„Web Portal for Political Memory“

188.939 Bachelorarbeit für Informatik und Wirtschaftsinformatik (PR, 5.0h, 10.0EC)

University	TU Wien
Curriculum	E 033 526, Wirtschaftsinformatik
Student	Michael Zawrel, 1025319
	e1025319@student.tuwien.ac.at
Supervisor	Ao.Univ.Prof. Mag. Dr. Horst Eidenberger

Table of Content

1. Introduction

Credibility is, besides the actual contents, an important decision criterion for voters at democratic elections. This credibility relies substantially on the consistency of the statements of a political protagonist. Despite rapid development in search techniques, the proof of past inconsistencies is still time-consuming and requires memorization of the exact diction. The "Web Portal for Political Memory" (hereinafter also referred to as "WP4PM" or "project") facilitates the inquiry by collecting articles about politics and extracting the most relevant statements. A link to the information source as well as a copy of the original report are thereby preserved. In this manner, it may not only serve as personal decision support but also as a source of evidence in political arguments.

2. Functional Requirements

The original statement of task for this bachelor thesis has been as follows:

"Statements of politicians - though relevant - are often forgotten on the next day. This web portal should help to organize and memorize important sayings of Austrian politicians. It should allow to semi-automatically define profiles for politicians (e.g. from teletext headlines) and to semi-automatically store their statements. The viewer should summarize sayings (if possible, on topics defined by a few keywords) of politicians together with timestamps. Furthermore, the portal should provide a search function. The implementation can be based on a free CMS." [1]

In personal meetings between supervisor and student, however, the goal has been shifted. New target became developing a prototype of the said web portal with focus on the process from information retrieval to evaluation of identified quotes, with highest possible degree of automation and precision. The user interface should be functional to support this task. This definition results in the following list of requirements:

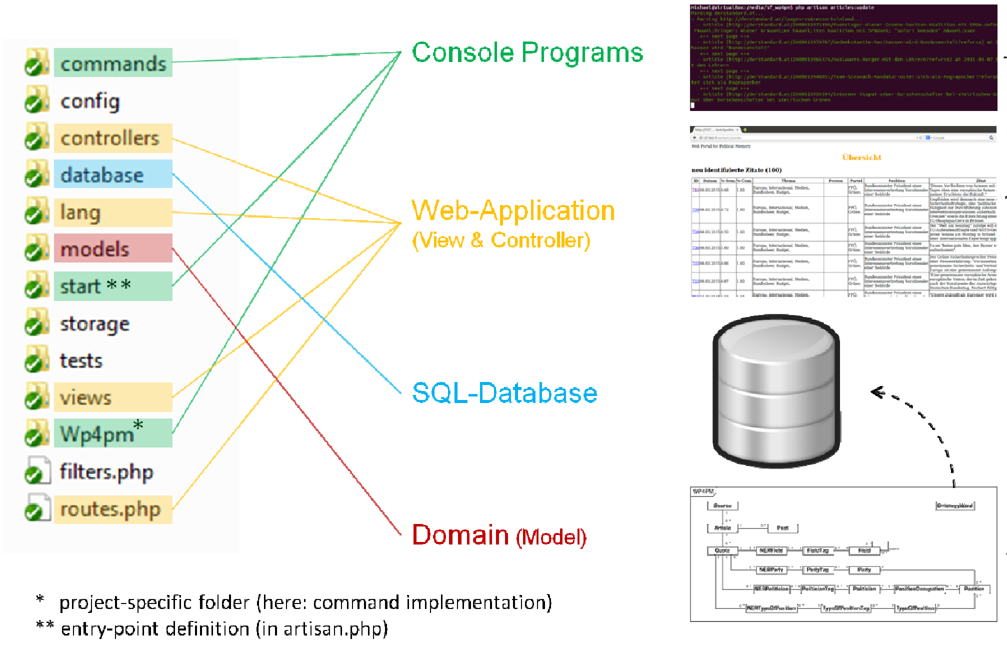
- automatic retrieval of data from internet sources
- automatic identification of direct quotes
- automatic identification of speaker and topic
- automatic evaluation of quote's relevance (decision: accept/reject)
- persistent storing of original articles, quotes and their metadata
- graphical representation of quotes within a web interface
- possibility for manual moderation of entries via graphical web interface

3. Description of Technologies

As a consequence, proper technologies had to be found to integrate all the database access, web representation as well as information retrieval and processing scripts.

The "Laravel" PHP framework comes up with an elegant answer to all those requirements. It allows easy creation of command line programs, simple rollout and migration of the database, has powerful O/R mapping, supports RESTful web applications, templating of views and facilitates the whole process of PHP development with its "convention over configuration" approach. All that leads to better maintainable applications. Additionally, Laravel performs a number of security measures (e.g. input validation) by default, allows easy implementation of access control and further improves convenience via custom URLs.

Additional information about system requirements and installation, as well as a free webcast can be found in the reference section at the end of this document. [2] [3]



* project-specific folder (here: command implementation)
** entry-point definition (in artisan.php)

Figure 1: Structure of a Laravel Project

4. Realization

The "Web Portal for Political Memory" features all the routines for performing the complete process from data retrieval to judgment about the relevance of quotes without any human intervention whatsoever. Yet, it has been decided that the final judgment should be incumbent on the administrative user, whereas the program provides assistance in form of two scores computed for the quote. Additionally, suggestions are made for the quote's speaker and the political field it deals with, which can also be edited by the administrator.

The process is split into five unattended steps performed by command line programs, and the manually performed decision about relevance. (Figure 2)

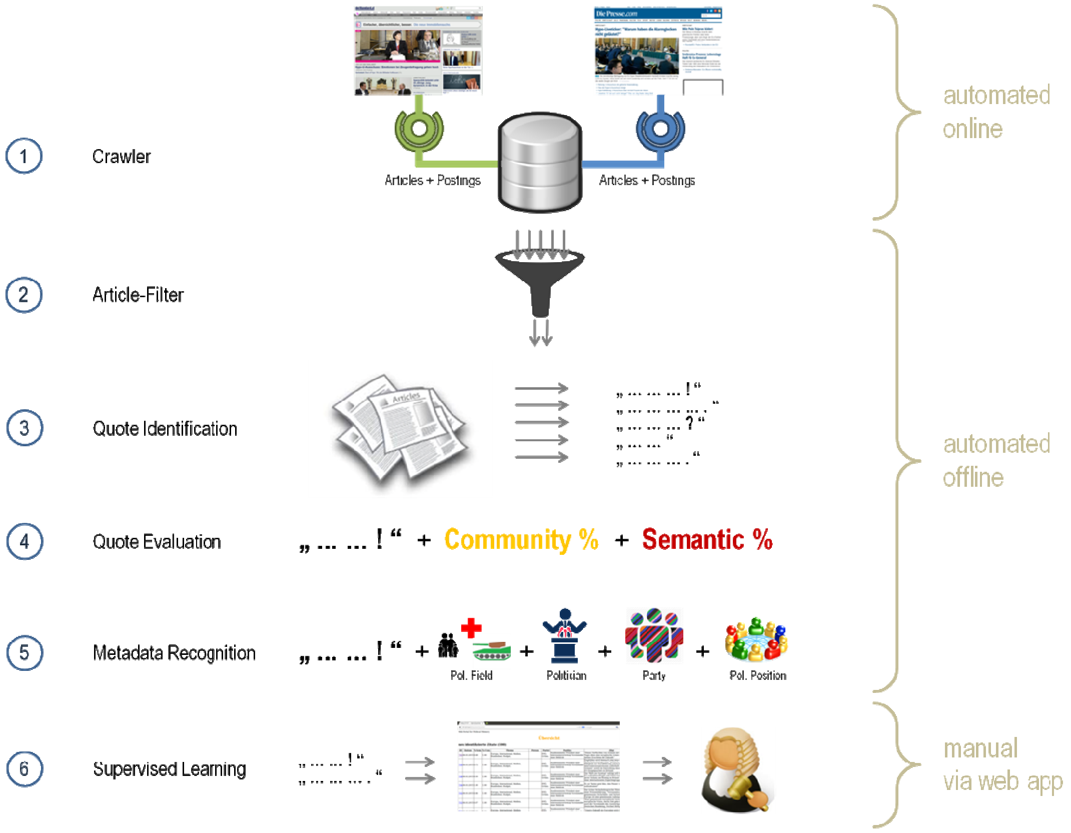


Figure 2: Web Portal for Political Memory - Process

Step 1: Crawler (Data Retrieval)

As a first step of the process, the raw data, of which quotes will be extracted, has to be gathered. Several providers have been in consideration being a potential source of information. *Notice: The following listing is valid for Austria, sources might have deviating characteristics in each country.*

- **Transcripts of parliamentary sessions:** The official transcripts provided by the web representation of the national houses of parliament would be the most reliable source possible. However, a long verification process leads to significant delay at publishing. [4] Session's summaries are also provided by the press service, but they are usually transformed into indirect speech and still extensive without any editorial filtering. [5]
- **ORF Teletext:** Teletext provides very compact up-to-date information but the absence of permanent links means the integrity of information cannot be proven permanently. [6]
- **APA OTS:** Via the Austrian Press Association's "Original Text Service", political parties may publish all kinds of messages. Since there is no adequate filtering, each sender would have to be subscribed separately and the list kept up-to-date. The number of publications via OTS is high. [7] Similar problems would arise for social media platforms like Twitter or Facebook.
- **Quality Newspapers:** The editorial staff ensures that the information content of all articles is very high. Information is up-to-date and key statements are mostly preserved in direct speech. Moreover, the community forum can be used for analysis of relevance.

Two popular newspapers have been chosen as data sources for the prototype, that are considered being comparatively factual despite having different political backgrounds: **"Die Presse"** [8] and **"Der Standard"** [9].

There is one crawler per data source which browses through the respective newspapers' politics sections and stores the articles and their reader's comments into the project's relational database. Entry points for the crawlers are the following URLs:

- http://diepresse.com/rss/Politik
- http://diepresse.com/rss/EU
- http://derstandard.at/?page=rss&ressort=inland

The command line program is started via: ***php artisan articles:update***

Step 2: Article Filter

Since the crawlers only gather information and do not analyze them, they might also collect political articles that are not within the scope of interest. The article filter scans the articles' full text for relevant tags such as abbreviations of local political parties and can easily be extended to searching for political positions and politicians existent in the database. Articles that do not contain any of the search terms are considered deletable. Additional criteria makes the filtering less restrictive.

Experience showed that scanning only for political parties leads to an acceptance rate below 50% and both, high precision (fraction of true positives) and recall (number of identified positives).

The command line program is started via: ***php artisan articles:check***

Step 3: Quote Identification

Articles identified as being relevant are scanned for direct quotes. At first, the articles formatting is removed. The pure text is then split into single sentences and analyzed. If quotes are encountered, the whole sentence is preserved and stored with the quote. If a quote spans over more than one sentence, each sentence is stored as own quote. A quote is defined as follows:

- A quote is always parenthesized by quotation marks
- A single parenthesized word is not considered a quote

Articles that do not contain quotes are considered deletable.

The command line program is started via: ***php artisan quotes:identify***

Step 4: Quote Evaluation

In order to support the accept/reject-decision, each quote is automatically analyzed. Target of the "Web Portal for Political Memory" is identifying quotes, that are either already remarkable or might have some importance in the future because of contradicting words or actions of the speaker.

The optimal way of measuring the social impact, would be a voting about the importance of every single quote by the community. However, something alike is usually not available. The closest equivalent present at practically each important news portal is a user comment section. The usage of direct quotes within user postings is relatively rare though. Matches between the quote's full text and a user comment have a high value. Positive rating signalizes that the discussed issue is interesting for other forum users as well. The general excitement of the community about an article can be measured by comparing its number of postings to the number at the most disruptive articles.

All those influences are combined into the "community score" of a quote as follows:

$$CS\ (q) = \frac{\ln ca_q}{\ln cmax_q} \times 0.7\ + \frac{(m_q + pr_q)}{\sqrt{cmax_q}} \times 0.7$$

$$Community\ Score\ (q) = \begin{cases} CS(q), & CS(q) \leq 1.0 \\ 1.0, & CS(q) > 1.0 \end{cases}$$

ca_q … number of comments of the article in which quote q has been identified
$cmax_q$ … highest number of comments of an article of the data source of the article of q
m_q … number of mentions of the direct quote q within the comments of its article
pr_q … sum of positive ratings for mentions of quote q

In most cases, the community score will only measure the response towards the article rather than the specific quote. Therefore, another measure is introduced which is independent of the respective community and deals with the linguistic and semantic features of a quote.

Research performed on the issue of semantic language analysis resulted in the conclusion, that there has not been found a reliable analytical way for deciding the question of relevance so far. As a consequence, a machine learning approach similar to those performed at big internet companies like "Google" has been chosen. [10]

$$Rsa(q) = \begin{cases} 0.1, & qas_q \leq 1\ OR\ qrs_q \leq 1 \\ \dfrac{qas_q}{qas_q + qrs_q}, & qas_q > 1\ AND\ qrs_q > 1 \end{cases}$$

$$W(q_w) = \begin{cases} Rsa(q), & q_w \notin wordlist \\ \dfrac{qa_w}{qa_w + qr_w}, & q_w \in wordlist \end{cases}$$

$$SS(q) = \frac{\sum_{i=1}^{n} W(q_i)}{n} / (Rsa(q) \times 2) + 0.2 \times t_q, \qquad t_q \in \{0,1\}$$

$$Semantic\ Score\ (q) = \begin{cases} SS(q), & SS(q) \leq 1.0 \\ 1.0, & SS(q) > 1.0 \end{cases}$$

qa_w … number of accepted quotes containing the w^{th} word of the quote
qas_q … number of accepted quotes of articles of the data source of the article of q
qr_w … number of rejected quotes containing the w^{th} word of the quote
qrs_q … number of rejected quotes of articles of the data source of the article of q
Rsa(q) … rate of quotes accepted of the data source of the article of q
t_q … indicator if the quote is present in the title of its article or not
$W(q_w)$ … rate of quotes accepted containing the w^{th} word of the quote

The semantic score is based on the experience, which words have been contained in quotes accepted in the past. The share of accepted quotes per word serves as a variable for calculating the score. The division by twice the acceptance rate of quotes of the respective data source is used for adjusting automatically to changing admission behavior and keeps the mean semantic score around 0.5 on the long run. A bonus of 0.2 points is granted if the quote is located in the headline of its article.

The command line program is started via: ***php artisan quotes:evaluate***

Step 5: Metadata Recognition

A very important function of an archive-type website is the searching for keywords. In order to later enable a search, the speaker of a quote, his/her political party, position and the issue that is talked about has to be stored in conjunction with the quote. The metadata recognition scans the complete article for matches with entries existing in the database. This procedure is called "list-based named entry recognition (NER)". The WP4PM allows the definition of multiple "tags" to entries of the mentioned kinds of metadata.

The command line program is started via: ***php artisan quotes:tag***

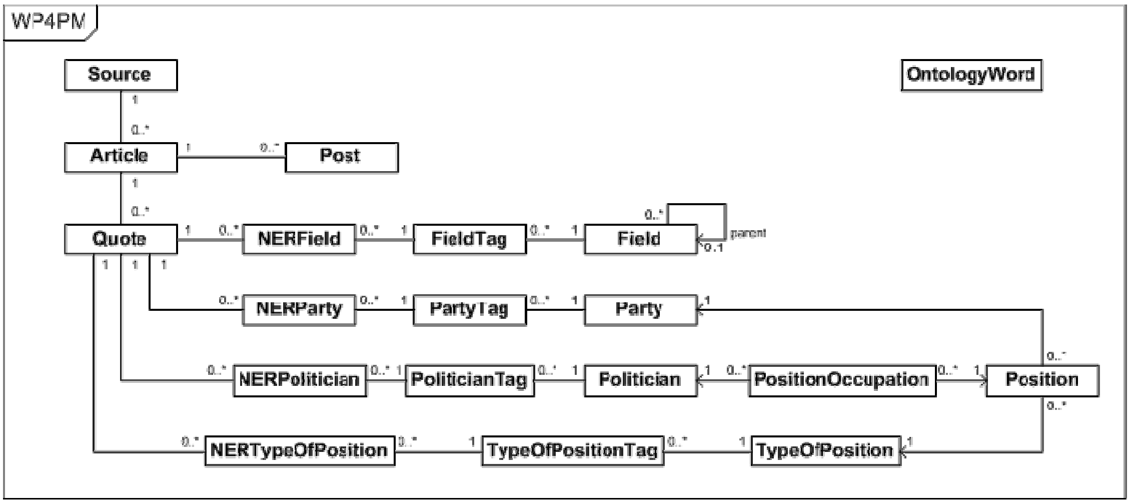
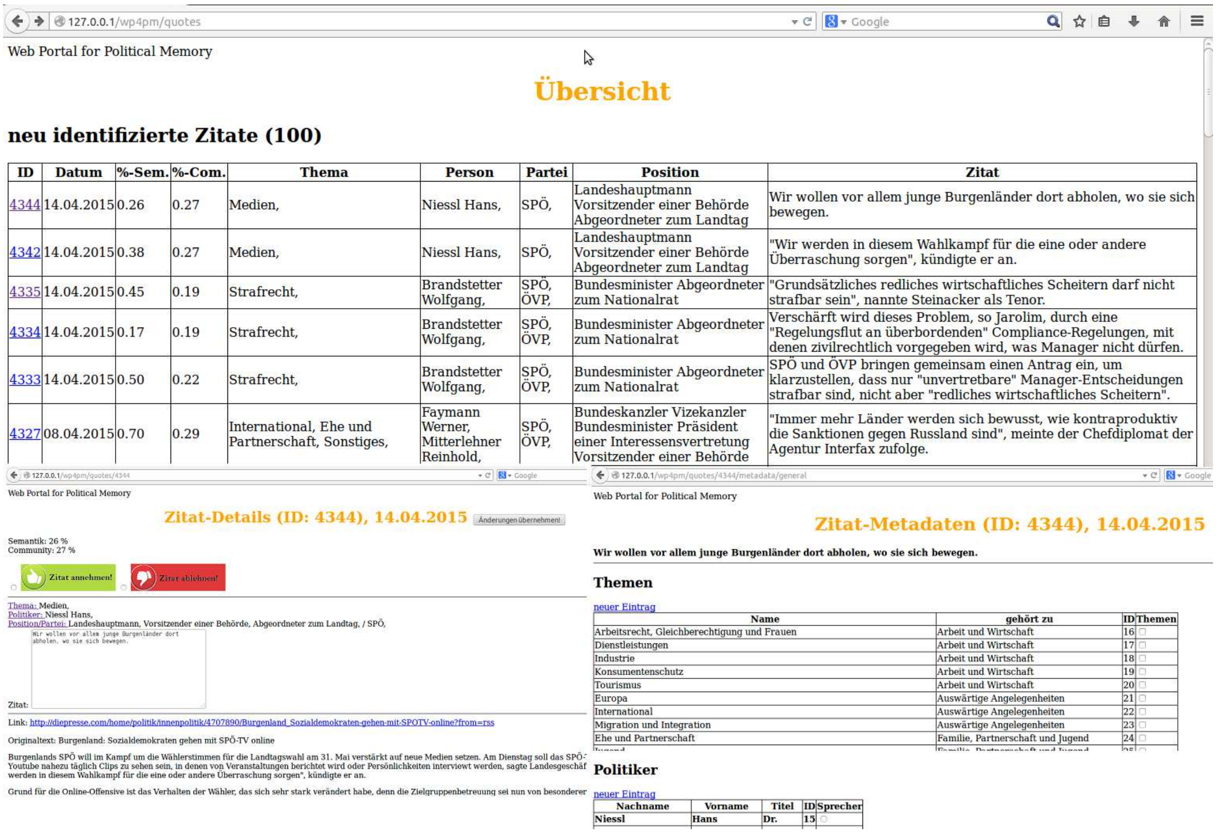


Figure 3: Web Portal for Political Memory - Domain (Model)

Step 6: Supervised Learning

The final decision about the relevance of a quote is incumbent to the administrative user. The data collected within Step 1 - 5 supports the decision making. Information is provided within a list of quotes on the main page of the web application. A click on the quote's id opens up the detailed information view. There, quotes can be edited, accepted and rejected. The view for editing the quote's metadata is accessible from this page as well. Each time, the user decides about a quote, the accepted- or rejected-counters of all its words within quotation marks are incremented by one. Moreover, the accepted- or rejected-counter of the data source is incremented by one and the quote's status set to "manually accepted" or "deletable", which removes it from the list on the main page. The changes of all those counters influence the semantic score of quotes evaluated in the future.



127.0.0.1/wp4pm/quotes

Web Portal for Political Memory

Übersicht

neu identifizierte Zitate (100)

ID	Datum	%-Sem.	%-Com.	Thema	Person	Partei	Position	Zitat
4344	14.04.2015	0.26	0.27	Medien,	Niessl Hans,	SPÖ,	Landeshauptmann Vorsitzender einer Behörde Abgeordneter zum Landtag	Wir wollen vor allem junge Burgenländer dort abholen, wo sie sich bewegen.
4342	14.04.2015	0.38	0.27	Medien,	Niessl Hans,	SPÖ,	Landeshauptmann Vorsitzender einer Behörde Abgeordneter zum Landtag	"Wir werden in diesem Wahlkampf für die eine oder andere Überraschung sorgen", kündigte er an.
4335	14.04.2015	0.45	0.19	Strafrecht,	Brandstetter Wolfgang,	SPÖ, ÖVP,	Bundesminister Abgeordneter zum Nationalrat	"Grundsätzliches redliches wirtschaftliches Scheitern darf nicht strafbar sein", nannte Steinacker als Tenor.
4334	14.04.2015	0.17	0.19	Strafrecht,	Brandstetter Wolfgang,	SPÖ, ÖVP,	Bundesminister Abgeordneter zum Nationalrat	Verschärft wird dieses Problem, so Jarolim, durch eine "Regelungsflut an überbordenden" Compliance-Regelungen, mit denen zivilrechtlich vorgegeben wird, was Manager nicht dürfen.
4333	14.04.2015	0.50	0.22	Strafrecht,	Brandstetter Wolfgang,	SPÖ, ÖVP,	Bundesminister Abgeordneter zum Nationalrat	SPÖ und ÖVP bringen gemeinsam einen Antrag ein, um klarzustellen, dass nur "unvertretbare" Manager-Entscheidungen strafbar sind, nicht aber "redliches wirtschaftliches Scheitern".
4327	08.04.2015	0.70	0.29	International, Ehe und Partnerschaft, Sonstiges,	Faymann Werner, Mitterlehner Reinhold,	SPÖ, ÖVP,	Bundeskanzler Vizekanzler Bundesminister Präsident einer Interessensvertretung Vorsitzender einer Behörde	"Immer mehr Länder werden sich bewusst, wie kontraproduktiv die Sanktionen gegen Russland sind", meinte der Chefdiplomat der Agentur Interfax zufolge.

127.0.0.1/wp4pm/quotes/4344

Web Portal for Political Memory

Zitat-Details (ID: 4344), 14.04.2015 Änderungen übernehmen!

Semantik: 26 %
Community: 27 %

○ Zitat annehmen! ○ Zitat ablehnen!

Thema: Medien,
Politiker: Niessl Hans,
Position/Partei: Landeshauptmann, Vorsitzender einer Behörde, Abgeordneter zum Landtag, / SPÖ,

Zitat: Wir wollen vor allem junge Burgenländer dort abholen, wo sie sich bewegen.

Link: http://diepresse.com/home/politik/innenpolitik/4707890/Burgenland_Sozialdemokraten-gehen-mit-SPOTV-online?from=rss

Originaltext: Burgenland: Sozialdemokraten gehen mit SPÖ-TV online

Burgenlands SPÖ will im Kampf um die Wählerstimmen für die Landtagswahl am 31. Mai verstärkt auf neue Medien setzen. Am Dienstag soll das SPÖ-
Youtube nahezu täglich Clips zu sehen sein, in denen von Veranstaltungen berichtet wird oder Persönlichkeiten interviewt werden, sagte Landesgeschäf
werden in diesem Wahlkampf für die eine oder andere Überraschung sorgen", kündigte er an.

Grund für die Online-Offensive ist das Verhalten der Wähler, das sich sehr stark verändert habe, denn die Zielgruppenbetreuung sei nun von besonderer

127.0.0.1/wp4pm/quotes/4344/metadata/general

Web Portal for Political Memory

Zitat-Metadaten (ID: 4344), 14.04.2015

Wir wollen vor allem junge Burgenländer dort abholen, wo sie sich bewegen.

Themen

neuer Eintrag

Name	gehört zu	ID	Themen
Arbeitsrecht, Gleichberechtigung und Frauen	Arbeit und Wirtschaft	16	☐
Dienstleistungen	Arbeit und Wirtschaft	17	☐
Industrie	Arbeit und Wirtschaft	18	☐
Konsumentenschutz	Arbeit und Wirtschaft	19	☐
Tourismus	Arbeit und Wirtschaft	20	☐
Europa	Auswärtige Angelegenheiten	21	☐
International	Auswärtige Angelegenheiten	22	☐
Migration und Integration	Auswärtige Angelegenheiten	23	☐
Ehe und Partnerschaft	Familie, Partnerschaft und Jugend	24	☐
Jugend	Familie, Partnerschaft und Jugend	25	☐

Politiker

neuer Eintrag

Nachname	Vorname	Titel	ID	Sprecher
Niessl	Hans	Dr.	15	○

Figure 4: WP4PM- Web Application (Top: Main Page, Left: Quote Details, Right: Quote Metadata)

5. Findings

In order to discuss the project's findings, it is important to define "relevance" of a quote in the given context. Considered being "relevant" are all the statements contributing to the shaping of the profile of a political player in terms of political agenda, credibility or style and therefore have the potential of significantly influencing voting decisions. Special attention has consequently been paid to:

- declarations/clarifications of a political standpoint (especially in case of inconsistencies)
- oppositions to common body of knowledge (e.g. knowledge gaps)
- violations of commonly agreed code of conduct (e.g. defamation)

The expectation has been that there exist strong (e.g. "absolut", "Schande") or definitive (e.g. "niemals", "immer", "kein", "sicher") words which generally increase the expressiveness (and therefore relevance) of a statement. This thesis could not be verified, mainly due to the inverting effect that negations have on such terms, like the example "man kann niemals absolut sicher sein" illustrates. One conclusion that can be drawn from such constructs is, that an effective semantic analysis should be able to determine if a statement is affirmative or restricting.

Another reason why words expected to be strong did not lead to a higher accept-reject-ratio was, that they were mainly used for criticizing the work of political contenders, often in (evidentially) exaggerated or polemic fashion. Own ideas, in contrast, have been expressed more considerate. Hence, a further asset would be functionality to detect if a speaker is referencing to own or foreign agenda.

Analysis of the used verb forms might serve as a factor as well. At least the usage of subjunctives seems to have an alleviative effect. For instance, quotes containing the indicative "ist" have been accepted 5 out of 20 times, whereas quotes containing the conditional form "wäre" have been accepted 0 out of 4 times.

The chosen machine learning approach for the semantic score leads to more stable results the more data is collected. Because of the discussed obstacles, it is expected that all accept-reject-ratios of the single words will converge towards a similar value around the overall accept-reject-ratio (i.e. are normally distributed with small variance). If the assumption that there exist words with a generally higher relevance than others were true, they could be easily identified with the chosen approach. The data collected in the course of this project is not sufficient for a statistically significant statement.

6. Conclusion

In the internet age, information is available about nearly everything. The challenge is filtering the most relevant pieces and archiving them, so they can be found when of use. Since there is no universal definition of relevance, a pure analytical way of measuring it, is out of reach. With the "Web Portal for Political Memory", a learning system has been introduced with the goal of automatically providing search, filtering and archiving in the domain of political discussion. As data sources, popular quality newspapers have been chosen.

For determining the relevance of political statements, two independent scores are calculated. The semantic score uses data of past decisions about relevance for its calculations, following the idea that the choice of words indicates a statement's level of significance. With the amount of data collected, this assumption could not be verified, but some criteria identified that might also signalize high information content. It looks like relevance and impact of a statement do only to a small extent depend on its usage of words. Much rather, significance arises out of the context and may even result from irony, sarcasm and polemics.

Semantic analyzers of the future will have to combine different approaches. In cases where linguistic analysis alone does not do the trick and not enough information about the historical viewpoint of the speaker towards the discussed issue is available, monitoring the community reaction is probably the most effective way of determining the relevance. This is the idea behind the community score which is a measure for the direct social impact of both: the quote and the article it has been extracted from.

Still the Web Portal for Political Memory is far away from making reliable decisions automatically. After sufficient observation though, could be analyzed, whether an ontology of "strong" words exist, that brings us closer to do so. Should there always be a user required for making a final decision, it is still a valuable orientation guide because democratic discourse requires the availability of information to the public.

7. References

[1] Bachelor thesis task definition: https://www.ims.tuwien.ac.at/topics/288

[2] Laravel setup information: http://laravel.com/docs/5.0

[3] Webcast "Laravel from Scratch": https://laracasts.com/series/laravel-from-scratch

[4] Stenographic protocols of parliamentary sessions: http://www.parlament.gv.at/PAKT/STPROT/

[5] Parlamentskorrespondenz (parliamentary press service): http://www.parlament.gv.at/PAKT/PR/

[6] ORF Teletext: http://teletext.orf.at/

[7] APA OTS: http://www.ots.at/

[8] Daily newspaper "Die Presse": http://diepresse.com

[9] Daily newspaper "Der Standard": http://derstandard.at

[10] Peter Norvig, "How Computers Learn":
https://www.youtube.com/watch?feature=player_detailpage&v=T1O3ikmTEdA#t=813