the two algorithms, the resulting network with Modularity and Degree Distribution should look like Figure 17.

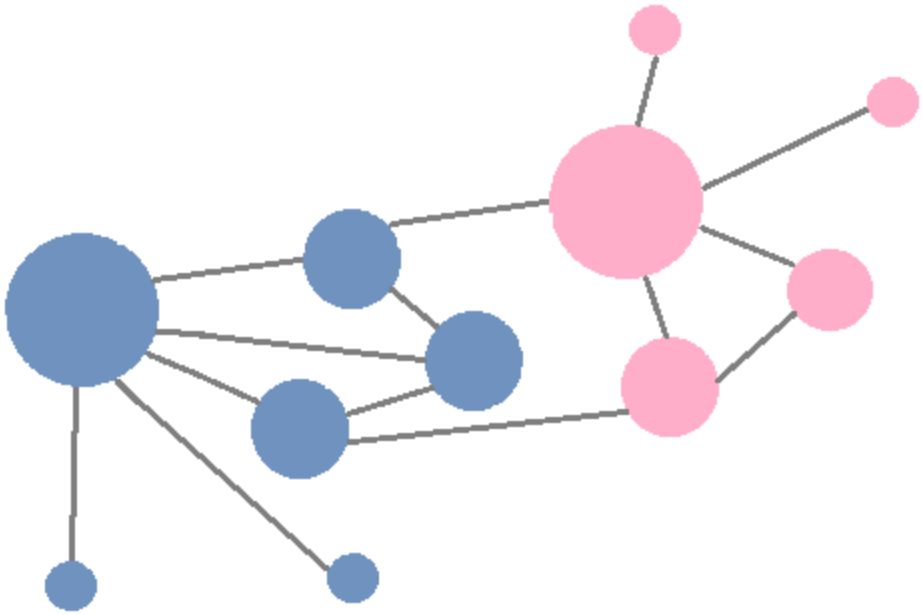


Figure 17: The resulting network

4.6.2. App interface

The app interface prototype has been created with the Axure [xii]software. Axure is interactive wireframe software that gives a way to quickly and easily deliver prototypes. It generates an interactive HTML website wireframe or UI mockup without coding. Here is a brief representation of the generated prototype. For the full prototype specifications go to Appendix A, also available online[xiii].

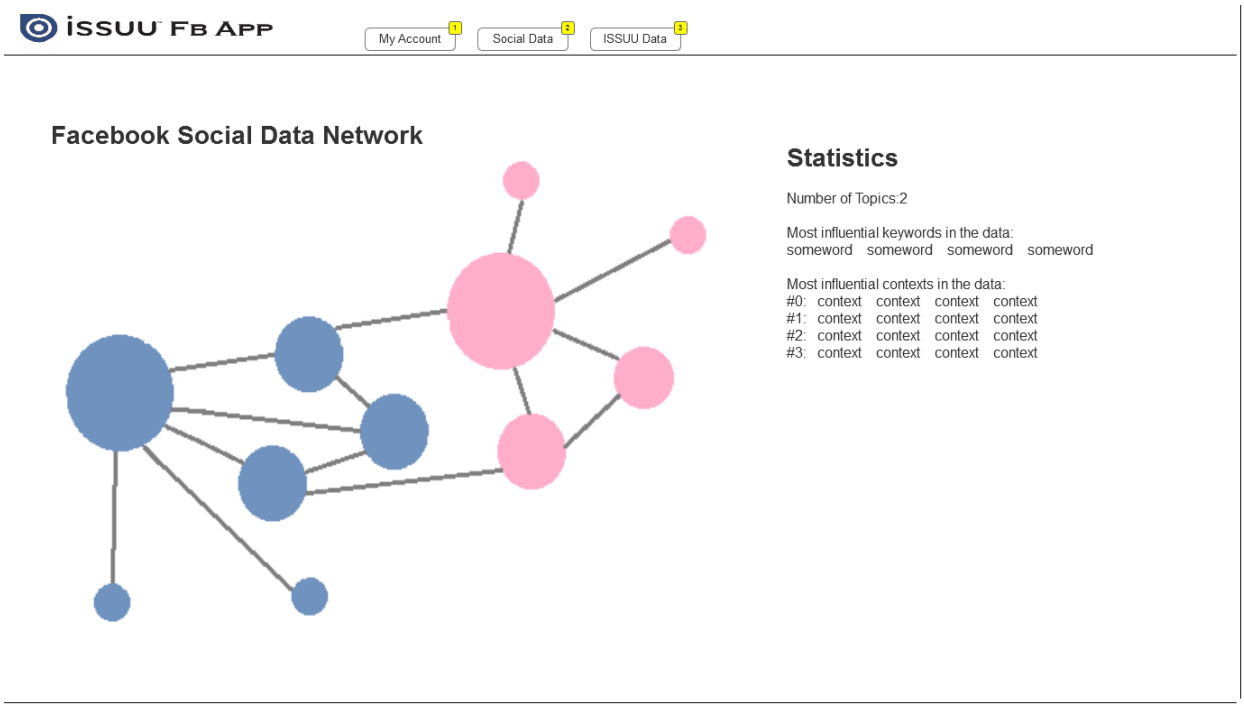


Figure 18: ISSUU Fb App interface

The above figure represents a prototype of the ISSUU Fb App interface. The End user will be able to see his content network visualization upon login. He will also be able to view and edit his account information and his ISSUU Data visualized as network.

4.7. Use case

In this chapter the author is going to discuss the use case for the End user and show the sequence diagrams for the ongoing processes.

4.7.1. Formal system Use case

Use case diagram (Figure 19) explains the actor's interaction with the system and depicting the specifications of a use case. It clearly shows difference in interactions and roles for each component.

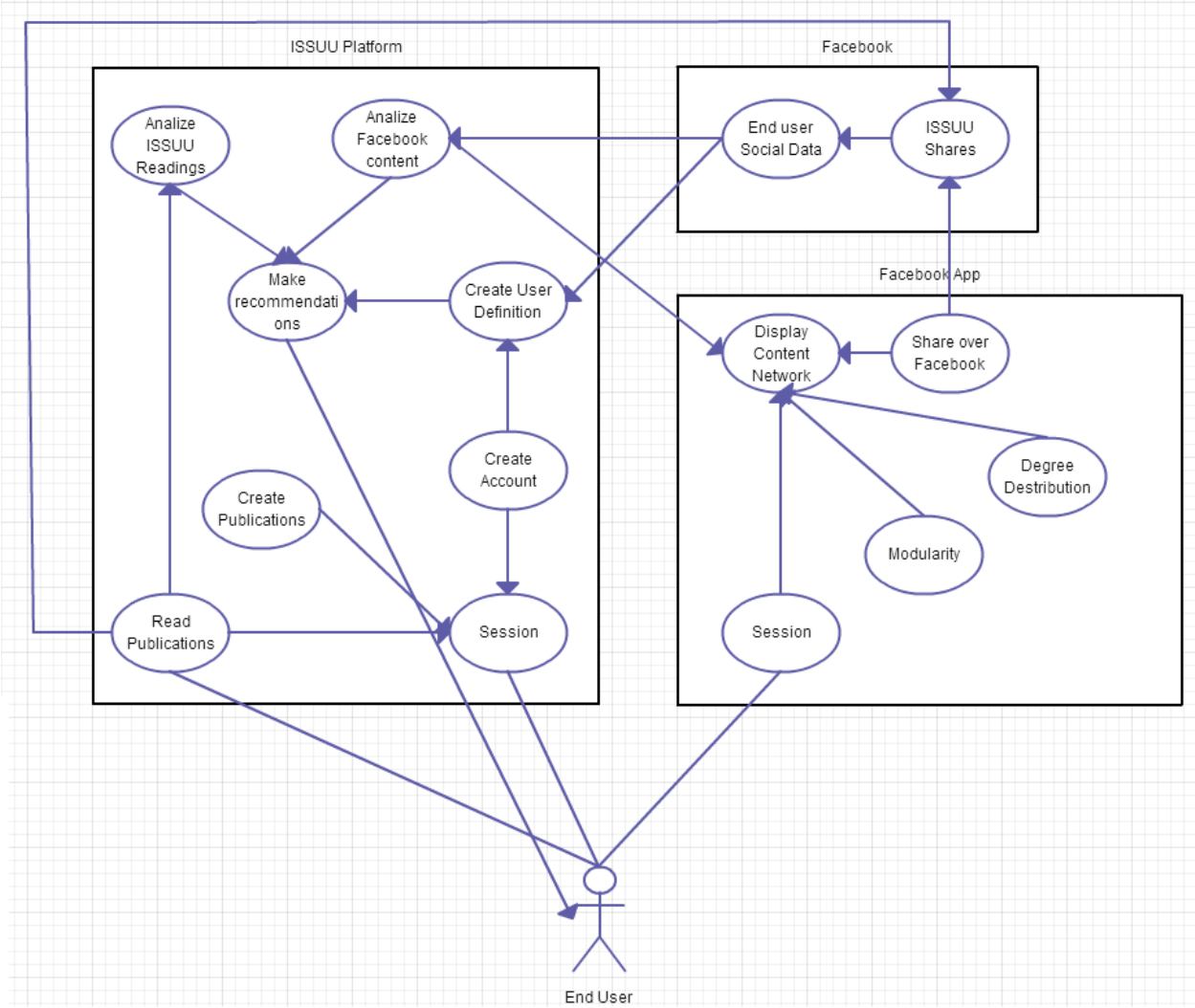


Figure 19: End user Use case

The above use case describes the use case for the End user. He has access to the ISSUU services, Facebook and the developed Facebook application. It clearly shows the interaction between actor and systems.

4.8. Sequence diagrams

The below sequence diagrams[xiv] show how processes operate with one another and in what order. They show object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. The parallel vertical lines (lifelines) show different processes or objects that live simultaneously and as horizontal arrows, the messages exchanged between them in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner.

4.8.1. ISSUU login

The below sequence diagram (Figure 20) explains the sequence of interaction between the end user and the ISSUU platform. The diagram is for End users with connected Facebook account to their ISSUU account – Connected users.

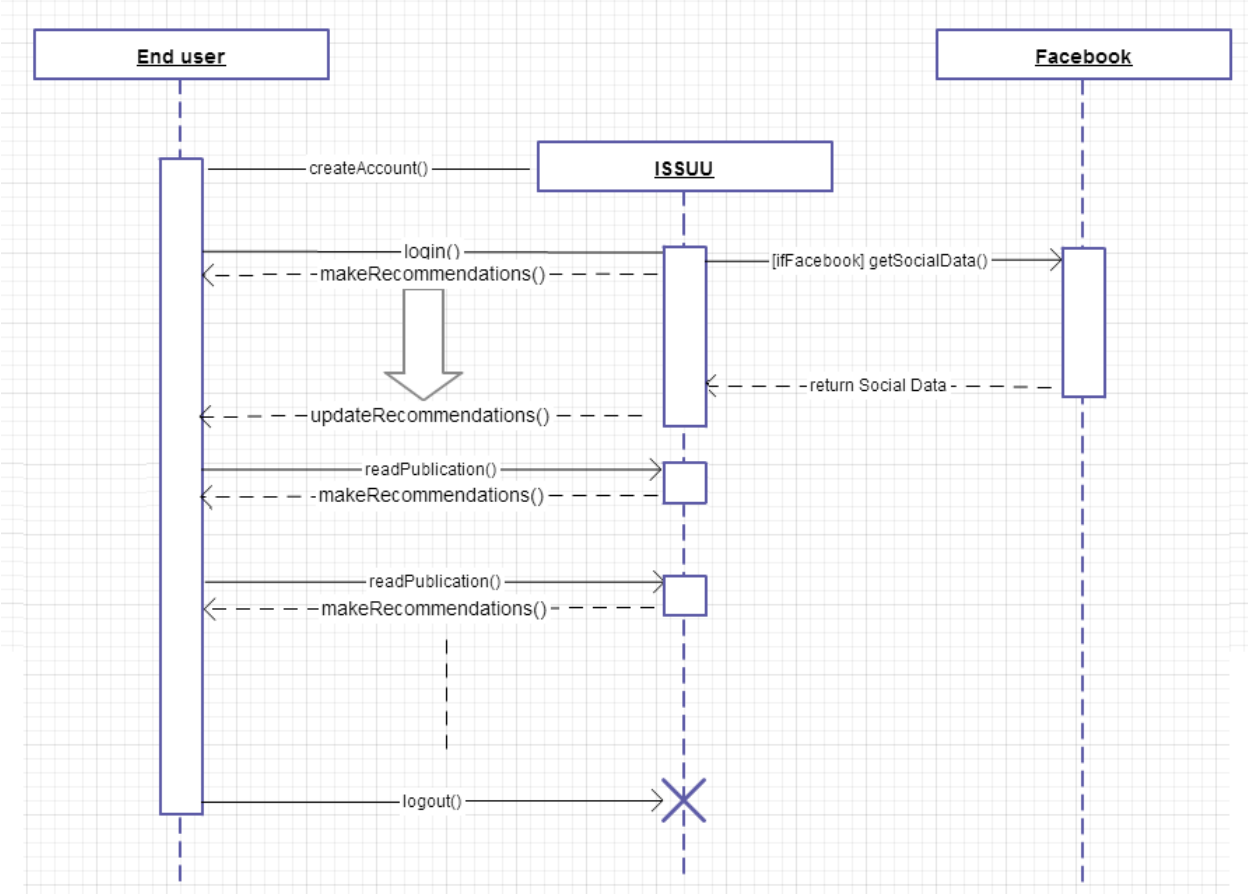


Figure 20: ISSUU login sequence diagram

The above sequence diagram is a simplified case where only events that are significant for the Project are displayed.

4.8.2. Facebook App

The below sequence diagram (Figure 21) explains the sequence of interaction between the end user and the ISSUU Facebook App. The diagram is for Connected users that are using the ISSUU Facebook App.

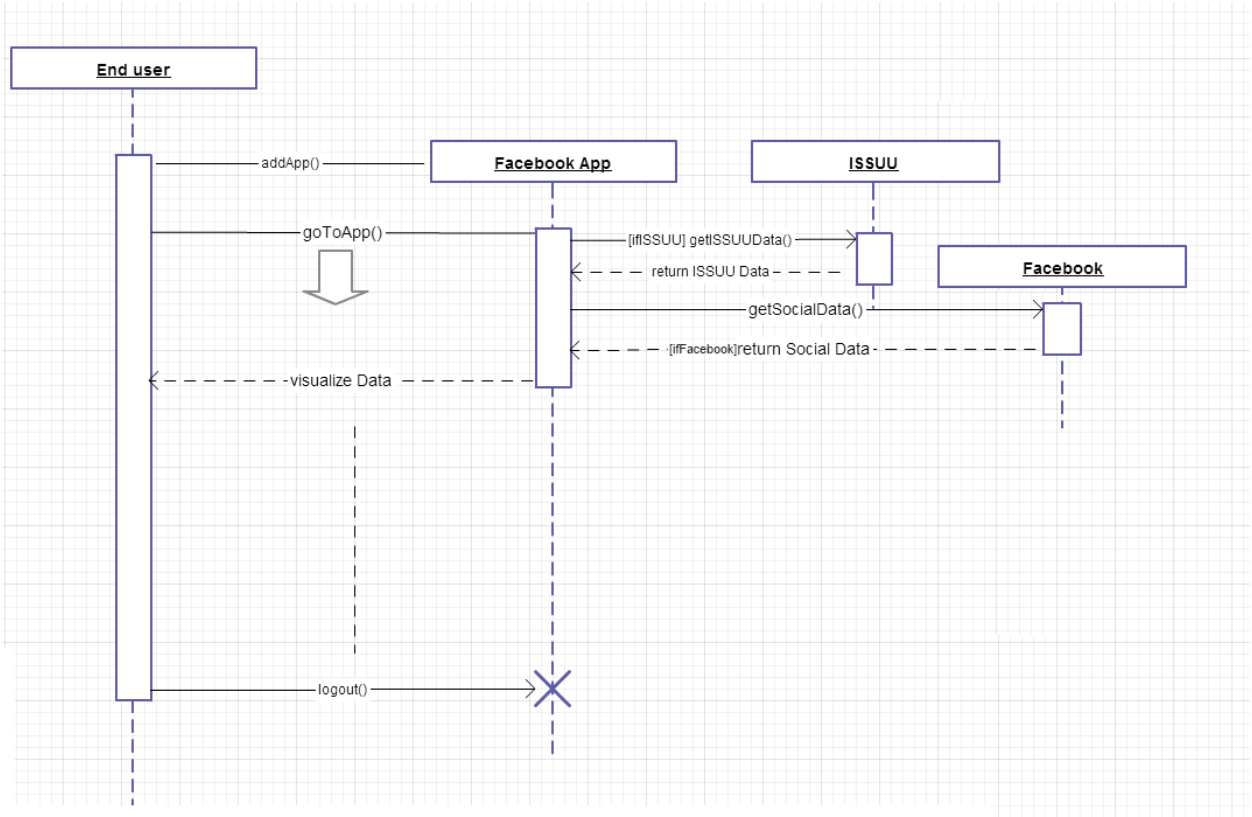


Figure 21: Facebook App sequence diagram

The diagram shows the interactions between the End user, ISSUU Facebook App, Facebook and ISSUU. An update of the ISSUU Data and the Social is performed if the available datasets are not up to date. Then the datasets are visualized.

5. Implementation

In this section the author is going to describe the used technologies for the implementation of the Facebook Application and how they were adapted to achieve the goals of the Master Thesis Project.

5.1. Adapting Technologies

The author's goal is not to "reinvent the wheel" in form of creating new technologies for the needed result, but to adapt already existing technologies to serve the intended purposes.

Implemented technologies table (Table 7) gives a brief overview of technologies used to achieve different goals. Each technology is described in more details later in this section.

Goal	Technology used
Retrieve Social Data from Facebook	Graph API and FQL
Application hosting	Heroku – cloud application platform
Programming language	Python
Applying Modularity algorithm	networkX
Applying Degree Distribution algorithm	networkX
Interface	HTML pages generated using Python scripts
Interface design	CSS
Visualize network	Sigmajs – JavaScript library

Table 7: Implemented technologies

5.1.1. Facebook Graph API

Much of the data that is going to be accessed via the Graph API requires the user to allow access to it as discussed in the design chapter. This includes reading anything beyond public data or writing data to a user's timeline.

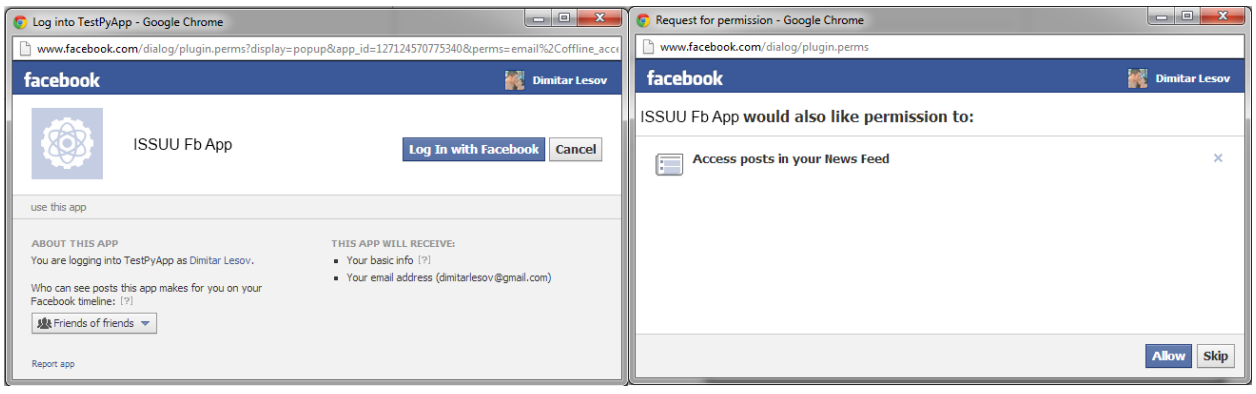


Figure 22: Fb App Access permissions

To access the Graph API fields the author is using FQL[xv]. Facebook Query Language enables the use of a SQL-style interface to query the data exposed by the Graph API. It provides for some advanced features not available in the Graph API, including batching multiple queries into a single call. It's important to batch a bunch of requests together and make them all at once,

instead of having to make a bunch of small requests - this can improve latency and overhead associated with making a large number of changes.

5.1.2. Heroku

Heroku [xvi]is a cloud application platform for web hosting of apps. It is a partner [xvii]of Facebook in creation of Facebook applications. Heroku is supporting the following environments:

- PHP
- Ruby
- Node.js
- Python

Apps hosted on Heroku are operated by processes called dynos (se Figure 23). A dyno, the basic unit of composition on Heroku, is a lightweight container running a single user-specified command. The commands run within dynos include web processes, worker processes and any process types declared in the app.

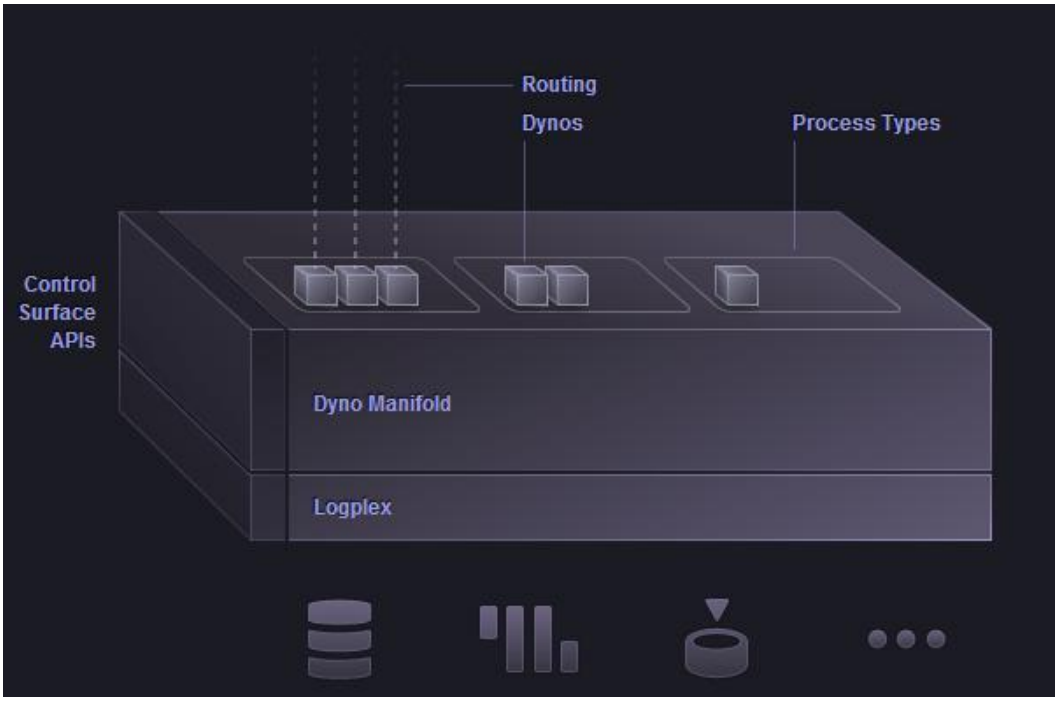


Figure 23: Heroku platform

For more information on the Heroku platform specifics look into How it Works[xviii].

For creating the ISSUU Facebook App prototype, Heroku platform is the perfect choice, because provides free hosting and it is tightly integrated with Facebook.

5.1.3. Python

Python[xix] is a general-purpose, high-level programming language whose design philosophy emphasizes code readability. It is the optimal programming language of choice for the Facebook App for the following reasons:

- Python's powerful network processing library NetworkX [xx] provides the necessary algorithms for the networks computations.
- Easy to create Web/Mobile apps with Python
- Python is supported by the hosting platform

Python is going to be used for the Facebook application scripts as well for the content network processing.

5.1.4. NetworkX

NetworkX is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

NetworkX Features:

- Python language data structures for graphs, digraphs, and multigraphs.
- Nodes can be "anything" (e.g. text, images, XML records)
- Edges can hold arbitrary data (e.g. weights, time-series)
- Generators for classic graphs, random graphs, and synthetic networks
- Standard graph algorithms
- Network structure and analysis measures
- Basic graph drawing
- Open source BSD license
- Well tested: more than 1500 unit tests
- Additional benefits from Python: fast prototyping, easy to teach, multi-platform

By representing the content network as NetworkX graph data structure the author is going to use its features to find out the Modularity and Degree Distribution of the network. Then the whole network is saved as gexf [xxi]file, ready to be visualized.

5.1.5. Sigmajs

Sigmajs [xxii]is an open-source lightweight JavaScript library to draw graphs, using the HTML canvas element. It has been especially designed to:

- Display interactively static graphs exported from a graph visualization software - like Gephi
- Display dynamically graphs that are generated on the fly

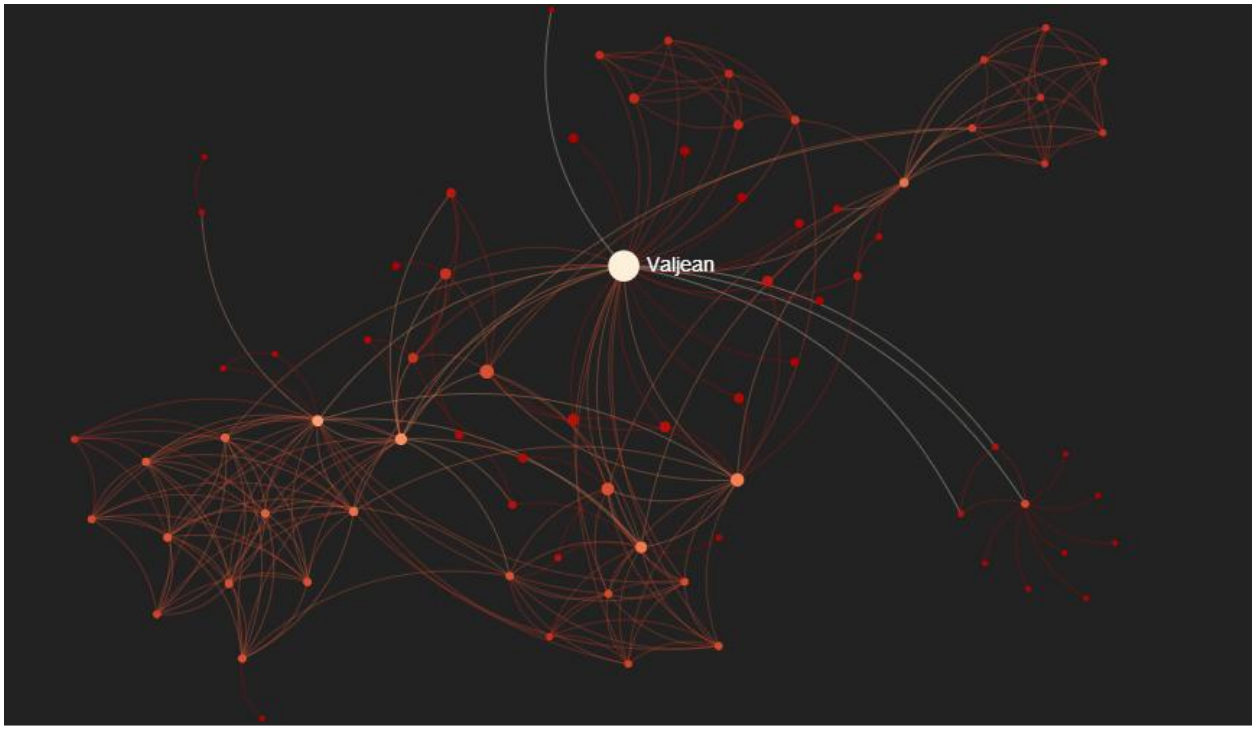


Figure 24: Sigmajs network visualization

5.2. Testing

Testing is the process of executing a program or system with the intent of finding errors. It involves any activity aimed at evaluating an attribute or capability of a program or system and determining that it meets its required results. This section explains more about performed tests and received results.

4.3.1 Functionality tests

Functionality tests are performed to detect possible errors in implementation/programming and to make sure system meets the requirements. Implementation of every functionality was tested during the Facebook App development sprints which helped to find programming errors.
Functionality tests in this project were performed to test:

- All social data acquiring (described in Section 3.4.5 Obtaining the Facebook End user Data);
- Correct functionality of system (described in Use cases described in Section 3.8.1 Formal System Use Case);

Functionality tests were performed three times:

- Two times during implementation sprints three and four. Such kind of tests help to find program and logic errors. Usually implementation tests fail and they should be used to find weak spots;

- One time after the Facebook App was implemented and all errors were fixed (final tests). Such tests are used to validate correct system workflow and make sure no errors were left.

This section describes several implementation tests, which showed potential problems in system workflow. Presented tests show a good example of typical errors that occurred during implementation. Tests were performed during implementation sprints and all errors found by those tests were fixed later. Section presents three types of tests:

- Specific case, when the End user uses non-English language posts in his Facebook.
- Validation test, when the latest social data is processed by the Facebook App
- Correct functionality of the system, when the automatic social data analysis works as intended

Test Scenario	End user has non-English language posts in Facebook
Tests steps	1. Create non-English language post in Facebook 2. Connect to ISSUU to perform Social Data analysis 3. Log in ISSUU Facebook App to see content network
Expected result	Social Data analysis performed if language is supported by ISSUU. ISSUU Facebook app ignores non-English posts.
Actual result	Social Data analysis performed. ISSUU Facebook App includes non-English posts to the content network, without applying proper data cleaning.
Applied fixes (if any)	Issue occurred because there was no language check when processing the social data. To fix the problem the author used networkX's english.pickle language pack to verify the language of the posts.

Table 8: Functionality test: Non-English posts

Test Scenario	End user's latest Facebook posts are represented in the ISSUU Facebook App
Tests steps	1. Go to Facebook and make a post 2. Log in the ISSUU Facebook App 3. Check if the newest post is visualized
Expected result	The newest Facebook posts are included into the content network visualization
Actual result	Newest posts not included
Applied fixes (if any)	A check_for_update() function was added before visualizing the content network. If there are new posts, they are added to the content network first and then the network is visualized.

Table 9: Functionality test: Latest Social Data

Test Scenario	End user's Social data is gathered from Facebook, stored as a raw data, then it is analyzed using Network Theory methodology and it is visualized.
Tests steps	1. Connect ISSUU account to Facebook account 2. Verify that the raw data is stored 3. Verify that network theory analysis is performed 4. Verify that the analysis results are visualized
Expected result	The raw and analyzed data are saved and available in the ISSUU database. Analysis performed correctly. Visualization available through the ISSUU Fb App.
Actual result	The data is available. Analysis performed correctly.
Applied fixes (if any)	

Table 10: Functionality test: Social Data analysis

3.3. Integration

A careful planning is needed to integrate the developed features into ISSUU platform. The integration process has to ensure a successful extension of the ISSUU platform with the designed improvements.

In the present time ISSUU is not ready to integrate the developed components into their system due to already agreed roadmap development for the first quarter of 2013. The company is seeing the provided by the Master Thesis project benefits and agrees that there is a potential gain laying unexplored in the Social Media data analysis. It is possible to integrate the project with ISSUU platform in the later stage, when the company is ready to start working on its Social Media agenda.

This chapter represents a detailed roadmap for implementing the integration process.

3.3.1. Prepare the system

In order the integration to work properly the ISSUU platform needs to be prepared to inhibit the developed functionalities. What needs to be ensured before the successful integration is:

- Prepare the ISSUU database (as described in section 3.5.5 Save the Social Data)
- Choose the Facebook App hosting according to the ISSUU standards and internal regulations.
- Calibrate Social Data gathering according to ISSUU system configuration (capacity, speed and overload limitations)

3.3.2. Select user segment

To make available the project to all of the ISSUU customers from the start presents a potential danger to the company. It is wise first to select a segment of the ISSUU users and introduce them to the new functionalities, perform adjustments, gather feedback and measure success of the project. If the project fulfills its goals, and after it has been properly adjusted to match the

user expectations, then it can be introduced to the rest of the users. In order to do that a user segment needs to be selected. The segment needs to be small enough not to change the general End user view of ISSUU as service, but large enough to gather the necessary data. A user segment from 10 000 to 100 000 End users (from 0.017% to 0.17% of the total amount) is going to be sufficient for the intended purposes.

3.3.3. Making a Pilot

A pilot project is a small scale preliminary project conducted in order to evaluate feasibility, time, cost, adverse events, and effect size (statistical variability) in an attempt to predict an appropriate sample size and improve upon the study design prior to performance of a full-scale research project. Performing the Pilot will require to communicate the change to the selected user segment and make them aware of the new features. Also it is important to gather feedback from them in order to improve the system. Performing a Pilot Project Integration of the designed and developed features will give an estimate of required effort for the full-scale project integration and the value of the acquired benefits.

3.3.4. Go Live

After performing the preliminary integration and gaining the necessary amount of confidence in the success of the project, a full-scale integration is going to be launched. The gathered knowledge from the pilot will enable a faster and smoother integration process for the rest of the users. The change needs to be communicated to the ISSUU customers both through the ISSUU announcements and Social Media campaign when going live.

6. Discussion

The Discussion chapter focuses on what has been attained with the project and relates to the user needs and the initial thesis goals.

6.1. What has been accomplished

The main goals of the thesis were:

- Design and implement prototype application which is able to acquire ISSUU End User Social Media Data, perform Data Analysis over it, prepare it as basis for personalization of ISSUU content and finally visualize it in form of a network graph.
- Create an app prototype, which acquires the End User Social Data. It should be available on the web and allow the performance of first tests of implemented algorithms.
- The prototype application should be available through the Social Web and through ISSUU.
- The prototype application should work on a web browser and it should be able to cooperate with server responsible for gathering and analyzing the data.
- Create server side Social Data gathering and analyzing mechanism. The gathered Social Data should be used for expanding the ISSUU User Definition and the analyzed Social Media content should be used for personalization of the recommended ISSUU content.
- The prototype application then should be able to visualize the analyzed data as network graph.
- Test and evaluate results of the Social Data analysis and visualization.

All of the goals have been addressed by the thesis. They could be summarized in accomplishing three main objectives – personalization, data analysis and visualization. The summary of the personalization part is described in section 6.1.1, summary of the Data Analysis is described in section 6.1.2 and summary of the visualization in section 6.1.3. Section 6.2 is ending this thesis and contains a description of planned future work.

6.1.1. Summary of Personalization

The objective of personalization has been achieved through two mechanisms.

The first mechanism includes enriching the ISSUU User Definition with the gathered Facebook personal data (through the created Facebook App). Now the User Definition includes fields like Location, Languages, Work, Education, etc. And not only that, but the data acquired is more precise and valid than the previously existing one.

The second mechanism includes analyzing the Facebook content posted by the End user. Defining somebody, based on what he posts on Facebook is a useful tool and adds perspective to the how an End user is viewed.

6.1.2. Summary of Data Analysis

The objective of Data Analysis has been achieved through the network theory method. A tool is created for Data Analysis of the content of the Social Media in the form of the Facebook App. The social data is cleaned according to the performed initial data analysis as described in section 4.5.3. Then the described algorithms for finding Modularity and Degree Distribute has been applied with the python language package NetworkX used for manipulation and study of the structure and functions of the complex networks. All of the network manipulations have been achieved through NetworkX. The resulting network has been visualized in a form of the main connected component of a graph.

Further Data Analysis could be achieved when the project is integrated as discussed in Chapter 3.3.

6.1.3. Summary of Visualization

After the performed data analysis the resulting network is visualized with the help of the created Facebook App. The dynamic interaction of the visualization has been achieved through the lightweight JavaScript library Sigmajs as described in Chapter 5.1.5.

The attractive visualization addresses the goals of increasing the ISSUU user base. It is not possible to measure the success of the achievement, because the Facebook App is not a part of the ISSUU platform yet (refer to chapter 3.3 Integration).

6.2. Future Improvement

This chapter discusses the future improvement of the project, both in expanding the scale of the developed features and proposing new features that may add value to the customers.

6.2.1. Visualize timeline

All data entries have a timestamp which enables them to create a timeline. The created timeline shows the change in the End user interests. For the time being that feature is used to analyze the Social Data and the timeline have not being visualized. Visualizing the timeline could give an advantage easy way of noticing trends in the data and making general conclusions about the dataset.

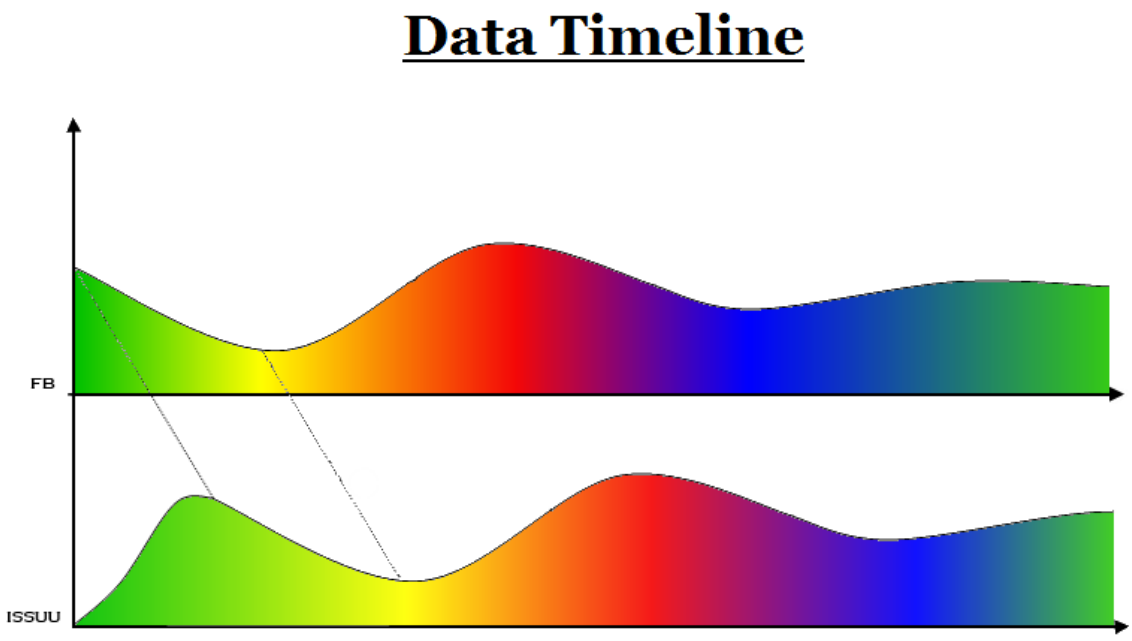


Figure 25: Example Data Timeline

The figure above is a prototype of the possible timeline visualization. The top graphic is the End user social data from Facebook and the bottom graphic is the End user readings data from ISSUU. The ordinate level represents the End user activity value in the time period and the abscissa value shows how the data is changing over time. The different colors represent the interests of the End user changing over time. In the idealized case it is possible to observe change in the End user interests first in Facebook, then in ISSUU. That enables a process of an early capturing of the End user interest and making the changes accordingly in the recommendations system.

6.2.2. Connect users

A significant amount of users read ISSUU publications without having an ISSUU account and half of the users that have an account haven't connected their Facebook to ISSUU. It is possible to improve the ISSUU service for based on the data analysis performed on the social media data. For example it is possible to establish with high probability which age group they are and which gender, based on what they are reading.

Since such a big percentage of the users anonymously are using the services, it is worth to invest time and effort into improving their user experience.

6.2.3. Publishers

Most of the developed functionalities are for the benefit of the End user. A important segment of the ISSUU customers are the publishers and it is a good idea to look further into how to increase the provided value for them through the Social Media.

6.2.4. Going Mobile

The ISSUU Facebook App is available through browser, even on mobile platforms. But it is a good idea to create an Android and/or iOS mobile app for it, increasing the options for access, thus increasing the value for the customers.

6.2.5. Facebook friends

As a possible development for the ISSUU Facebook app is to analyze the End user's Facebook friends social data (the ones having ISSUU accounts) and include it into the data analysis. That could be useful to understand how individuals influence each other and how their interests correlate.

Also it is possible to extend the app functionality into making reading competitions – who has read the most magazines, who has the most variety in topics, etc.

6.2.6. Facebook Graph Search

Facebook Graph Search[xxiii] has been introduced during writing of the Master Thesis Project and it is a very useful tool that could be used to improve the performed services. With Graph Search it is possible to look up anything shared with the End user on Facebook.

The Graph Search is still in beta version, but as soon it is released it is worth investing time into extending the developed functionalities through it.

7. References

[i] http://issuu.com/explore
[ii] http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
[iii] http://en.wikipedia.org/wiki/Collaborative_filtering
[iv] https://developers.facebook.com/docs/concepts/login/
[v] https://developers.facebook.com/docs/reference/plugins/like/
[vi] https://www.facebook.com/about/timeline
[vii] http://barabasilab.neu.edu/networksciencebook/
[viii] http://developers.facebook.com/docs/reference/api/
[ix] http://developers.facebook.com/docs/reference/api/
[x] http://en.wikipedia.org/wiki/Degree_distribution
[xi] http://en.wikipedia.org/wiki/Modularity_(networks)
[xii] http://www.axure.com/
[xiii] http://share.axure.com/DEAOBK/
[xiv] http://en.wikipedia.org/wiki/Sequence_diagram
[xv] http://developers.facebook.com/docs/reference/fql/
[xvi] http://www.heroku.com/
[xvii] https://developers.facebook.com/blog/post/558/
[xviii] http://www.heroku.com/how
[xix] http://en.wikipedia.org/wiki/Python_(programming_language)
[xx] http://networkx.github.com/
[xxi] http://gexf.net/format/
[xxii] http://sigmajs.org/
[xxiii] https://www.facebook.com/about/graphsearch

Appendix A

ISSUU Fb App Specification

Design Prototype

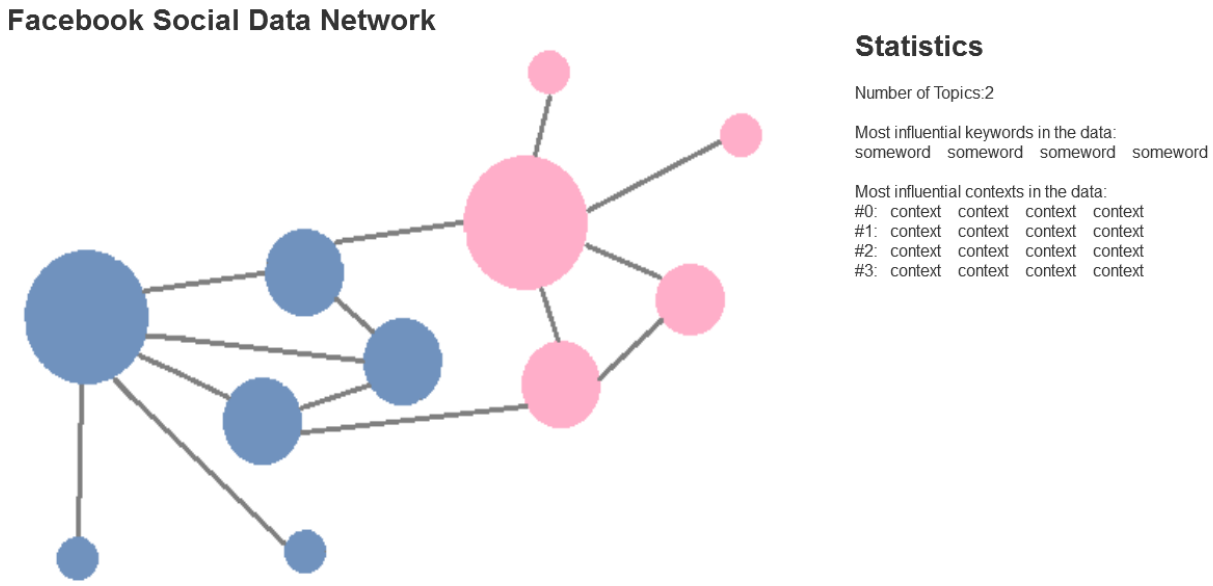
Dimitar Lesov

08-Nov-12

1.1. ISSUU Fb App

The Main Page of the Facebook App. Here are visialized the Account details, the Social Data Network and the ISSUU Data Network.

OnPageLoad:
Case 1:
Set MainArea state to State2



1.1.1. Widget Table

Footnote	Label	Interactions
1		OnClick: Case 1: Set MainArea state to State1
2		OnClick: Case 1: Set MainArea state to State2
3		OnClick: Case 1: Set MainArea state to State3

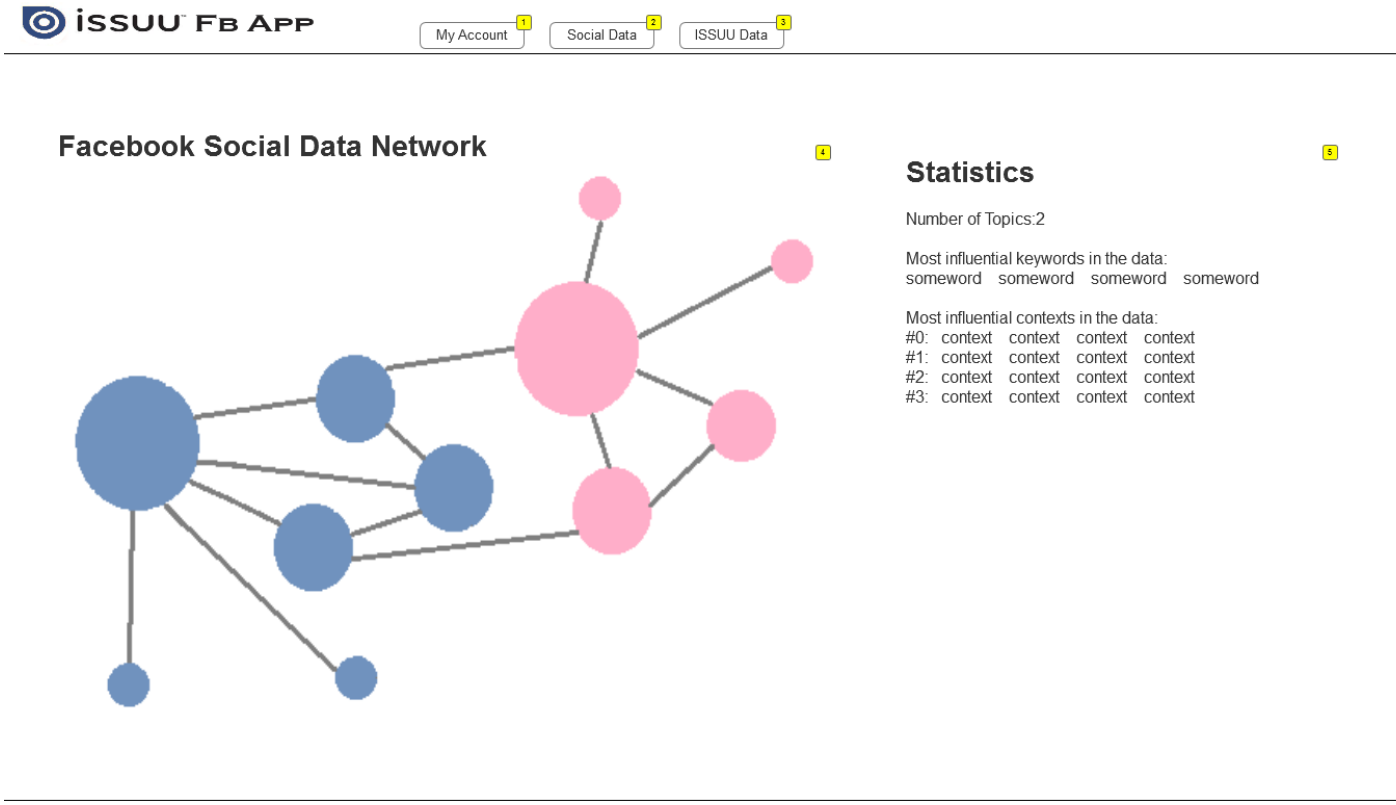
1.1.2. MainArea

1.1.2.1. User Account

1.1.2.2. Widget Table

Footnote	Label	Interactions	Description
1		OnClick: Case 1: Set MainArea state to State1	End user details. When pressed displays the account details - End user Facebook profile picture and the End user ISSUU User Definition details.
2		OnClick: Case 1: Set MainArea state to State2	The Facebook Data visialization of the End user - posts, likes, etc. (all described ing chapter3.5.1 Data Selection)
3		OnClick: Case 1: Set MainArea state to State3	ISSUU Data visualization of the End user ISSUU publications readings.
4			The End user Facebook Profile Picture.
5	text		The End user details gathered from the ISSUU User Definition.

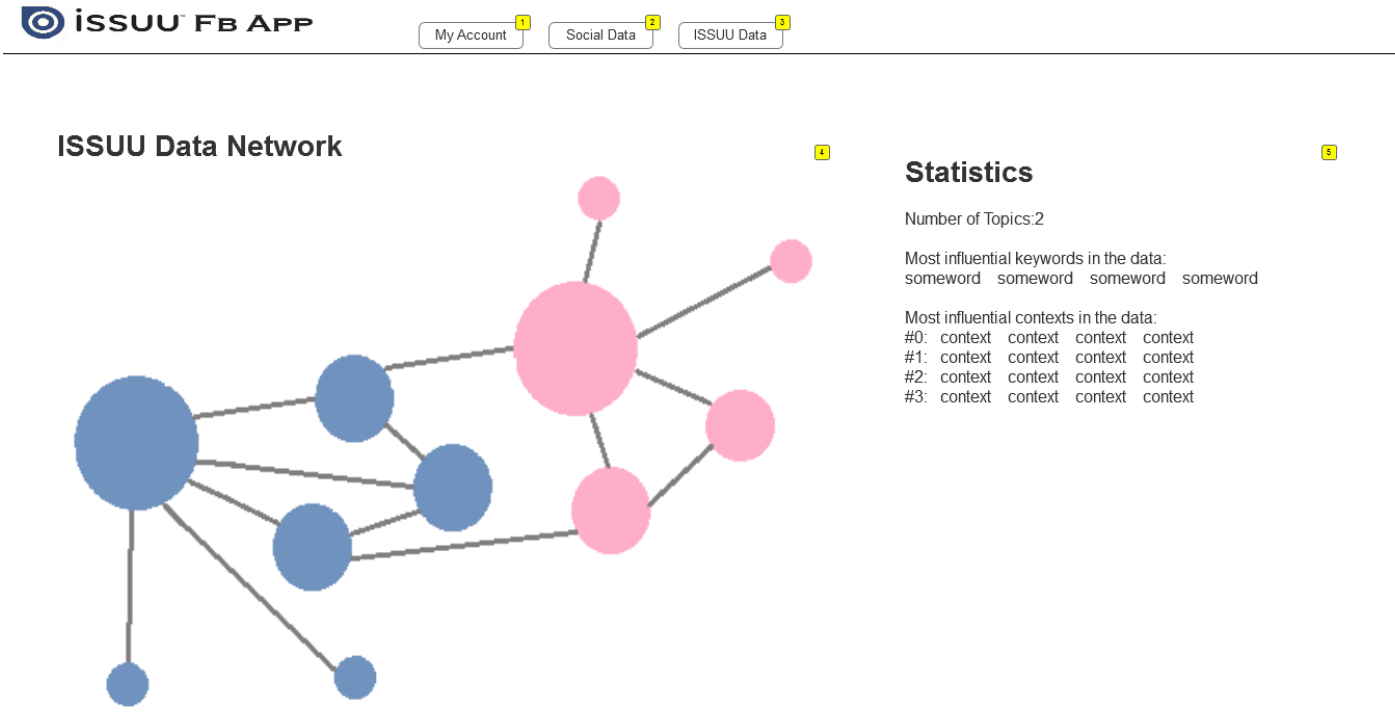
1.1.2.3. Social Data



1.1.2.4. Widget Table

Footnote	Label	Interactions	Description
1		OnClick: Case 1: Set MainArea state to State1	
2		OnClick: Case 1: Set MainArea state to State2	
3		OnClick: Case 1: Set MainArea state to State3	
4			A visualized Facebook content network with Modularity and Degree distribution.
5	text		Provided statistics for the visualized network.

1.1.2.5. ISSUU Data



1.1.2.6. Widget Table

Footnote	Label	Interactions	Description
1		OnClick: Case 1: Set MainArea state to State1	
2		OnClick: Case 1: Set MainArea state to State2	
3		OnClick: Case 1: Set MainArea state to State3	
4			A visualized ISSUU read publications content network with Modularity and Degree distribution.
5	text		Provided statistics for the visualized network.

Appendix B

Approach

In order to have flexibility and increase efficiency with the project deliverables the author decided on using an agile approach for the Master Thesis Project. Below is description of the adopted project format, timeframes and deliverables of the project.

Development Framework

The author using the gained knowledge and experience from various DTU courses took into consideration the needed activities to achieve the Master Thesis goals and decided on SCRUM[xxiii] software development framework as the most efficient approach for the project. Like other agile development methodologies, Scrum can be implemented through a wide range of tools. Many companies use universal tools, such as spreadsheets to build and maintain artifacts such as the sprint backlog. There are also open-source and proprietary packages dedicated to management of products under the Scrum process. For this project the author has selected Microsoft Project Professional 2013 [xxiii] which includes the needed project planning functionalities and completely suffices the project planning requirements.

During the project span the author has adopted the following Roles: Product Owner, Designer, Developer and Tester.

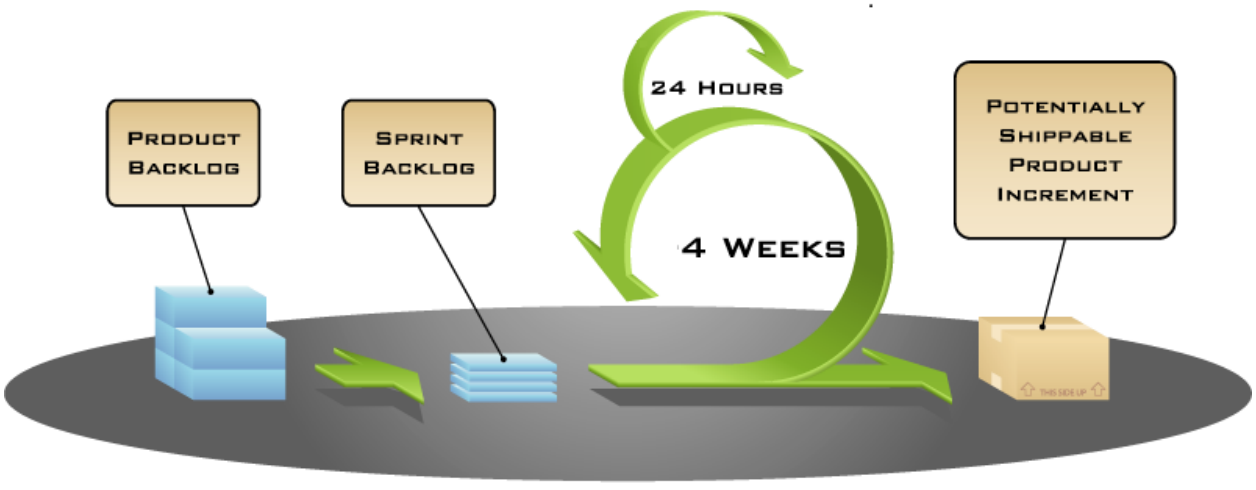


Figure 1: SCRUM Process

Product Backlog

The product backlog is an ordered list of requirements that is maintained for a product. It contains Product Backlog Items that are ordered by the Product Owner based on considerations like risk, business value, dependencies, date needed, etc. The features added to the backlog are commonly written in Story format. The product backlog is the "What" that will be built, sorted in the relative order it should be built in. The product backlog contains rough estimates of both business value and development effort. Those estimates help the Product Owner to gauge the timeline and may influence ordering of backlog items.

Sprint Backlog

The sprint backlog is the list of work for development that must be addressed during the next sprint. The list is derived by selecting stories/features from the top of the product backlog until the Developer feels it has enough work to fill the sprint. The Developer should keep in mind the velocity of its previous Sprints when selecting stories/features for the new sprint, and use this number as a guide line of how much "effort" he can complete.

The stories/features are broken down into tasks by the Developer, which, as a best practice, should normally be between four and sixteen hours of work. With this level of detail the Developer understands exactly what to do. Often an accompanying task board is used to see and change the state of the tasks of the current sprint, like "to do", "in progress" and "done".

Sprint

A sprint is the basic unit of development in Scrum. The sprint is a "timeboxed" effort, i.e. it is restricted to a specific duration. The duration is fixed in advance for each sprint and is normally between one week and one month. For the Master Thesis project the author has selected the 4 week span of the Sprint.

Each sprint is preceded by a planning, where the tasks for the sprint are identified and an estimated commitment for the sprint goal is made, and followed by a review, where the progress is reviewed and lessons for the next sprint are identified.

During each sprint, a finished portion of the product is created. The set of features that go into a sprint come from the product backlog. Which backlog items go into the sprint is determined during the sprint planning. The sprint goals should not be changed during the sprint. Development is timeboxed such that the sprint must end on time; if requirements are not completed for any reason they are left out and returned to the product backlog.

A key principle of Scrum is its recognition that during a project the customers can change their minds about what they want and need, and that unpredicted challenges cannot be easily addressed in a traditional predictive or planned manner. As such, Scrum adopts an empirical approach—accepting that the problem cannot be fully understood or defined, focusing instead on maximizing the ability to deliver quickly and respond to emerging requirements.

Project Plan

Using the described methodology the author divided the thesis into six Sprints with four weeks duration each. By Project Plan completion of the Master Thesis will take 6 months.

In the presented below project plan the detail level is limited to Work Items. This is done due to the fact that all Work Items are broken down into tasks and subtasks and listing them all is going to take up a big portion of the report.

Sprint 1 – Requirements

The first sprint will take 4.6 weeks – from 01.10.2012 to 31.10.2012. Below are displayed the Sprint Time table and the Sprint Gantt chart.

	Item Type	Name	Actual Work	Start	Finish
1	Sprint	Sprint 1	81 hrs	Mon 10/1/12	Wed 10/31/12
2	Work Item	Define Goals of the Thesis	15 hrs	Mon 10/1/12	Thu 10/4/12
3	Work Item	Define Scope of the Thesis	12 hrs	Thu 10/4/12	Sun 10/7/12
4	Work Item	Collaborate with ISSUU	24 hrs	Sun 10/7/12	Wed 10/31/12
5	Work Item	Gather Requirements	30 hrs	Sun 10/21/12	Wed 10/31/12

Figure 2 Error! Main Document Only.: **Sprint 1 Time table**

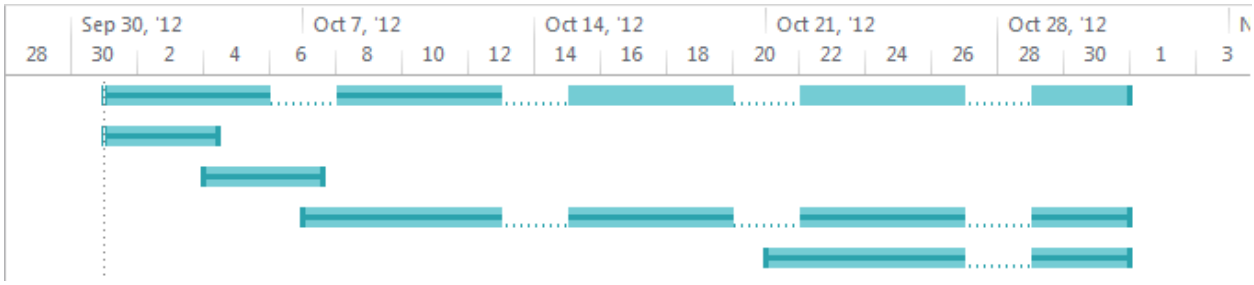


Figure 3: Scrum 1 Gantt chart

The Sprint includes the following Work Items:

- Define Goals of the Thesis: Find out the general direction of the thesis and what should be accomplished by the project. Think of ways to generate more value and improve ISSUU services.
- Define Scope of the Thesis: Decide upon the coverage of the project, what is feasible and what is possible to achieve in the given timeframe.
- Collaborate with ISSUU: Make contact with ISSUU and discuss goals and scope of the project. Understand user needs and gather requirements. Discuss roles, involvement and resources from both sides. Agree on timeframes and project schedule. Make commitments.

- Gather Requirements: Create a requirements list pursuant with the user needs, timeframes and workload amount.

Sprint 2 – First Prototype

The second sprint will take 4.4 weeks – from 01.11.2012 to 30.11.2012. Below are displayed the Sprint Time table and the Sprint Gantt chart.

	Item Type	Name	Actual Work	Start	Finish
6	Sprint	Sprint 2	176 hrs	Thu 11/1/12	Wed 11/28/12
7	Work Item	Design Phase	84 hrs	Thu 11/1/12	Tue 11/13/12
8	Work Item	Produce Product Backlog	57 hrs	Fri 11/16/12	Wed 11/28/12

Figure 4: Scrum 2 Time table

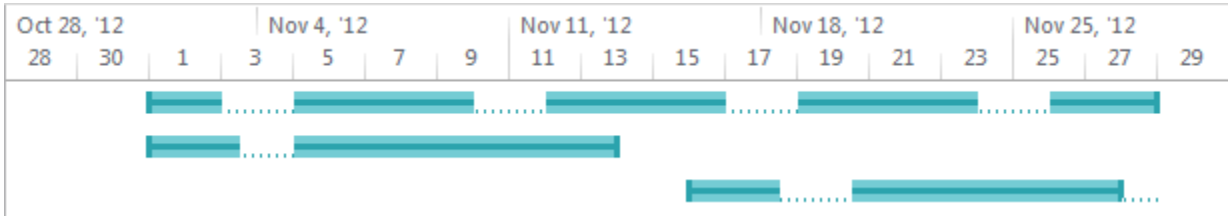


Figure 5: Sprint 2 Gantt chart

The Sprint includes the following Work Items:

- Design phase: After the purpose and requirements are determined a plan for a solution is developed. It includes low-level component selection and design as well as the architectural view.
- Produce Product Backlog: Generate Stories and add features to the Product Backlog. Segmentation of Stories into Work Items and Work Items into Tasks.

Sprint 3 – Implementation

The third sprint will take 4.4 weeks – from 01.12.2012 to 31.12.2012. Below are displayed the Sprint Time table and the Sprint Gantt chart.

	Item Type	Name	Actual Work	Start	Finish
9	Sprint	Sprint 3	268 hrs	Sat 12/1/12	Mon 12/31/12
10	Work Item	Decide what data to use	18 hrs	Sat 12/1/12	Wed 12/5/12
11	Work Item	How to take the data	35 hrs	Tue 12/4/12	Sat 12/8/12
12	Work Item	Get data from Social Media	66 hrs	Mon 12/10/12	Thu 12/20/12
13	Work Item	First Prototype	149 hrs	Wed 12/19/12	Mon 12/31/12

Figure 6: Sprint 3 Time table

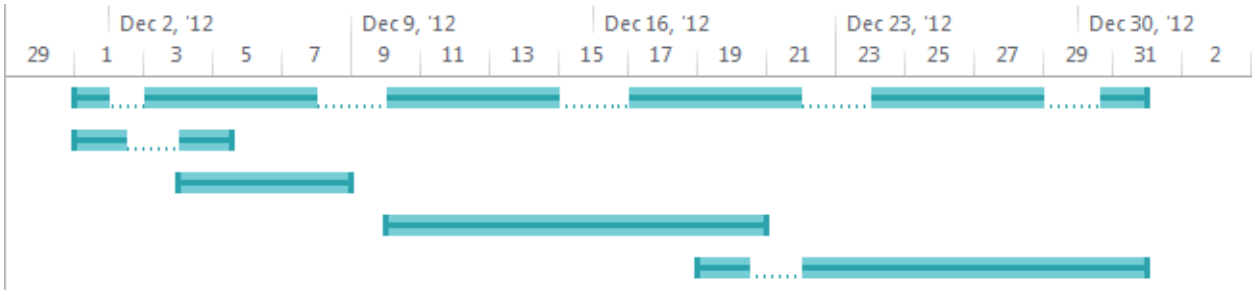


Figure 7: Sprint 3 Gantt chart

The Sprint includes the following Work Items:

- Decide what data to use: Select what social data to use. Not all available data is relevant and a filtering must be applied based on different criteria.
- How to take the data: Define the mechanisms by which the data will be taken from the Social Media site.
- Get data from Social Media: Apply the decided approach and the selected data gathering mechanisms and mine the social data.
- First Prototype: Adapt technologies and create Social Media Application as first prototype of the solution.

Sprint 4 – Implementation

The forth sprint will take 4.6 weeks – from 01.01.2013 to 31.01.2013. Below are displayed the Sprint Time table and the Sprint Gantt chart.

	Item Type	Name	Actual Work	Start	Finish
14	Sprint	Sprint 4	178 hrs	Tue 1/1/13	Thu 1/31/13
15	Work Item	Content Network	29 hrs	Tue 1/1/13	Sat 1/5/13
16	Work Item	Modularity of Content Netwo	36 hrs	Sun 1/6/13	Sat 1/12/13
17	Work Item	Degree Distribution of Conte	24 hrs	Sun 1/13/13	Thu 1/17/13
18	Work Item	Tag Cloud	25 hrs	Sat 1/19/13	Wed 1/23/13
19	Work Item	Second Prototype	65 hrs	Mon 1/21/13	Thu 1/31/13

Figure 8: Sprint 4 Time table

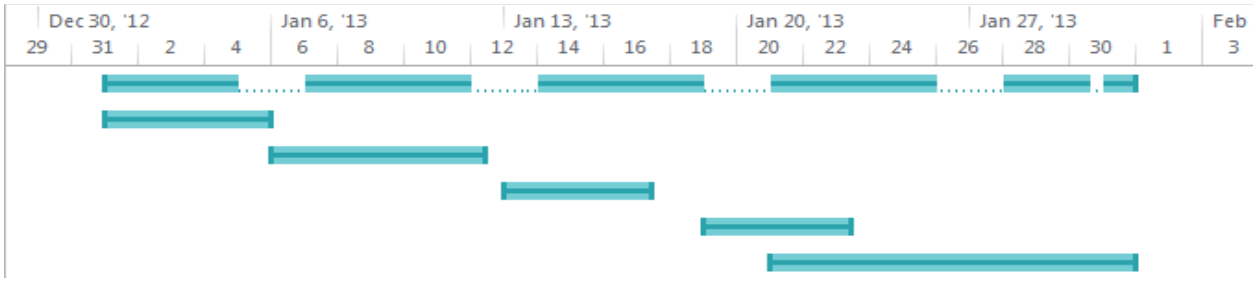


Figure 9: Sprint 4 Gantt chart

The Sprint includes the following Work Items:

- Content Network: Create a content network based on the gathered social data. Use ISSUU LDA practices for the content. Create visualization of the network in the Social Media Application.
- Modularity of Content Network: Create mechanism to run Modularity algorithm over the content network in order to discover content topics. Visualize the Modularity of the Network in the Social Media Application.
- Degree Distribution [xxiii]of Content Network: Create mechanism to run Degree Distribution algorithm over the content network in order to calculate the nodes' importance in the network. Visualize the Degree Distribution of the Network in the Social Media Application.
- Second Prototype: Combine the developed features of the Social Media Application into the second prototype of the solution.

Sprint 5 – Integration

The fifth sprint will take 4 weeks – from 01.02.2013 to 28.02.2013. Below are displayed the Sprint Time table and the Sprint Gantt chart.

	Item Type	Name	Actual Work	Start	Finish
20	Sprint	Sprint 5	159 hrs	Fri 2/1/13	Thu 2/28/13
21	Work Item	Deploy Solution to ISSUU	37 hrs	Fri 2/1/13	Wed 2/6/13
22	Work Item	Analyze gathered Data	90 hrs	Fri 2/8/13	Fri 2/22/13
23	Work Item	Calibrate	48 hrs	Thu 2/21/13	Thu 2/28/13

Figure 10: Sprint 5 Time table

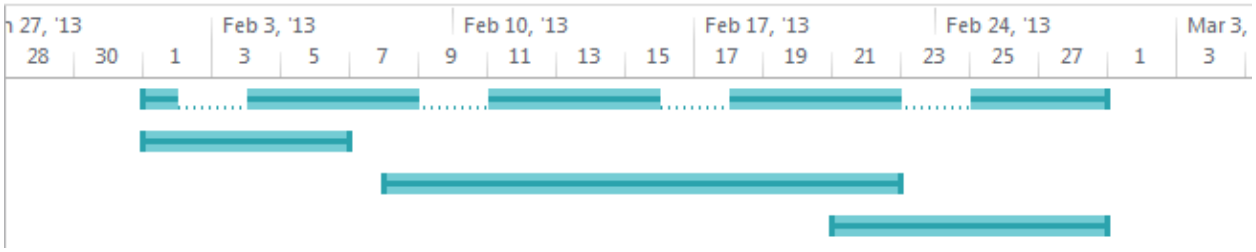


Figure 11: Sprint 5 Gantt chart

The Sprint includes the following Work Items:

- Deploy Solution to ISSUU: Collaborate with issuu and deploy the created prototype on the ISSUU platform.

- Analyze gathered Data: Perform analysis on how much of the performed Data Analysis on the End user's Social Data is useful and discover trends and shortcomings. Compare the real findings to the expected findings.
- Calibrate: Calibrate the performed Social Data Analysis in needed. Maximize the generated value from the produced prototype.

Sprint 6 – Report

The sixth and last sprint will take 4.2 weeks – from 01.03.2013 to 29.03.2013. Below are displayed the Sprint Time table and the Sprint Gantt chart.

	Item Type	Name	Actual Work	Start	Finish
24	Sprint	Sprint 6	180 hrs	Fri 3/1/13	Sun 3/31/13
25	Work Item	Generate Report Structure	15 hrs	Sat 3/2/13	Sun 3/3/13
26	Work Item	Apply Template Formats	16 hrs	Tue 3/5/13	Wed 3/6/13
27	Work Item	Create Visuals	31 hrs	Thu 3/7/13	Mon 3/11/13
28	Work Item	Describe Project Processes	98 hrs	Sun 3/10/13	Mon 3/25/13
29	Work Item	Review Report	12 hrs	Tue 3/26/13	Wed 3/27/13

Figure 12: Sprint 6 Time table

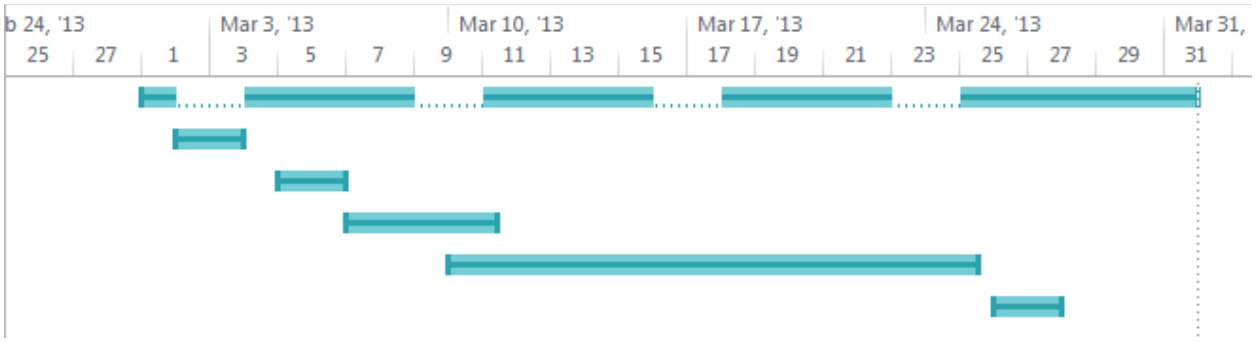


Figure 13: Sprint 6 Gantt chart

The Sprint includes the following Work Items:

- Generate Report Structure: Create the structure of the report.
- Apply Template Formats: Apply the approved DTU template formats.
- Create Visuals: Create visuals for the report.
- Describe Project Processes: Write down the applied processes and experiences.
- Review Report: Final review before hand-in.