AFEL

Analytics for Everyday Learning

EC Project 687916

D1.3 – Feature definitions and extraction methods

Deliverable Coordinator: Ilire Hasani-Mavriqi (KNOW)

Contributors: Ilire Hasani-Mavriqi, Dominik Kowald, Elisabeth Lex (KNOW), Alessandro Adamou, Mathieu d'Aquin (OU), Susana López Sola, Ricardo Alonso Maturana, Esteban Sota (GNOSS), Ujwal Gadiraju (LUH)

Reviewers: Peter Holtz (IWM), Besnik Fetahu (LUH)

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Change Log

Version	Date	Amended By	Comment
0.1	14/11/2016	Ilire Hasani-Mavriqi	Outline and stub
0.2	18/11/2016	Ilire Hasani-Mavriqi	First draft
0.3	15/12/2016	Ilire Hasani-Mavriqi, Dominik Kowald	Consolidated structure and first use case, incorporated first feedback from OU
0.4	19/01/2017	Ilire Hasani-Mavriqi, Dominik Kowald	Incorporated feedback from OU and GNOSS
0.5	30/01/2017	Elisabeth Lex	Sanity check and feedback
1.0	31/01/2017	Ilire Hasani-Mavriqi, Dominik Kowald, Elisabeth Lex, Alessandro Adamou, Susana López Sola, Esteban Sota	First comprehensive draft
2.0	10/02/2017	Ilire Hasani-Mavriqi, Dominik Kowald, Alessandro Adamou	Second draft
2.5	13/02/2017	Ilire Hasani-Mavriqi, Dominik Kowald, Alessandro Adamou	Second draft for internal QA
3.0	28/02/2017	Ilire Hasani-Mavriqi, Dominik Kowald, Elisabeth Lex, Susana López Sola, Alessandro Adamo, Mathieu d'Aquin	Final version

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Executive Summary

This document provides an outline on methods that are necessary for defining and extracting features that are relevant of learning activities in data sources captured within the AFEL project.

We contribute with a methodology on how to tackle the problem of feature definition and extraction in everyday learning settings. We tackle this problem from a use case perspective. Thus, for selected data sources identified in D1.1 and GNOSS-Didactalia data sources identified in D5.1, we construct use cases by describing characteristics of a particular dataset, such as, the type and format of the data provided within the dataset. Apart from that, we provide methods for using and combining this data by defining features corresponding to learning activities identified in WP4. Finally, we provide approaches to extract the defined features from the raw data. These approaches include state-of-the-art clustering, natural language processing and data re-factoring methods.

This document should serve as an initial step for the data enrichment procedure of the AFEL project and therefore, combines outcomes from different work packages. Specifically, it applies the results achieved in WP1 and WP4 as valuable inputs. The outcomes of this deliverable will serve as inputs for WP2, WP3 and WP5.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Table of Contents

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Introduction

In the earlier deliverable D1.1 [AFGYL+16], the AFEL project drew an initial set of guidelines for the identification of data sources that are candidates for providing data of relevance to online learning activities. This resulted in a taxonomy of data source categories based on their assumed relevance for the project and a specification of actual data sources that instantiate these categories. Because D1.1 was just a preliminary study, the classification effort was mainly grounded on the combination of existing learning activity models with the arising requirements from other work packages in the project, including the use cases of WP5. Prior experiences of the consortium in the interaction with some of these data sources, as well as earlier pilot studies, also contributed to this endeavour of classifying data sources from the standpoint of actors in knowledge transfer (e.g., learners or teachers). However, determining the characteristic dimensions (features), which allow learning activities to be detected and described, and consequently the attributes that instantiate them in each dataset, was beyond the scope of that initial study. It is instead the goal of the present study to propose single out an initial specification of these features and present an instantiation of them on some key data sources.

The contribution of this study to other activities in AFEL is threefold:

- To provide visual analytics tools (WP3) with the properties required for generating visualisations of learning activities.
- To guide the data enrichment activities of WP2.
- To aid the data integration activity of WP1 through an indication of the pivotal properties that should be used for aggregating input from multiple datasets and serving the result through the AFEL Data Platform API[1].

In this deliverable, we first analyse the captured data sources presented in D1.1 and GNOSS-Didactalia data sources introduced in D5.1 [ALSGD+16] and describe features that are related to learning activities. We define such features based on two key entities: (i) learners and (ii) digital artifacts that are evident in the context of everyday learning in terms of the co-evolution of learning and knowledge construction [HKYCD+16]. Naturally, the interactions between key entities play also a very important role on the process of feature definition. We apply the AFEL data source taxonomy defined in D1.1 to classify features in relation to users (learners), resources (digital artifacts), and groups or communities.

The establishment of feature definitions and extraction methods is conducted in form of use cases. Thus, for selected data sources from D1.1 and GNOSS-Didactalia we construct use cases by, first, describing characteristics of a particular dataset, such as, what kind of data and in which format is provided within a dataset. Second, we provide methods on how this data

[1] AFEL Data Platform, http://data.afel-project.eu

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

can be used and combined to define features corresponding to learning activities that are evident in the dataset. Third, we provide approaches to extract the defined features from the raw data. These approaches include state-of-the-art clustering [EB11], [FHW16], [JMF99] and natural language processing [MSBFB+14] methods.

We contribute methodologically with this deliverable by providing a guidance for feature specifications in informal learning settings.

After a short introduction to the topic of feature extraction from online resources, we analyse and discuss a number of use cases, to finally provide a feature specification that can be applied in general settings.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Feature Engineering

In this section, we provide some background information on the key concepts used within the AFEL project that are already outlined in D1.1, D1.2 and D4.1, to better relate to the work presented throughout this deliverable.

A formal definition of the 'feature' concept is given by Bishop, who defines a feature as "*an individual measurable property of a phenomenon being observed*" [BIS06].

Everyday learning often takes place in a networked environment [TLLP13], thus, it is of interest to identify features that are typical for such environments. Authors of the work presented in [MMGDB07] provide an overview of social network features and measurements in general. They conducted the first large-scale measurement study and analysis of the structure of multiple online social networks. Results show that there are five basic characteristics that differentiate a social network from a regular website [DUB09]: *user-based*, *interactive*, *community-driven*, *relationships* and *emotion over content*.

While defining and specifying features that are characteristic for everyday learning, we have in mind the co-evolution model of learning and knowledge construction defined within the AFEL project, which consolidates the theoretical basis of the project [CK08] [HKYCD+16]. The co-evolution model of learning and knowledge construction identifies the key entities within the context of everyday learning: (i) persons and (ii) (digital) artifacts [HKYCD+16]. Persons can, for example, write or edit a digital artifact, which could be characterized as an active interaction, or they can only read and consume an artifact that could be noted as a passive interaction. Persons that interact (passively or actively) with a digital artifact together form an online community. Online groups and communities are signified with persons sharing common interests, beliefs or opinions for a certain topic or issue and exchanging opinions with other community members.

These key entities and interactions between them are also main actors that guide the process of feature definition and specification in this deliverable.

Assumptions of learning activities are defined in D4.1 [HKYCD+16], whereas the AFEL Glossary [AFELG16] gives definitions of key terms such as knowledge, learning, learning activities, learning scopes, and learning trajectories.

We define and extract features that are related to learning activities based on data about: (i) users' activities, (ii) how users behave, and (iii) how they interact with each other and with contents [AFGYL+16] [AFELG16].

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Data source category	*Feature types*	*Examples*
Users		
User profile	- Topics of interest - Competences - Motivation (to learn/teach)	*Facebook* (likes, group memberships) *LinkedIn* (skillsets) *StackOverflow* (tags on posted/answered questions), *Wikipedia* (edits) *GitHub/Bitbucket* (issues, commits)
Social status indicators	- Recognition of competences - Influence - Job title	*StackExchange* (reputation score) *Twitter* (followers, retweet stats) *GitHub* (accepted pull requests) *Didactalia* (rank in games) *SlideShare* (view/share stats) *Google Scholar* (h-index)
Relations to resources	- Authoring - Rating	*WikiHow* (edits) *Facebook* (reactions, comments) *Wikipedia* (edits, talk pages) *SlideShare* (shared slides) *YouTube* (posted videos, votes, comments)
Resources		
Basic metadata	- Authors - Edit history - Topics - Privacy and access control	*Wikis*, *DBLP*, *SlideShare* etc.
Social context indicators	- Complexity - Heterogeneity - General appeal - Safety-for-work - Controversiality - Author bias	*Facebook* (reactions, closed groups) *YouTube* (upvotes/downvotes, appearance of video in listings)
Popularity and authority indicators	- Shares - Citations - Ratings - Cross-references	*Twitter* (retweets of linked resource) *SlideShare* (shares, views) *CiteSeer* (citations) *Google Scholar* (citations) Authority records
Cross-resource connections	- Derivatives - Subresources - Co-authorship - Topic similarity - Grouping	Hyperlinks *GitHub* (forks) MOOC modules
Groups and communities		

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Demographics	- Size - Age ranges - Language of communication	*Facebook* (groups and managed pages) *Didactalia* (communities) *LinkedIn* (communities)
Indicators of group heterogeneity	- Education of members - Origin of members - Consumed resources - Activities	*Facebook* (individual info of group members) *LinkedIn* (skillsets of individual community members)
Network-related metrics	- Affiliated groups - Collaborations	*Reddit* *Didactalia* *StackExchange* (across QA sites)
Network dynamics	- Inter-group communication	*Facebook* (groups and managed pages)
Activity records		
Contributions	- Initialization - Editing - Curating revisions	*GitHub* (pull requests made/moderated)
Expressions of user needs	- Questions - Skills in training	*StackExchange* *Quora*
Browsing	- Sites - Specific pages - Login status - Intentionality	Browser storage (local/sync) Cookies Redirects
Consumption	- Visits - Read/Playthrough - Local downloads	*SlideShare* (views) *YouTube* (views)
Communication	- Shares - Comments	*Facebook* (shares) *Pinterest* (pins) *Twitter* (retweets)

Table 1. Aggregated features based on data identified within AFEL [AFGYL+16] [AFELG16].

Table 1 serves as a guidance while defining and extracting features in our use cases, but the categorization depicted in this table may not necessarily be relevant for each use case.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Related Approaches

In this section, we provide an overview of the state of the art approaches with regard to the selected topics that are relevant for the feature extraction methods.

Similarity Measures

Here we provide a short overview of similarity measures used to determine the similarity between entities (i.e., learners or digital artifacts). One possible application scenario would be to find most similar learners to a particular learner, which is a prerequisite for the well-known Collaborative Filtering recommendation algorithm. In practice, measuring the similarity between two vectors can be used to determine similar entities for a given entity.

Similarity Between Vectors

One often-used measure for determining similarity scores between two non-zero vectors is the Cosine similarity score, which is given by:

$$sim(A,B) = \frac{A * B}{||A|| * ||B||} = \frac{\sum Ai * Bi}{\sqrt{\sum Ai * Ai}\sqrt{\sum Bi * Bi}}$$

Here, A and B are two vectors and Ai / Bi are components of these vectors. For example, these could be learning resources with whom the users have interacted with.

The measure returns a value between -1 and 1, where -1 means exactly the opposite, 1 means exactly the same and 0 means no correlation at all.

Classification Algorithms

In Machine Learning, classification is the problem of assigning a category (from a set of categories) to an entity (i.e., a learner or a digital artifact). In the field of unsupervised learning the classification of patterns into groups or categories is known as clustering [JMF99]. One of the tools that can be applied in this context is provided within the free Weka 3 framework[2].

In the course of the AFEL project, we build on classification algorithms in order to assign users (learners) or resources (digital artifacts) to categories. One example would be to identify “admin”-like users in collaborative editing environments such as Wikipedia in order to distinguish these “caretakers” from traditional Wikipedians. A similar approach can be

[2] http://www.cs.waikato.ac.nz/ml/weka/

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

applied in StackExchange, for example, to distinguish between users that provide helpful solutions to certain problems (act as tutors) and users that provide only minor comments.

Community Detection

In informal learning environments, individuals form groups or communities that are characterized with persons sharing common interests, beliefs or opinions for a certain topic or issue.

In the graph theory and network science, communities in a network are characterized as (overlapping) sets of nodes, which are densely connected internally [Luh95].

In some use cases, group or community membership of the user is made explicit (i.e., user registration to a particular community). If this is not the case there are algorithms and methods to detect potential communities in a network [F10]. One of the widely used algorithms is the stochastic block-model algorithm [KN11]. One advantage of this approach is that it attempts to find a block partition without the need to specify the partition size in advance.

Graph and Network Analysis

There are tools available that enable efficient graph / network processing and analysis. Among others, the graph-tool[3] is a very efficient tool implemented in the programming language Python.

After constructing the collaboration networks from particular datasets based on user interactions, these networks can be stored in the graph-tool binary format (.gt) for further processing. Furthermore, extracted features can be stored as node or edge properties for later analysis.

Graph-tool is also very powerful on the following characteristics: pickling, graph statistics (degree/property histogram, vertex correlations, average shortest distance, etc.), centrality measures, standard topological algorithms (isomorphism, minimum spanning tree, connected components, dominator tree, maximum flow, etc.), generation of random graphs with arbitrary degrees and correlations, detection of modules and communities via statistical inference, and much more.

Additionally, it provides a very sophisticated graph visualisation by using a variety of algorithms and output formats.

[3] https://graph-tool.skewed.de/

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Use Cases

This section addresses a series of use cases identified within the AFEL project, both at a macroscopic level as identified by the output of WP5, and as detailed use cases within common practices of online activities. The use cases (cf. Figure 1) are then matched with some of the data sources identified in D1.1 in order to provide a by-example methodological insight into the feature extraction processes. To that end, at least one instance per use case is discussed in greater detail.

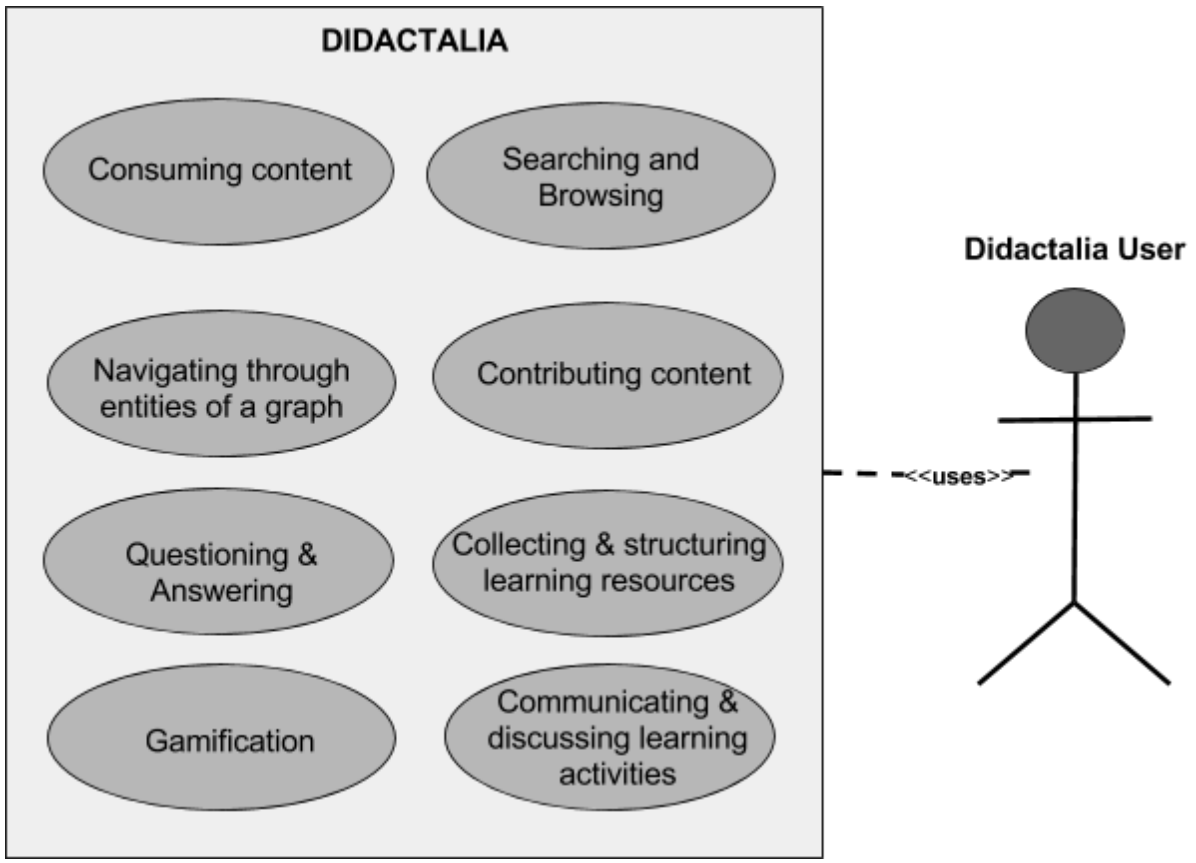

Figure 1. Didactalia use case.

Questioning and Answering

Individuals turn to question & answering (Q&A) sites mostly to seek online help while trying to find solutions to certain problems but also to provide answers to difficult questions. Such platforms enable more experienced users to play a role of a tutor at a certain point in time for specific topics. The Q&A use case is identified as an important use case within Didactalia, since users of Didactalia are provided with a Q&A tool. Thus, discussing how features can be extracted from data sources belonging to this use case is of importance for the AFEL project.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

In the following, StackExchange is discussed in more details as one of the examples of Q&A sites.

StackExchange

Informal learning usually occurs on Q&A sites while seeking for help online and interacting with our peers. So, we define and discuss certain problems online, exchange opinions and try to reach some kind of consensus on a best provided solution. Such interactions between a help-seeker and more experienced peers are also characterized as negotiation processes in networked online environments. Negotiation processes include the involvement of learners on understanding and defining certain problems, identifying the experts in the field, and finding possible solutions. As such, negotiation processes affect the constructed knowledge accumulated in the online community as a whole.

Such online communities are most visible in Q&A portals such as StackExchange, in which users collaborate by asking questions and providing answers to particular problems, indicating that users exchange their opinions while trying to agree on the best suggested solutions.

StackExchange manages multiple Q&A sites each around a specific theme, the largest and oldest one being by far the programming-oriented StackOverflow[4]. For each site, StackExchange releases anonymized data dumps through the Internet Archive every three months[5]. These data dumps consist of XML files that aggregate data about the main artifacts of the platform, such as: *Posts*, *Users*, *Votes*, *Comments*, *PostHistory* and *PostLinks*. There is no XSD schema available, but details of the applied schema can be found in readme files included with the dumps. For instance, these provide information on how a vote category should be interpreted (e.g. upvote, downvote, rate as spam or offensive or approve edit). A post on the StackExchange documentation service also gives some more insights into the structure of XML files[6].

The StackExchange dataset was applied in one of our previous works [HGCLH16], where we investigated the role of users' social status and network structure on opinion dynamics and consensus building in collaboration networks.

For illustration purposes, we provide a listing of the core XML files and the corresponding elements and attributes.

Fields	*datatype*	*Description*

[4] StackOverflow, http://stackoverflow.com/
[5] StackExchange data dumps on The Internet Archive, https://archive.org/details/stackexchange
[6] StackExchange meta - Database schema documentation for the public data dump and Data Explorer, http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Users.xml Personal profile of a user of a site and link to StackExchange		
Id	integer	Identifier of the user, *locally unique within a single Q&A site*.
AccountId	integer	Another user identifier, presumed to be globally unique.
DisplayName	string	Customizable username.
Reputation	integer	User score weighted on their activities (weights and algorithm are given).
CreationDate	ISO 8601	Timestamp of account creation.
LastAccessDate	ISO 8601	Timestamp of latest user login.
WebsiteUrl	URL	External user page (e.g. on LinkedIn)
Location	string	Custom name of the location (can be in English or in the native language of the user or location).
AboutMe	string	Self-description of the user in natural language.
Age	integer	User's stated age in years at the time of the snapshot.
Views	integer	Number of times the user's profile has been viewed.
UpVotes | DownVotes	integer	Quantitative feedback *contributed by* the user throughout the site.
ProfileImageUrl	URL	User's avatar image
Posts.xml Content and metadata of a question or answer in its state at the time of the snapshot		
Id	integer	Identifier of the post, *locally unique within a single Q&A site*.
PostTypeId	integer	Classifies a post as e.g. a question or answer
AcceptedAnswerId	integer	If a question, Id of the post accepted by the owner as the primary answer.
Score	integer	Weighted on the feedback received (weights and algorithm are given).
ViewCount	integer	Number of times the post has been viewed.
Body	string	The content of the post.
Title	string	The title of the post.
Tags	HTML-encoded string	List of tag *names* for the post, each '<>'-enclosed.
OwnerUserId	integer	Id of the post author (can be inferred from PostHistory).
LastEditorUserId	integer	Id of the author of the last edit (can be inferred from PostHistory).
LastEditDate	ISO 8601	Timestamp of latest edit.
LastActivityDate	ISO 8601	Timestamp of latest access.
AnswerCount	integer	Number of answers at the time of the snapshot (can be inferred from Posts).
CommentCount	integer	Number of comments at the time of the snapshot (can be inferred from Comments).
Favorite	integer	Number of times the post was marked as favorite.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

PostHistory.xml Revision history of a question or answer (user-resource relations)		
Id	integer	Identifier of the revision.
PostHistoryTypeId	integer	The type of revision (editing, closing, rollback etc.)
PostId	integer	Reference to the original post.
RevisionGUID	integer	Grouping of multiple history records for a single revision.
CreationDate	ISO 8601	Timestamp of the revision.
UserId	integer	The author of the revision.
Text	string	The new value of the revision content.
PostLinks.xml Cross-resource relations established by contributors		
Id	integer	Links have unique identifiers.
CreationDate	ISO 8601	Timestamp of the link creation.
PostId	integer	Subject of the link.
RelatedPostId	integer	Object of the link.
LinkTypeId	integer	Whether the link is general relatedness or a “mark as duplicate”.
Comments.xml User comments attached to posts that do not qualify as answers or posts by their own right		
Id	integer	Identifier of the comment.
PostId	integer	Reference to the post the comment is attached to.
Score	integer	Weighted on the feedback received by the comment.
CreationDate	ISO 8601	Timestamp of the comment.
UserId	integer	The author of the comment.
Text	string	The text content of the comment.
Badges.xml Awards indicating a user’s role or influence within a site or topic		
Id	integer	Identifier of a specific award assignment.
Name	string	A semi-formal title assigned to the award.
Class	integer	An enumeration 1 to 3, 1 being the top (Gold) award
Date	ISO 8601	Timestamp of the award.
UserId	integer	The author of the comment.
TagBased	boolean	Whether the award results from the activity on a certain topic.

Table 2. Main artifacts of StackExchange data dumps and lists of all their features.

Feature definition and extraction

Based on the key entities identified within AFEL and the available data sources from StackExchange, we define features that are related to learning activities. We also show how these features can be extracted from the StackExchange dataset.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

User-based features

User related data is mostly provided within the User.xml file, but also in PostHistory.xml, Badges.xml and Comments.xml.

One of the first user related features is the **user profile** feature that provides insights into user's interests, competence, expertise, role within a community or user's motivation. The attributes *AboutMe* and *Age* of the User.xml file provide personal information of users. By applying natural language processing algorithms, such as extracting keywords and concepts from the text provided in *AboutMe* attribute we can construct a user profile. Additionally, the *Name* and *Class* attributes from the Badges.xml file provide information about the role of the user within the community. In terms of the learning activity, the user profile feature would enable to identified particular experienced users as well as experts in specific fields.

User social status features makes it possible to identify the most influential learners in online communities. StackExchange provides explicit user reputation scores that can be used as a proxy for a user's social status. Such Q&A sites have a reputation system which rewards users based on their contributions. Based on the policies of this reputation system, users get appropriate scores for giving good answers, asking good questions or for voting on questions/answers of other users. Reputation scores can be extracted from the *Reputation* attribute in the User.xml file.

The StackExchange platform does not indicate associations between users or friendship links. For that reason, we turn our attention to collaboration networks which we extract by analyzing co-posting activities of users in order to have **social ties** between them. In Q&A sites, a co-posting activity between two users refers to a scenario under which two users comment on the same post. Thus, if two users contributed in any way to a same post, they are connected via an edge in the collaboration network. This feature can be extracted from the PostHistory.xml file by processing and analysing the *PostId* and *UserId* attributes. Social ties in the collaboration network indicate that between users connected through an edge the exchange of opinions, knowledge or experiences is taking place.

Social tie features can be further enhanced by investigating the strength of the collaboration that is taking place between two users. Thus, we define the **social tie strength** feature and extract it from the PostHistory.xml file by investigating the number of posts that two users edited together. The social tie strength feature, even though it is constructed on the individual level, can be applied for identification of (overlapping) communities within a collaboration network. For example, if we visualize a network and emphasize the social tie strengths between users, we would see users (i.e., overlapping nodes) that have strong social ties with users (i.e., members) of more than one community.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

User contributions feature shows which percentage of users provided which amount of the content or contributed to which number of posts. This feature can be extracted from the collaboration network, which means that graph analysis is applied. The node degree (in/out) distribution determines the user contributions feature. For example, if the node degree distribution fits to a power law or heterogeneous distribution, it indicates that a low number of users contributes with a high number of posts or comments in StackExchange.

User relations feature or user-to-user connections determines if similar users (i.e., with regard to social status or node degree) tend to connect together or if they rather connect to dissimilar users. This can be conducted by calculating the assortativity coefficient of the node degree or reputation score distributions, which means that graph analysis should be applied. Typically, in StackExchange the assortativity coefficient of the reputation scores is negative. Negative assortativity coefficient indicates a negative correlation between reputation scores over the network edges. In other words, users with lower reputation scores are more likely to connect to users with higher reputation scores. In particular, a typical post in StackExchange has many users with low reputation scores, e.g., who post a question, and only a few or even only a single user with a high score, e.g., who answers the question. This finding is in line with the assumptions from the social status theory, which states that it is our natural predisposition to interact with people who have a high social status in our social communities. High social status users also provide high quality content. Thus, determining if low status users tend to connect to high status users, indicates that the learning rate would be higher.

To find out which users interact with what kind of resources we defined the **user-resource relations** feature. In this way we can find out, for example, on what topics is a particular user interested. This feature can be extracted from PostHistory.xml and Comments.xml files by analysing the *UserId* and *PostId* attributes.

A **user meta-knowledge** resource shows how aware are users of the available content in a Q&A site. This feature can be extracted by analysing the *PostId* and *RelatedPostId* attributes in PostLinks.xml file.

Feature	How to extract	Objective / possible learning activity indicator
User profile – interest, competence, motivation, role	User.xml – *AboutMe, Age* attributes contain user personal information Badges.xml – *Name* and *Class* attributes provide information about the role of the user	Identify experienced users, experts in the field

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

User social status	User.xml – *Reputation* attribute provides explicit reputation scores that can be used as a proxy for social status	Identify most influential users
Collaboration network, social ties Social tie strength	PostHistory.xml – co-posting activities (if two users contributed to the same post, they are connected with an edge) Number of posts that two users edited together	Collaboration between users - exchange of opinions, knowledge, experience The intensity of the collaboration
User contributions	Node degree (in/out) distribution in the collaboration network	Identify which percentage of users provide the most content
User relations (if similar users tend to connect together)	Degree assortativity (positive/negative) of the collaboration network	High degree users provide high quality content. Determine if low degree users tend to connect to high degree users (negative assortativity), indicating that the learning rate would be higher
User – resource relations	PostHistory.xml and Comments.xml	Which users interact with which kind of resources (e.g., based on topics)
User meta-knowledge	PostLinks.xml – *PostId* and *RelatedPostId* attributes	Indicator to what extent users are aware of the available content, they usually refer to a related post

Table 3. Overview of the user based feature definition, extraction and relation to learning activities in Q&A.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Resource-based features

Resource-based features are mostly constructed from the Posts.xml and PostLinks.xml files.

Resource type (e.g., question or answer) feature can be extracted from the *PostTypeId* attribute contained within the Posts.xml file. In this way we could determine the question/answer ratio within a Q&A site and identify, for example, questions with the higher number of answers.

The Posts.xml file contains *Score* and *ViewCount* attributes that provide information on the **resource popularity**. This feature enables us to identify most popular and high quality resources.

Natural language processing and clustering algorithms can be applied on the text snippets extracted from *Title* and *Tags* attributes of the Posts.xml, to obtain the **resource topic** feature.

Question-best answer relation is one of the most important resource based features. This can be extracted from the *AcceptedAnswerId* attribute, which could also be related to the user that provided the best answer. By identifying the best accepted answer or solution to a particular problem we can confirm that users collaboratively contributed to solve a particular problem. Furthermore, by finding the user that provided the best answer we can also relate this information to user's role or social status and check the correlation.

Cross-resource relations feature indicates to what extent users are aware of the available content within the online community. If they know that a similar issue was handled in the past they refer the user asking a question to the corresponding post. This feature can be extracted by analysing the *PostId* and *RelatedPostId* attributes in PostLinks.xml file. Furthermore, from this feature we can derive a **resource similarity** feature, which implies that two resources are similar based on users meta-knowledge.

Feature	**How to extract**	**Objective / possible learning activity indicator**
Resource type (question, answer)	Posts.xml – *PostTypeId* attribute	Determine question/answer ratio Identify questions with the higher number of answers
Resource popularity	Posts.xml – *Score* and *ViewCount* attributes	Identify most popular resources
Resource topic	Posts.xml – *Title* and *Tags* attributes could be analysed	Cluster resources based on topics

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Question – best answer relation	Posts.xml – *AcceptedAnswerId* attribute could also be related to the user that provided the best answer	Identify the answer that is accepted as the best one, indicating that users contributed to solve a particular problem. Match the user that provided the best answer and correlate with user's role or social status
Cross-resource relations	PostLinks.xml – *PostId* and *RelatedPostId* attributes	Indicator to what extent users are aware of the available content, they usually refer to a related post
Resource similarity		Determine the similarity between resources based on users meta-knowledge

Table 4. Overview of the resource based feature definition, extraction and relation to learning activities in Q&A.

Group / community-based features

It is possible to cluster users based on personal information (e.g., by extracting keywords from the *AboutMe* attribute in Users.xml) or based on resources they consumed (e.g. *Title* attribute in Posts.xml) and assign them **community labels**. Another approach to detect communities is based on graph analysis by applying **stochastic block model** algorithm. **Modularity score** is a feature that gives a measure of strength of the community structure in a network. A high modularity score indicates the existence of strong communities in the network, while a low modularity score means that the community structure is not that strong.

Feature	**How to extract**	**Objective / possible learning activity indicator**
Community label	By processing personal information of users or analysing the type of resources they interact with	Detect how learners in a network are grouped, which are the largest communities or how communities overlap (based on users' data or topics of interest)
Community structure -	Apply block-model	Obtain the community

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

block model	algorithm from the graph theory on a constructed network	structure of the learners' network
Modularity score	Apply Newman modularity algorithm from the graph theory on a constructed network	Indicates the existence of strong/weak communities in the network

Table 5. Overview of the group / community based feature definition, extraction and relation to learning activities in Q&A.

Examples of feature correlations

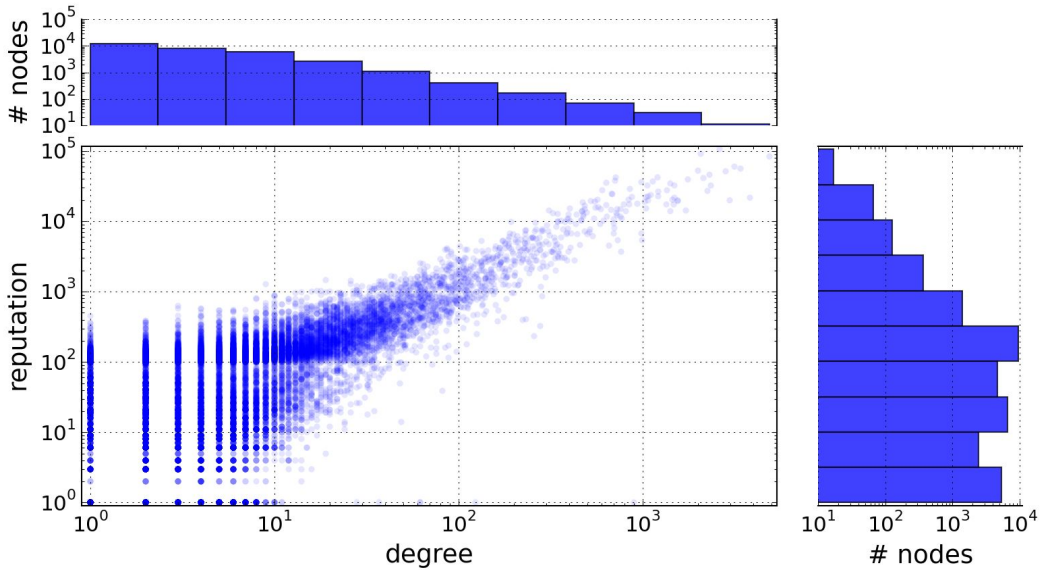

Figure 2: Distribution of reputation scores and node degrees. The plot on the right shows the heterogenous distribution of reputation scores in the StackExchange English network. The plot on the top presents the heterogenous distribution of node degrees. In the middle, the scatter plot of reputation scores vs. node degrees is shown. The Pearson correlation coefficient between the degree and the reputation score is 0.88. All other StackExchange datasets have comparable distributions and correlation coefficients.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

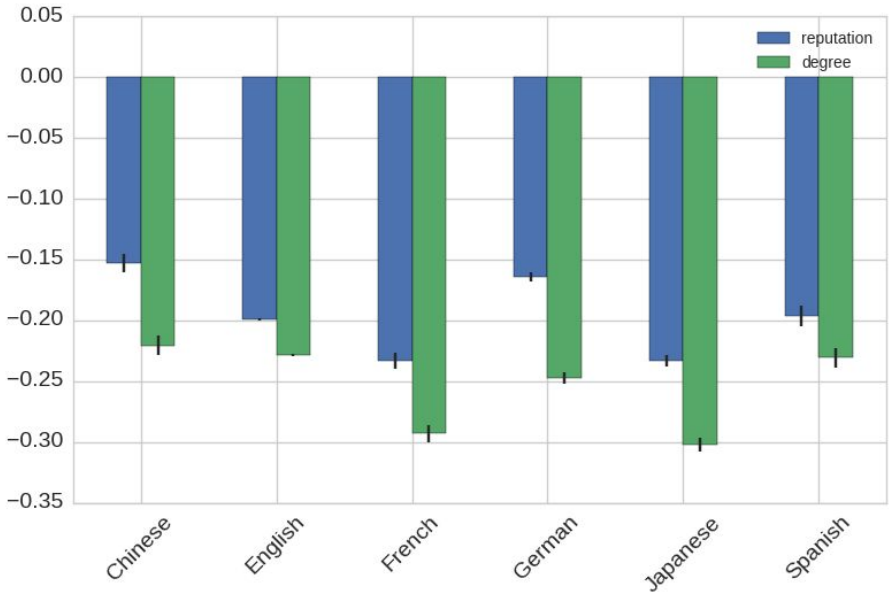

Figure 3. Reputation and degree assortativity coefficients of StackExchange datasets.

Collaborative Editing

Collaborative editing is writing done by more than one person, which typically involves various discussions among the collaborators. Since wikis are the most common tools for collaborative editing, we focus on Wikipedia in this section. By collaboratively editing a digital artifact, the contributors create knowledge whenever they overcome information that 'irritates' the common knowledge of the group in a productive way ('productive friction'). The structurally coupled epiphenomenon of knowledge creation is learning - that is the increasing ability to successfully integrate information from the environment into a learner's cognitive system [HKYCD+16].

Wikipedia

Wikipedia is a wiki-based encyclopedia that enables its users to collaboratively create and edit content. Wikipedia follows a very general goal of creating and managing the largest repository of human knowledge to date. It constitutes the Web's largest general reference work and has become one of the ten most popular pages on the Web. Additionally, the Wikimedia Foundation, which hosts Wikipedia, provides complete database dumps that can be freely used. Thus, Wikipedia has become of high interest for researchers trying to understand the underlying processes and dynamics of this socio-technical system. Furthermore, Wikipedia can be seen as a "living laboratory" to study the collaboration between users and within communities [SCCP09].

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Although there are no direct connections between users in Wikipedia (e.g., friendship or follower relationships), it is possible to create collaboration networks on basis of the Wikipedia edit logs. Therefore, we treat nodes as users and an edge between two nodes is introduced if the two corresponding users have contributed to at least one common article. In order to focus on "real" contributions, we ignore edits to the "*MainPage*" and to special pages (i.e., only content pages are considered). Furthermore, we focus on interactions, in which at least *100* words have been added. Furthermore, we also remove all bots, anonymous users and admins from the data. However, we recognize that there are still many users available that act in a similar way as admins. To be specific, these users perform a lot of edits to many different articles very frequently. In order to identify these users, we follow a classification method as described in Section Classification Algorithms.

Feature definition and extraction

To this end, we use complete dumps[7] of four Wikipedia language editions of different sizes: (i) *Danish* with *248,998* users, *209,364* articles and *8,619,237* edits, (ii) *Indonesian* with *727,668* users, *365,538* articles and *11,107,563* edits, (iii) *Finnish* with *296,094* users, *378,333* articles and *16,025,301* edits and (iv) *Norwegian* with *338,022* users, *418,518* articles and *14,304,527* edits.

Such a Wikipedia edit log is represented in XML format containing all revision contents and metadata of each available article. In order to efficiently extract the relevant attributes of the revisions, we use the *WikiEvent* Java library provided by the University of Konstanz[8]. This gives us access to revision attributes such as the title of the article, the timestamp of the revision in milliseconds, the type of the edit (i.e.,*added*, *deleted*, *restored* or *undeleted*), the number of words affected by the edit, and the username or IP address of the active user.

Hence, the following attributes are available in the Wikipedia dataset for each revision:

- **PageTitle**: title of the article
- **Time**: timestamp in milliseconds
- **InteractionType**: type of the edit (i.e., *added*, *deleted*, *restored* or *undeleted*)
- **WordCount**: number of words affected by the edit
- **ActiveUser**: username for logged-in or IP address for anonymous users
- **Target**: article title for edits of type *added* or the active user otherwise

[7] https://dumps.wikimedia.org/backup-index.html
[8] http://algo.uni-konstanz.de/software/wikievent/

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

User-based features

This enables us to gather features that account for the contribution of users as described in Table 6.

Feature	How to extract	Objective / possible learning activity indicator
Number of edited pages	Count the pages edited by a user	Indicates the number of learning resources this user has contributed to
Total number of edits	Count the edits done by a user	Indicates the number of edit-activities this user has provided
Edit frequency	Calculate the average time span between two subsequent edits for a user	Indicates if a user performs a lot of edits to many different articles very frequently
Edit length	Calculate the average word count over all edits for a user	Indicates if a user mainly provides minor edits (e.g., just corrections) or major contributions

Table 6. Overview of the user based features with relation to collaborative editing.

Resource-based features

Feature	How to extract	Objective / possible learning activity indicator
Page title	Can be extracted from the .csv file created by the *WikiEvent* library	By analyzing the titles of the pages a user has edited, we can identify the topics this user is interested in
Number of edits	Count the edits done to a specific page	Indicates if a learning resource (i.e, page) is maintained
Number of users	Count the distinct users that edited a specific page	Indicates the interest in a specific learning resource (i.e., page)

Table 7. Overview of the resource based features with relation to collaborative editing.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Group / community-based features

The identified community-based features are summarized in Table 8.

Feature	How to extract	Objective / possible learning activity indicator
Community structure - co-edits	By connecting users that collaboratively edited a page	Indicates the community structure based on collaborative editing (i.e., learning/working together)
Community label	By processing personal information of users or analysing the type of resources they interact with	Detect how learners in a network are grouped, which are the largest communities or how communities overlap (based on users' data or topics of interest)
Community structure - block model	Apply block-model algorithm from the graph theory on a constructed network	Obtain the community structure of the learners' network
Modularity score	Apply Newman modularity algorithm from the graph theory on a constructed network	Indicates the existence of strong/weak communities in the network

Table 8. Overview of the community based feature definition, extraction and relation to collaborative editing.

Communicating and Discussing Learning Activities

Different online platforms, available to date, enable learners the consumption of learning resources (i.e., reading tweets in Twitter or accessing a link posted in Reddit) and facilitate the communication between learners and discussion of learning resources and learning activities. This use case is also identified in terms of Didactalia (i.e., visiting content, reading, clicking on external links, following users and topics). In this section we outline the cases of two example platforms: Twitter and Reddit.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Twitter

Over the past years, the microblogging platform Twitter has become one of the most popular social networks on the Web. Users can build a network of follower connections to other Twitter users, which means that they can subscribe to content posted by their *followees*. Twitter was also the first social platform that adopted the concept of *hashtags*. Hashtags are freely-chosen keywords starting with the hash character "#" to annotate, categorize and contextualize Twitter posts (i.e., *tweets*). The advantage of hashtags is that anyone with an interest in a hashtag can track it and search for it, thus receiving content posted by somebody outside of their own Twitter network. For example, users can retrieve tweets created during the European football championship by searching for the hashtag *#euro2016*, even if they do not have a social link to the tweet producers. Meanwhile, many social platforms, such as Instagram and Facebook, have adopted hashtags as well.

By extracting hashtags (i.e., considered as artifacts) that particular users applied and analysing relations between users and hashtags we define features that are related to learning activities. The following subsections provide details from one of our recent studies [KPL17], in terms of which such features were extracted,

Feature definition and extraction

We crawl two datasets using the Search API of Twitter[9]. The first one (i.e., "*CompSci*" dataset) consists of researchers from the field of computer science and their followees, while the second one (i.e., "*Random*" dataset) consists of random people and their followees. These datasets represent two different network settings: (i) a domain-specific one, in our case the domain of computer scientists, and (ii) a more general one consisting of random Twitter users. Our crawling strategy for both datasets comprises of the following four steps:

1) Crawl seed users. We start with identifying and crawling a list of seed users for each dataset. In the case of the "*CompSci*" dataset, we take the users who were identified as computer scientists in the work of [HJ14]. In the case of the "*Random*" dataset, we used the Streaming API of Twitter[10] in October 2015 to get a stream of tweets and extracted the user-ids to get our list of random seed users. From both user lists, we remove all users with more than 180 followees, which results in 2,551 seed users for the "*CompSci*" dataset and 3,466 seed users for the "*Random*" dataset. The threshold of using a maximum of 180 followees is chosen because the Twitter Search API only allows 180 requests per 15 minutes, which gives us the possibility to crawl the tweets of all followees of a seed user within this reasonable time window.

[9] https://dev.twitter.com/rest/public/search
[10] https://dev.twitter.com/streaming/overview

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

2) Crawl followees. Next, we use these follower relationships to crawl the followees of the seed users in order to create a directed user network for analyzing the social influence on hashtag reuse. Based on the number of seed users, the average number of followees per seed user is 94 in the case of the "*CompSci*" dataset and 72 in the case of the *"Random"* dataset. Overall, our crawling procedure gives us 91,776 distinct users for the "*CompSci*" dataset and 127,112 distinct users for the *"Random"* dataset.

3) Crawl tweets. In the third step, we crawl the 200 most recent tweets of all the users and remove the tweets in which no hashtags are used. The threshold of a maximum of 200 most recent tweets is set because of another restriction of the Twitter Search API that only allows 200 tweets to be received per a single request. This crawling procedure results in 5,649,359 tweets for the "*CompSci*" dataset with an average number of tweets per user of 61, and 8,157,702 tweets for the *"Random"* dataset with an average number of tweets per user of 64. Our crawled tweets cover a time range from 2007 to 2015.

4) Extract hashtag assignments. Finally, we extract the hashtag assignments by searching for all words that start with a "\#" character. This results in 9,161,842 hashtag assignments for 1,081,403 distinct hashtags in the "*CompSci*" network and 13,628,750 for 1,507,773 in the *"Random"* network. Thus, in both datasets, each distinct hashtag is used approximately 9 times on average and each user uses approximately 100 hashtag assignments in her tweets on average. Examples for popular hashtags are #bigdata, #iot and #ux in case of the "*CompSci*" dataset, and #shahbag, #ff and #art in case of the *"Random"* dataset.

The statistics of theses datasets are summarized in Table 9.

Dataset	$\|U_S\|$	$\|F\|$	$\|U\|$	$\|T\|$	$\|HT\|$	$\|HTAS\|$
CompSci	2,551	241,225	91,776	5,649,359	1,081,403	9,161,842
Random	3,466	252,219	127,112	8,157,702	1,507,773	13,628,750

Table 9. Features of our "*CompSci*" and *"Random"* Twitter datasets. Here, |Us| is the number of seed users, |F| is the number of followees of these seed users, |U| is the number of distinct users, |T| is the number of tweets, |HT| is the number of distinct hashtags and |HTAS| is the number of hashtag assignments.

This data enables us to gather features such as (i) the follower relationship between users, (ii) the set of hashtags used by a user, and (iii) the set of users interested in a specific hashtag. These features are summarized in Tables 10 to 12.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Reddit

Reddit is a portal, in which registered users can post content such as direct links or texts, known as submissions. They can also use the comment function to discuss and and to up or down vote these submissions. Users can also create new subreddits (i.e., communities) that focus on a particular topic. A collaboration network can be extracted by processing users' contributions that are included in publicly available submissions'[11] and comments'[12] [13] dumps. If a user contributed to a submission of another user by providing a comment, or if they both commented to a thread (submission), an edge is created between them. The particular subreddits can be used as (ground-truth) communities in the collaboration network and each user can be assigned to the subreddit, to which she contributed the most.

Feature definition and extraction

The publicly available submissions and comments dumps are provided for a period of 8 years (2006/07-2015/16) as monthly files compressed with bzip2 compression. After decompression, each file is comprised of series of JSON blocks delimited by new lines. Main fields of JSON files correspond to the fields documented in the Reddit API[14], however, there is a lack of documentation included in the dumps.

A JSON block of a monthly Reddit comments file contains 22 fields, each of them containing "either some text (the body of the comment), an integer (the number of upvotes), a boolean (whether the comment has been archived), or a timestamp (when the comment was made)"[15]. A JSON block of a monthly Reddit submissions files is constructed in a similar way containing 32 fields.

For illustration purposes, we include the following snippets of a JSON block from Reddit comments and submission files from year 2014, month 5:

Reddit comments:

{u'subreddit_id': u't5_2zhuq', u'removal_reason': None, u'subreddit': u'TsundereSharks', u'id': u'ch6fmzc', u'gilded': 0, u'archived': True, u'author': u'feline_crusader', u'parent_id': u't1_ch6eg27', u'score': 2, u'retrieved_on': 1433596941, u'controversiality': 0, u'body': u"It is!! But that baka scuba-kun didn't even notice!", u'edited': False, u'author_flair_css_class': u'pinkbow', u'downs': 0, u'link_id': u't3_24ebbv', u'score_hidden': False, u'name': u't1_ch6fmzc', u'author_flair_text': u'', u'created_utc': u'1398902400', u'ups': 2, u'distinguished': None}

[11] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/
[12] https://www.reddit.com/r/datasets/comments/3mg812/full_reddit_submission_corpus_now_available_2006/
[13] https://archive.org/details/2015_reddit_comments_corpus
[14] https://www.reddit.com/dev/api/
[15] http://blaze.pydata.org/blog/2015/09/08/reddit-comments/

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Reddit submissions:

{u'domain': u'imgur.com', u'banned_by': None, u'media_embed': {}, u'subreddit': u'cumsluts', u'selftext_html': None, u'selftext': u'', u'secure_media': None, u'link_flair_text': None, u'id': u'24i0tq', u'gilded': 0, u'secure_media_embed': {}, u'stickied': False, u'author': u'ddolson', u'media': None, u'score': 183, u'retrieved_on': 1407829466, u'over_18': True, u'thumbnail': u'nsfw', u'subreddit_id': u't5_2sl16', u'edited': False, u'link_flair_css_class': None, u'author_flair_css_class': None, u'downs': 0, u'is_self': False, u'permalink': u'/r/cumsluts/comments/24i0tq/sharing/', u'url': u'http://imgur.com/KPiw7X3', u'author_flair_text': None, u'title': u'Sharing', u'created_utc': 1398988799, u'ups': 183, u'num_comments': 3, u'distinguished': None}

The detailed description of the fields and their meaning is provided in the data structure documentation[16]. After examining the JSON blocks, the fields that are relevant to user interactions related to learning activities could be extracted and stored in a csv format, for example, for more convenient processing and analysis.

User-based features

To extract user based features both submissions and comments dumps should be analysed and processed. There is no sufficient user data included in the dumps that can be analysed to extract a user profile.

User social status - karma. In Reddit, users can accumulate so-called "karma" scores that rise if a user receives good ratings for her posts. Karma scores represent more a reflection of her vibes in the community and also what the community thinks about her. Karma scores could be applied as a proxy for social status. Karma scores are not included in the publicly available Reddit dumps, but they can be crawled via the public API using the python-based PRAW API wrapper[17].

Social ties between users are created if a user contributed to a submission of another user by providing a comment, or if they both commented to a thread (submission). By matching the *'author'* and *'id'* fields in both submissions and comments dumps, social ties feature can be extracted and a collaboration network can be build. By identifying how often a user replied to another user, we can detect the **strength of the social tie** between them.

As already introduced in the StackExchange use case, algorithms from the graph theory can be applied here as well, to capture features such as node degree, or degree assortativity, which emphasize users that contribute the most in the network (i.e., users with the high node

[16] https://github.com/reddit/reddit/wiki/JSON

[17] https://praw.readthedocs.io/en/stable/

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

degree) or how users are connected to each other (i.e., if they tend to connect to similar or dissimilar users).

User-resource relation feature can be extracted by analysing the unique number of authors (through the *'author'* field) in both submissions and comments corpus and by detect the ratio of users contributing with content (i.e., submissions) and users commenting, discussing or voting submissions.

Feature	How to extract	Objective / possible learning activity indicator
User social status - karma	'link_karma' and 'comment_karma' fields can be crawled via the public API using the python-based PRAW API wrapper	Identify most influential users in terms of karma - a reflection user's vibes in the community
Collaboration network, social ties	Match *'author'* and *'id'* fields in both submissions and comments dumps	Collaboration between users - discuss posts, exchange of opinions, knowledge, experience
Social tie strength	How often a user replied to another user	The intensity of the interactions between users
Follower relationship between users	*GET friends/ids* call of the Twitter Search API	The list of followers / followees of an user
User – resource relations	Identify unique authors in both submissions and comments corpuses from the *'author'* field	Detect the ratio of users contributing with content (i.e., submissions) and users commenting, discussing or voting submissions
User hashtags	*GET search/tweets* call of the Twitter Search API and extract hashtags (starting with # character)	The set of hashtags used by an user describing the topics this user is interested in

Table 10. Overview of the user based feature definition, extraction and relation to learning activities in Reddit.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Resource-based features

Resource based features can be extracted by analysing the fields available in the public submissions and comments corpuses and by crawling other corresponding fields through the Reddit public API.

Feature	How to extract	Objective / possible learning activity indicator
Resource type (e.g., comment, link, subreddit)	By crawling the public API to extract the *'kind'* field	Determine submission/comment (per subreddit) ratio Identify subreddits with the higher number of contributions
Submission/subreddit - comment relation	By processing the *'num_comments'* field in the submissions corpus	Identify submissions or subreddits with a higher number of comments, indicating the user engagement to a particular topic
Resource popularity	*'score'*, *'ups'*, *'downs'* fields in submissions and comments corpuses *'accounts_active'* field of a subreddit (crawled from the public API) gives the number of active users in last 15 minutes *'subscribers'* field gives the number of users subscribed to a subreddit	Identify most popular resources (links, text, comments or subreddits)
Resource topic	*'titel'* and *'subreddit'* fields could be analysed	Cluster resources based on topics / subreddits

Table 11. Overview of the resource based feature definition, extraction and relation to learning activities in Reddit.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Group / community-based features

Reddit is organized into so-called areas of interest (i.e., the subreddits). Since subreddits can in fact be regarded as explicit semantic communities, we treat **subreddits as communities**. This feature can be extracted by investigating '*subreddit*' and '*subreddit_id*' fields included in the public dumps and by matching a particular user with a subreddit to which the user contributed the most.

Similarly to the use cases introduced in previous sections, algorithms from the graph theory can be applied on the constructed collaboration network to detect communities and to determine the modularity score.

Feature	How to extract	Objective / possible learning activity indicator
Community - subreddit	Subreddits (i.e., topic based content categorization) provide ground-truth communities. Extracted through '*subreddit*' and '*subreddit_id*' fields	Detect how learners in a network are grouped, which are the largest communities or how communities overlap
Community structure - block model	Apply block-model algorithm from the graph theory on a constructed network	Obtain the community structure of the learners' network
Community based on hashtag	Gather all users that have used a specific hashtag	The set of users (i.e., the community) interested in a specific hashtag
Modularity score	Apply Newman modularity algorithm from the graph theory on a constructed network	Indicates the existence of strong/weak communities in the network

Table 12. Overview of the group / community based feature definition, extraction and relation to learning activities in Reddit.

Collecting and Structuring Learning Resources

Learners can collect and structure learning resources by applying available tools that are discussed in this section. For example, in Bibsonomy learners are able to store and organize their bookmarks and publication entries. Learners can tag their resources, which helps them to structure and re-find information. In Didactalia educational resources are gathered,

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

organized and published as structured data, mainly by the Didactalia team, but also by community members.

Microsoft Academic Graph

Microsoft Academic Graph is a very large dataset containing scientific publication data from different disciplines with detailed information about citations, authors, institutions, publication venues (journals or conferences) and field of studies[18]. Data pre-processing is time consuming due to the size of this dataset. During our investigations on this dataset, we encountered problems with author name disambiguation.

Feature definition and extraction

The publicly available dataset[19] consists of text files that represent tables containing data columns separated by tabs. The six main entities, such as authors, publications, affiliations (institutions), venues (journals and conferences), fields of study and events (specific conference instances), derived from the dataset and their relations (e.g., citations, authorship) are presented in Figure 3.

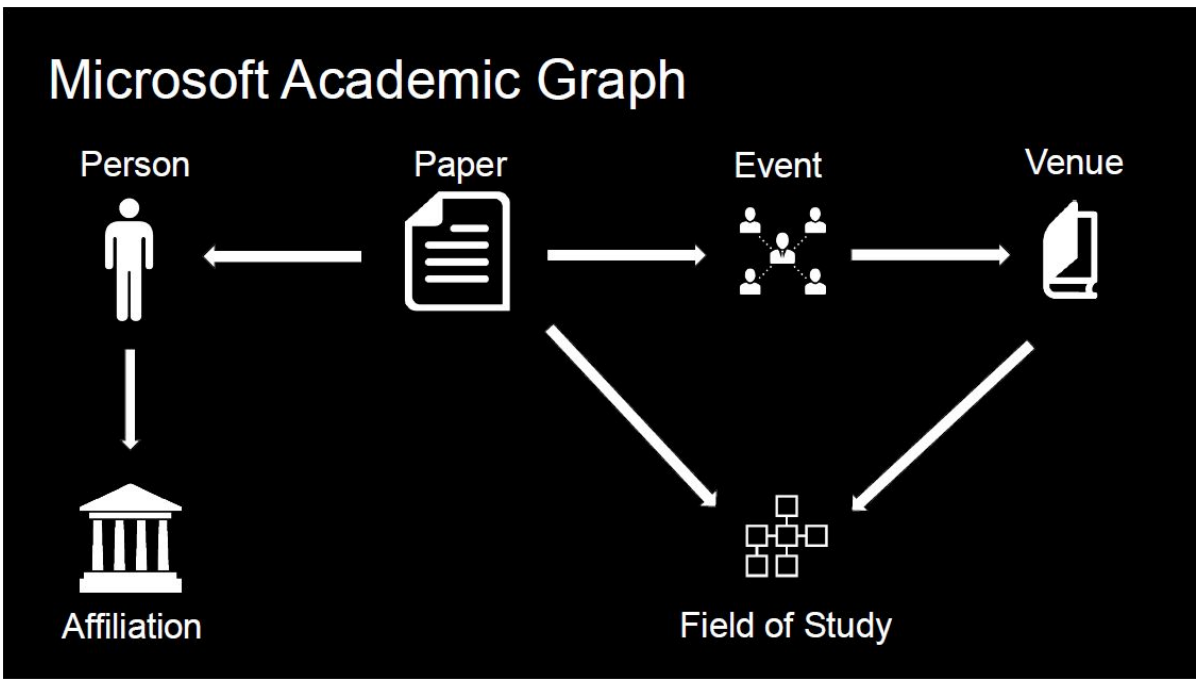

Figure 4. Entity relations in Microsoft Academic Graph[20]

[18] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[19] https://academicgraph.blob.core.windows.net/graph-2015-08-20/index.html

[20] http://humanities.uva.nl/~kamps/gsb15/keynote/AlexWade.pdf

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

The relationships between entities of the Microsoft Academic Graph provide opportunities for the extraction of features characteristic for each of the entities and also for measuring scholarly impacts and influences [HK16], [SSSME+15]. An overview of the complete schema can be found under [21].

To extract user based features from the Microsoft Academic Graph, the relations between the *Author* entity (i.e., table in the dataset) and the *PaperAuthorAffiliations* table should be analysed first. In this case, the three columns: Paper ID, Author ID and Affiliation ID should be extracted to detect, for example, which papers a particular author has published or which is the author's affiliation. By further investigating the Paper ID and Author ID columns a co-authorship network for an author can be constructed.

By mapping the Paper ID to the corresponding columns of *ConferenceSeries*, *Journals* and *FieldsOfStudy* tables, information such as the prefered venues and areas of interest for an author can be extracted.

Data on resources respectively papers is mostly included in the *Papers* table, but through relations with other tables many characteristic features can be detected. Such features could be, publication age, distribution of publications through venues or fields of study, paper rank, or number of co-authors.

Bibsonomy

BibSonomy[22] is a social bookmarking and publication sharing system. Its dataset is freely available and can be downloaded for scientific purposes[23]. We utilize the database dump from 2015-01-01 to gather all tags assigned to resources (i.e., Web links and academic references). This results in 772,112 bookmarks, 10,180 users, 683,482 resources, 199,594 tags and 2,981,038 tag assignments.

Feature definition and extraction

For this purpose three files of this dataset are investigated in detail, namely *tas*, *bookmark* and *bibtex*. Since the *bookmark* and *bibtex* files provide metadata for the resources, we focus on the *tas* file, in which the user interactions are provided.

This file consists of the following attributes for each bookmark:

- user (number; user names are anonymized)
- tag

21 https://raw.githubusercontent.com/halolimat/Microsoft-Academic-Graph/master/Microsoft_Academic_Graph_files/images/Microsoft_Academic_Graph.jpg

22 http://www.bibsonomy.org/

23 http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

- content_id (matches bookmark.content_id or bibtex.content_id)
- content_type (1 = bookmark, 2 = bibtex)
- date

This enables us to extract both, user-based and resource-based features, based on the given tag assignments, which are summarized in Tables 13, 14 and 15.

Apart from that, in [KL16], we showed that the same features can be extracted from other social bookmarking systems, such as CiteULike, Delicious, Flickr, MovieLens and LastFM, as well.

Didactalia

Didactalia is a global community and a content storage Website for teachers, students and parents to create, share and find open educational resources; it has about 186,000 structured items (100,000 in the collection) with semantic contexts and over 300,000 registered users (more info: D5.1, Section What is Didactalia?).

Didactalia could also belong to the other previous use cases, as it includes tools for collaborative editing, Q&A, and communicating and discussing learning activities. It is being presented here, in the section of 'collecting and structuring learning resources' because the core component of Didactalia are the resource collections, where educational resources are gathered, organized and published as structured data, mainly by Didactalia team, but also by community members.

In this use case, user-artifact interactions would be:

- Consumption: Visiting content
  - Reading
  - Watching videos
  - Downloading content
  - Clicking on external link
  - Exploring the graphic visualization of the resource knowledge graph
- Browsing (see use case 'Searching and Browsing')
- Searching (see use case 'Searching and Browsing')
- Contributions (posts, comments, likes...)
  - Sharing: Collecting and structuring learning resources
  - Posts
  - Collaborative editing (not currently used in Didactalia collection, although the tool is available)
  - Comments
  - Likes
  - Questioning

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

    - Answering
- Playing: learning games (see specific use case ‘Gamification’)
- Formal activities
    - Tests
    - Tasks
- Communication
    - Communicating and discussing learning activities: proactive communication (following users and topics)

In the next tables, we enumerate the Didactalia features:

- User-Based, related with each user and the effects of the relations between users.
- Resource-Based, related with each resource, and with the classification and relations of the resources.
- Community based, related with the activity of formal groups of users.

User-based features

Feature	How to extract	Objective / possible learning activity indicator
User profile	See D5.2, sections 3 and 4. a) User registered data, from the GNOSS API calls: community subscriptions, user subscriptions and community membership. Furthermore, information from the biography or short biography could be used.. See D5.2, section 6 and 7. b) Additional information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog. The complete set of data will be developed in AFEL project.	see Table 3
User social status	See D5.2, sections 3, 4, 6 and 7. a) Resources data, from the GNOSS API calls: editors, readers, comments.	see Table 3

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

	b) Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the JavaScript code. The complete set of data will be developed in AFEL project.	
Social ties & Social tie strength	See D5.2, sections 3, 4, 6 and 7. a) Resources data, from the GNOSS API calls: editors, readers, comments. b) Information about social learning activities, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the JavaScript code. The complete set of data will be developed in AFEL project.	see Table 3
User contributions	See D5.2, sections 3, 4, 6 and 7. a) Resources data, from the GNOSS API calls: editors, readers, comments. b) Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project.	see Table 3
User-to-user relations	See D5.2, sections 3 and 4, 6 and 7. a) User registered data, from the GNOSS API calls: user subscriptions. b) Inferred information from the data of social actions.	see Table 3
User-resource relations	See D5.2, sections 3, 4, 6 and 7. a) Resources data, from the GNOSS API calls: editors, readers, comments, tags, categories, etc. b) Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code.	see Table 3

Table 13. User-based features extractions in Didactalia

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Resource-based features

Features	How to extract	Objective / possible learning activity indicator
Resource type	See D5.2, section 2 Resource information from GNOSS API	see Table 4
Resource popularity	See D5.2, section 2,3, 5 and 6 Information to be inferred from the API data and the social actions data, to be developed in AFEL project.	see Table 4
Resource topic	See D5.2, section 2 a) Resource information from GNOSS API b) Automatic detection of topics, to be developed in AFEL project.	see Table 4
Question-best answer relation	We could use the number of votes of each comment. This information should be related with the user information, in order to calculate his or her social status (to be developed in AFEL project).	see Table 4
Cross-resource relation	To be developed in AFEL project. We could exploit linked resources. The link can be explicit (link functionality or html link in the description of the resource), or implicit (contexts, graph navigation).	see Table 4

Table 14. Resource-based feature specifications in Didactalia.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Group / Community Based

The identified community-based features are summarized in next table:

Feature	How to extract	Objective / possible learning activity indicator
Community	See D5.2, sections 3 and 4, 6 and 7. a) User registered data, from the GNOSS API calls: user subscriptions. b) Inferred information from the data of social actions.	see Table 5
Community structure - block model	See D5.2, sections 3 and 4, 6 and 7. a) User registered data, from the GNOSS API calls: user subscriptions. b) Inferred information from the data of social actions: Apply block-model algorithm from the graph theory on a constructed network	see Table 5
Modularity score	See D5.2, sections 3 and 4, 6 and 7. a) User registered data, from the GNOSS API calls: user subscriptions. b) Inferred information from the data of social actions: Apply Newman modularity algorithm from the graph theory on a constructed network	see Table 5

Table 15. Community feature specifications in Didactalia.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Searching and Browsing

The searches and contexts are the usual expression of the Knowledge Graph of the GNOSS communities, such as Didactalia. Both have been explained in detail in section 5 of D5.1.

Didactalia

Didactalia offers navigation systems based on knowledge graphs and a context generation system that are thought for users to enjoy a search experience based on personal reasoning and not in previously administrated processes to present information. Didactalia search engine meets three main characteristics of faceted search engines: concatenated searches based on properties of the items in the result list, summarization of results and refinement options that offer only possible results.

The faceted search engine of Didactalia, accessible both in the home and in the 'Educational Resources' page, contains the most relevant filters to look for educational material (subject, age, language, resource type and tags).
Semantic contexts use the educational graph of Didactalia to enrich and complement every resource from the perspective of other approaches or disciplines.

In this use case, user-resource interactions would be:

- Browsing
    - Navigating to other contents
    - Browsing to a new search
    - Following a pre-established learning path/graph (suggested by the teacher, the machine or other user)
    - Browsing/navigating through entities (including the study the influence of the presentation of the graph)
- Searching
    - Text box (free text search). Examples of use cases in searchers:
        - Successive searches:
            - Ex. 1: search 1 - concept 1, search 2 - synonym of concept 1.
            - Ex. 2: search 1 - concept 1, search 2 - subclass of concept 1.
    - Iterative searches: Filters (use of faceted search)
    - Repeated searches (1st, 2nd, 3rd… time of search for the same concept/facet)
    - Saved searches

In this use case, we focus on user-based features as shown in Table 16.

User-based features

Features	How to extract	Objective / possible

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

		learning activity indicator
Topics of search	See D5.2, sections 6 and 7. Information for registered users and anonymous users, from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project.	Identify what the user is interested in
Repeated searches by the same user (same search)	See D5.2, sections 6 and 7. Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project	Interesting to establish a leveling of learner interest
Successive searches	See D5.2, sections 6 and 7. Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project	It shows which terms the learner considers to be related and could help to determine their 'learning scope'
Popular searches	See D5.2, sections 6 and 7. Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project	To construct learning recommendations and search suggestions: to help learner to find out common topic that other people ('Model learner') have searched for
Saved searches	To be developed in the AFEL Project	- Leveling of learner interest

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

		-Evolution of user interest over time
Use of faceted searches	See D5.2, sections 6 and 7. Information for registered users and anonymous users, from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project	-It can help discern if learners browse the platform without a clearly pre-established goal or with it - Leveling of learner interest
Relation between previous and currently visited Webpage (topic/NER/popularity/users involved…)	See D5.2, sections 6 and 7. Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project.	Identify whether the learner can navigate in breadth (dispersion) or depth (targeting)
Actions when browsing/ browsing path	See D5.2, sections 6 and 7. Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project.	To recommend learning paths that have worked for other learners
Progress in a pre-established learning path/graph	See D5.2, sections 6 and 7. Information for registered users and anonymous users, inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in AFEL project.	The learner's navigation through his pre-established learning path give information on his progress and commitment to achieve the learning objectives

Table 16. User-based features specifications in Didactalia (Searching and browsing)

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Navigation Through Entities of a Graph

At the end of the each content description, Didactalia shows an interactive graphical visualization of the graph of related topics for exploring the connections between DBpedia articles and other educational Didactalia contents. The graph invites the user to keep track of the relationship between connected entities, so that it offers a frame and helps to understand the learning object that was the starting point (See D5.1, Section 'Didactalia Knowledge Graph')

Extractors developed for AFEL now provide information about the navigation that the user follows informing about resource from a context (the user reaches a resource following a link in a context) but it is necessary too to provide information about the navigation through the entities of the graph. In the current visual representation of the Didactalia Graph it would be interesting to know: (i) the route of nodes (entities) clicked by the visitor (that gives access to the reduced view of the Wikipedia article corresponding to the entity); (ii) the connections between nodes followed and read; (iii) the entities consulted in greater depth, the reduced view of the article can be navigated to the page with the complete article; (iv) once in the page of the article, it would be necessary to know if navigation between entities of this new graph is continued (See D 5.1, Section 'Gnoss data context')

The study of navigation through entities is one of the research studies to be designed and carried out in Didactalia. The identification of specific features for this use case will be one of the results of the project. The Didactalia specific extractors for AFEL, with the corresponding required improvements, will be used to provide the necessary data according to the study. It must be considered a possible change in the visual representation of the entities graph.

This navigation information will be inferred from the data of social actions pushed from Didactalia to the AFEL Catalog by the Javascript code. The complete set of data will be developed in course of the AFEL project and therefore the extracted user-based, resource-based and community-based features will be presented in a future deliverable.

Gaming

Didactalia has created 450 interactive multi language games (English and Spanish at this moment) as learning games to acquire geography and anatomy knowledge. The geography games include questions about continents, countries, capitals, flags, regions, states, mountains, oceans, earth, atmosphere, waters and other topics[24]. The anatomy games[25] include questions about the cellular parts (animal and vegetable cells) as well as about the

[24] https://mapasinteractivos.didactalia.net/en/community/mapasflashinteractivos/MapasDidactalia
[25] https://cienciasnaturales.didactalia.net/en

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

human body (organs and systems).

For each game, there are two types of play: “What's the name?” and “Where is it?”, both with information about scoring attempts per game. Once the game is started the user has the possibility to activate the mode “Study”, a training mode that allows students to memorize easily the main items of the game. Players can play from any type of mobile device or computer, either against the clock or against other players thanks to the services of Challenges and Tournaments, which make it possible to compete against other players of the platform in one-on-one mode (Challenges: from the ranking of each game) or in group mode (Tournaments[26] [27]).

The study about gamification in learning games is one of the research studies to be designed and carried out in Didactalia. Thus, the corresponding features will be defined in a future deliverable.

[26] https://cienciasnaturales.didactalia.net/torneos

[27] https://mapasinteractivos.didactalia.net/en/community/mapasflashinteractivos/torneos

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Feature Specification

In this section, we specify features independently of data sources and use cases that apply in general settings and describe them taking the data sources es examples. Thus, this section serves as a summary of the features identified in the various use cases. Furthermore, this section outlines the top-bottom perspective of feature engineering indicating that features identified here are applicable in different use cases in general and can also be extracted. Identified features can also be included in the AFEL Schema.

User-Based Features

Feature	Description	Relation to use cases
User profile	Information on user's interests (e.g., given by the set of used tags / hashtags), competences, expertise, role within a community and user's motivation	Questioning and Answering, Communicating and Discussing Learning Activities, Collecting and Structuring Learning Resources
User social status	Identify the most influential learners in online communities	Questioning and Answering
Social ties	Collaboration (co-posting) activity between two users	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing
Social tie strength	Strength of the collaboration taking place between two users	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing
User contributions	Which users provided which amount of the content or contributed to which number of posts (e.g., the number of edited pages, the total number of edits,	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

	and the average time span between two subsequent edits)	editing, Collecting and Structuring Learning Resources
User-to-user relations	Determines if similar users tend to connect together or if they rather connect to dissimilar users	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources
User-resource relations	Find out which users interact with which resources, topics of resources (e.g., tags or hashtags) or searches of resources	Questioning and Answering, Collaborative editing, Communicating and Discussing Learning Activities, Collecting and Structuring Learning Resources
User meta-knowledge	Shows to what extent are users aware of the available content within the online community (referring other users to a similar issue that was handled in the past)	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources
Topics of search Repeated searches by the same user (same search) Successive searches Popular searches Saved searches	Important information to identify user's interest, leveling of user's interest, user's learning scope, evolution of user's interest over time and to construct learning recommendations and search suggestions	Searching and Browsing
Use of faceted searches	Shows if users browse the platform without a clearly pre-established goal or with it	Searching and Browsing

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Relation between previous and currently visited Webpage	Identify whether the user can navigate in breadth (dispersion) or depth (targeting)	Searching and Browsing
Actions when browsing/ browsing path	Relevant information to recommend learning paths that have worked for other users	Searching and Browsing
Progress in a pre-established learning path/graph	Determine user's progress and commitment to achieve the learning objectives	Searching and Browsing

Table 17. User based feature specifications.

Resource-Based Features

Feature	Description	Relation to use cases
Resource type	The type of a posting, e.g., question or answer in StackExchange	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources
Resource popularity	Identify most popular and high quality resources	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources
Resource topic	Describes the topic of a resource (e.g., based on assigned tags)	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources
Question-best answer relation	By identifying the best accepted answer or solution	Questioning and Answering

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

	to a particular problem we can confirm that users collaboratively contributed to solve a particular problem. Furthermore, by finding the user that provided the best answer we can also relate this information to user's role or social status and check the correlation.	
Cross-resource relation	Detects relations between resources in a corpus. It also indicates to what extent are users aware of the available content within the online community, if they provide a reference to a similar issue that was handled in the past (e.g., Q&A sites).	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources
Resource similarity	Based on such meta-knowledge of users, we could find out the similarity between resources	

Table 18. Resource based feature specifications.

Groups / Communities-Based Features

Feature	Description	Relation to use cases
Community label	Based on user's personal information or their topics of interest, reflected through the subject of resources they interact with, it is possible to detect how learners in a	Questioning and Answering, Communicating and Discussing Learning Activities

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Community - subreddit	network are grouped, which are the largest communities or how communities overlap Special case: Reddit-subreddits (i.e., topic based content categorization) provide ground-truth communities	
Community structure - block model	Enables us to obtain the community structure of the learners' network by applying the stochastic block-model algorithm from the graph theory	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources
Modularity score	Determines the strength of the community structure	Questioning and Answering, Communicating and Discussing Learning Activities, Collaborative editing, Collecting and Structuring Learning Resources

Table 19. Groups and communities based feature specifications

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Conclusion

With this document we made the first step towards determining the characteristic features that are relevant to online learning activities from the data sources captured within the AFEL project.

This deliverable facilitates the feedback loop between visual analytics, data enrichment and modeling tasks within the project. Therefore, it applies the results achieved in WP1, WP4 and WP5 as valuable inputs and provides results for for WP2, WP3, WP4 and WP5.

Our methodological contribution gives insights on how to tackle the problem of feature definition and extraction in everyday learning settings. We tackle this problem from a use case perspective. We provide an initial specification of the features relevant to learning activities by presenting an instantiation of them on some of key data sources identified in D1.1 and GNOSS-Didactalia data sources identified in D5.1.

Finally, we outline the top-bottom perspective of feature engineering indicating that features identified here are applicable in different use cases in general and can also be extracted. Identified features can also be included in the AFEL Schema.

For future work, we plan to to update our feature tables as soon as more data is inserted into the AFEL platform. Furthermore, we will already build upon our presented feature engineering work by implementing a learning graph recommender algorithm for Didactalia. Therefore, especially resource-based features as well as user-based features (e.g., user-specific interactions with resources) are relevant.

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

References

[AFGYL+16] Alessandro Adamou, Besnik Fetahu, Ujwal Gadiraju, Ran Yu, Elisabeth Lex, Matthias Traub, Peter Holtz, Ricardo A. Maturana, Esteban Sota, Susana López-Sola, and Mathieu d'Aquin. *D1.1 - Specification of data to be collected*. AFEL project deliverable (2016).

[AFELG16] AFEL Glossary. (2017). https://docs.google.com/document/d/14riOroN6uKSTrT_alWtEjEKhV3UGO7VWdbkF6jJ_FNo/edit#heading=h.8gbr5eyu0i22

[ALSGD+16] Ricardo Alonso Maturana, Susana López-Sola, Esteban Sota, Ujwal Gadiraju, Stefan Dietze, Mathieu d'Aquin, Alessandro Adamou. *D5.1 - Description of the available social environments and data.* AFEL project deliverable (2016).

[BIS06] Christopher Bishop. *Pattern recognition and machine learning*. Berlin: Springer. ISBN 0-387-31073-8 (2006).

[CK08] Cress, Ulrike, and Joachim Kimmerle. *A systemic and cognitive view on collaborative knowledge building with wikis*. International Journal of Computer-Supported Collaborative Learning 3, no. 2 (2008): 105-122

[DDA16] Mathieu d'Aquin, Keyur Dave and Alessandro Adamou, The AFEL deliverable template, AFEL document (2016).

[DUB09] Ryan Dube. Characteristics of Social Networks. (2009). http://socialnetworking.lovetoknow.com/Characteristics_of_Social_Networks

[EB11] Everitt, Brian (2011). *Cluster analysis*. Chichester, West Sussex, U.K: Wiley. ISBN 9780470749913.

[F10] Santo Fortunato. *Community detection in graphs*. Physics Reports, Volume 486, Issues 3–5, February 2010, Pages 75-174, ISSN 0370-1573, http://dx.doi.org/10.1016/j.physrep.2009.11.002.

[FHW16] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for *"Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann, Fourth Edition, 2016.

[HJ14] Hadgu A., and Jäschke R. *Identifying and Analyzing Researchers on Twitter*. In Proceedings of the 2014 ACM conference on Web science (WebSci'14). ACM, New York, NY, USA, 23–30. (2014)

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

[HGCLH16] Ilire Hasani-Mavriqi, Florian Geigl, Subhash Chandra Pujari, Elisabeth Lex, and Denis Helic. *The Influence of Social Status and Network Structure on Consensus Building in Collaboration Networks.* Social Network Analysis and Mining 6, 1 (2016), 1-17

[HK16] D Herrmannova, P Knoth. *An analysis of the Microsoft Academic Graph*. D-Lib Magazine, 2016

[HKYCD+16] Peter Holtz, Joachim Kimmerle, Seren Yenikent, Ulrike Cress, Mathieu d'Aquin, Stefan Dietze, Besnik Fetahu, Ujwal Gadiraju, Belgin Mutlu, and Peter Hasitschka. *D4.1- Report on the analysis of learning and cooperation*. AFEL project deliverable (2016).

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. *Data clustering: a review*. *ACM Comput. Surv.* 31, 3 (September 1999), 264-323. DOI=http://dx.doi.org/10.1145/331499.331504

[KL16] Kowald, D., and Lex, E. *The influence of frequency, recency and semantic context on the reuse of tags in social tagging systems*. In Proceedings of the 27th ACM Conference on Hypertext and Social Media (pp. 237-242). ACM. (2016).

[KPL17] Kowald, D., Pujari, S., & Lex, E. (2017). *Temporal Effects on Hashtag Reuse in Twitter: A Cognitive-Inspired Hashtag Recommendation Approach.* In Proceedings of the 26th International World Wide Web Conference (WWW'2017). ACM. (Arxiv: https://arxiv.org/pdf/1701.01276v1.pdf)

[KN11] B. Karrer and M. E. J. Newman. *Stochastic blockmodels and community structure in networks.* Phys. Rev. E, Jan 2011.

[Luh95] Luhmann, Niklas, *Social systems*, Stanford University Press (1995)

[MMGDB07] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. *Measurement and analysis of online social networks*. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC '07). ACM, New York, NY, USA, 29-42. DOI=http://dx.doi.org/10.1145/1298306.1298311 (2007).

[MSBFB+14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

[SCCP09] Suh B., Convertino G., Chi E. H., and Pirolli P.. *The singularity is not near: slowing growth of wikipedia.* In Proceedings of OpenSym '09. ACM. (2009)

[SSSME+15] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. *An Overview of Microsoft Academic Service (MAS) and*

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246.

[TLLP13] Tammets K., Laanpere M., Ley T., Pata K. (2013) *Identifying Problem-Based Scaffolding Patterns in an Online Forum for Construction Professionals*. In: Hernández-Leo D., Ley T., Klamma R., Harrer A. (eds) Scaling up Learning for Sustained Impact. EC-TEL 2013. Lecture Notes in Computer Science, vol 8095. Springer, Berlin, Heidelberg

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org