FutureTDM
Explore . Analyse . Improve

TOOLS

IMPROVING UPTAKE OF TEXT AND DATA MINING IN THE EU

Facts

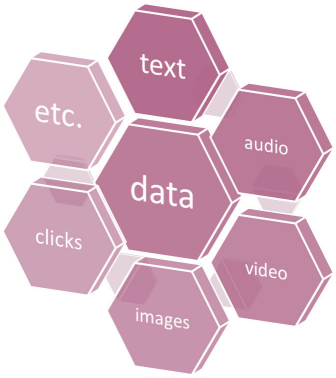
Project No: 665940

Program: H2020 | CSA | GARRI-3-2014

Duration: 09/2015 - 08/2017

Techniques, Tools and Technologies **FOR TDM IN EUROPE**

Data comes in a variety of forms i.e. text, audio, video, images, graphs, numbers, chemical compounds, likes, etc. and can be presented in a structured, unstructured or semi-structured way. Depending on the data medium and structure, different techniques are being deployed to extract information and knowledge.



What do we mean by data, structured or unstructured?

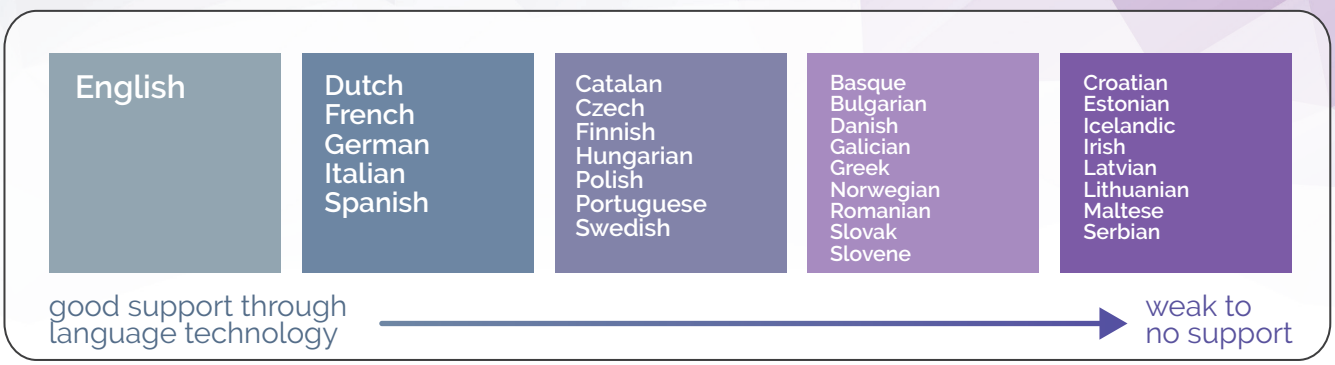
Structured data resides in fixed fields within a record or files (e.g. data in spreadsheets and/or relational databases). **Unstructured data** cannot be so easily classified and "understood" by machines, such is i.a. files with running text, audio, image and video data, graphics, webpages and emails, blog posts and tweets. In-between stands **semi-structured data**, which although does not confirm to rigidly defined text fields, contains tags and other annotation mark-up features.

Which are the TDM techniques employed?

Text Mining techniques are applied to textual, transcribed audio data and multimedia data converted into textual representations so as to turn it into structured data for further analysis. Data is usually linguistically annotated at multiple levels, both externally (e.g. by assigning a domain classification label to a document) and internally (e.g. by annotating spans of text as referring to concept(s) in an ontology terminological database, or as referring to a named entity, or as being the subject/agent of verb/predicate).

Data Mining techniques aim at extracting patterns by combining statistics and statistical analysis with machine learning and database management, using methods such as **association rule learning** (to discover relationships between variables in large databases, e.g. in recommendation systems and applications), **cluster identification and analysis** (to segregate data into smaller groups based on their common characteristics, e.g. TV-viewers or web application users for targeted marketing), **classification** (to identify the predefined category(-ies) in which particular data belong, e.g. mobile subscribers that will stop their subscription) and **regression** (to estimate how the value of a dependent variable will change when an independent

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.



variable changes, usually used to predict e.g. growth of an entity based on a number of macro/micro economic parameters).

What are the tools used for these techniques?

Two types of tools are central for TDM, namely **annotation workflows** and **editors**. Almost all text mining applications are in the form of **workflows** of operational modules, whereby each module's output is the input to the next module. Some modules can be common to many pipelines (e.g. sentence splitting), whereas other modules are task specific. Obviously, these modules need to be interoperable and based on a common structure. An indicative list of software packages, including analytics tools (standalone NLP tools), collections of components, workbenches with graphical user interfaces, as well as interoperability frameworks, though most of them can be multiply classified, is presented below.

Annotation editors are tools used for editing annotations in text and for their visualization. Annotation editors are central to many TDM tasks, such as the creation of corpora for system training and evaluation; visualization and error analysis of application output; manual correction of automatically created annotations; and validation of annotations prior to release for use. Tools falling under the heading annotation editors may cover various levels, notably text level annotations, document level metadata, and/or annotations using external resources such as ontologies, thesauri, gazetteers etc. Annotation editors typically constrain the end user to create and edit annotations according to some fixed schema or type system. Some editors also allow for ad-hoc, schema-less annotation, and others support ontology-backed schemas. Examples of such editors include Argo (Manual Annotation Editor), WebAnno, the GATE Developer and GATE Teamware, Brat, Alvis AE, Egas and WorkFreak.

Are all tools and technologies free for use?

A quite large number of tools are open-source while there are also proprietary frameworks catering for TDM, either across sectors or specializing in domain specific applications.

How does language affect tools and technologies?

Tools and technologies can be either language-independent or dependent. However, not all languages are supported equally well, as indicated in META-NET White Papers. English claims models and resources for almost all software packages; further well-supported languages include German, French, Spanish, Chinese and Arabic, followed by many languages with limited support.

Name	Implemented in
Alvis	Java
Apache cTAKES	Java/UIMA
Apache OpenNLP	Java
Apache UIMA	Java
Argo	Java/UIMA
Bluima	Java/UIMA
ClearTK	Java/UIMA
DKPro Core	Java/UIMA
GATE Embedded	Java
Heart of Gold	Java + Python
JCoRe	Java/UIMA
NLTK	Python

An indicative list of software packages