Characterizing Regulatory Documents and Guidelines based on Text Mining
(Supplementary Material)

Karolin Winter[1], Stefanie Rinderle-Ma[1], Wilfried Grossmann[1], Ingo Feinerer[2], Zhendong Ma[3]

[1] Faculty of Computer Science, University of Vienna, Vienna, Austria
{karolin.winter, stefanie.rinderle-ma, wilfried.grossmann}@univie.ac.at
[2] University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria
ingo.feinerer@fhwn.ac.at
[3] Center for Digital Safety & Security, Austrian Institute of Technology, Vienna, Austria
zhendong.ma@ait.ac.at

Abstract. The following case study is added as supplementary material for the Paper "Characterizing Regulatory Documents and Guidelines based on Text Mining" submitted to CoopIS17.

1 Description of Privacy documents

A preliminary study [1] applied content analysis to a text corpus consisting of 5 technical documents on privacy in video surveillance i.e., the EDPS Video-Surveillance Guidelines (EDPS)[4], the OECD Privacy Guidelines (OECD)[5], the Guidelines for Public Video Surveillance (Video)[6], the Data protection and privacy ethical guidelines (Data)[7], and the Operational Guidance on taking account of Fundamental Rights in Commission Impact Assessments (Guide)[8].

1. EDPS contains suggestions on the design and operation of video-surveillance systems for European Institutions and bodies.
2. Document OECD addresses privacy protection concerning the exchange of personal data within OECD countries.
3. Video describes guidelines for video surveillance of public places.
4. In Data ethical guidelines concerning research proposals are outlined.

[4] https://secure.edps.europa.eu/EDPSWEB/webdav/shared/Documents/Supervision/Guidelines/10-03-17_Video-surveillance_Guidelines_EN.pdf

[5] http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

[6] http://www.constitutionproject.org/wp-content/uploads/2012/09/54.pdf

[7] http://ec.europa.eu/research/participants/data/ref/fp7/89827/privacy_en.pdf

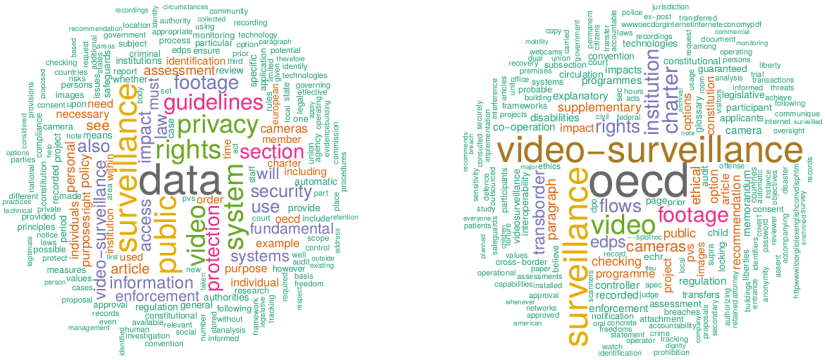
[8] http://ec.europa.eu/justice/fundamental-rights/files/operational-guidance_en.pdf

5. Providing advice for assessing the impact of the Charter of Fundamental Rights is the aim of the Guide.

The number of pages in the selected document set ranges from 18 (Data) to 104 (Video) and it is the basis for the second case study (cf. Sect. 2).

2 Case Study 3: Privacy Documents

Privacy documents, described in 1, are the subject of the following evaluation. Like in the first case study on ISO documents a corpus *PrivacyAll* containing all 5 documents is constructed and preprocessed. Frequent terms using *weightTf* as well as *weightTfIdf* are determined in a **1st analysis** step. The results are displayed in Figure 1a and 1b.



(a) Word cloud for *privacyAll*, weightTf (b) Word cloud for *privacyAll*, weightTfIdf

Fig. 1: Unigram word clouds for *privacyAll*

Emerging unigrams are

- `Data, surveillance, rights, privacy, system, protection, oecd, video, institution` and
- `video surveillance, public video, fundamental rights, data protection, impact assessment, personal data, law enforcement, privacy framework, ethical guidelines` as well as `surveillance systems` are significant bigrams.

These terms describe the topics present in the selected documents (cf. Sect. 1).

Figure 2 and 3 indicate that the Guide (grey) treats a different topic than the other four documents because they have not so many terms in common. For the Video document (orange) and the EDPS (red) the opposite holds. This observation should be reflected by the clustering.

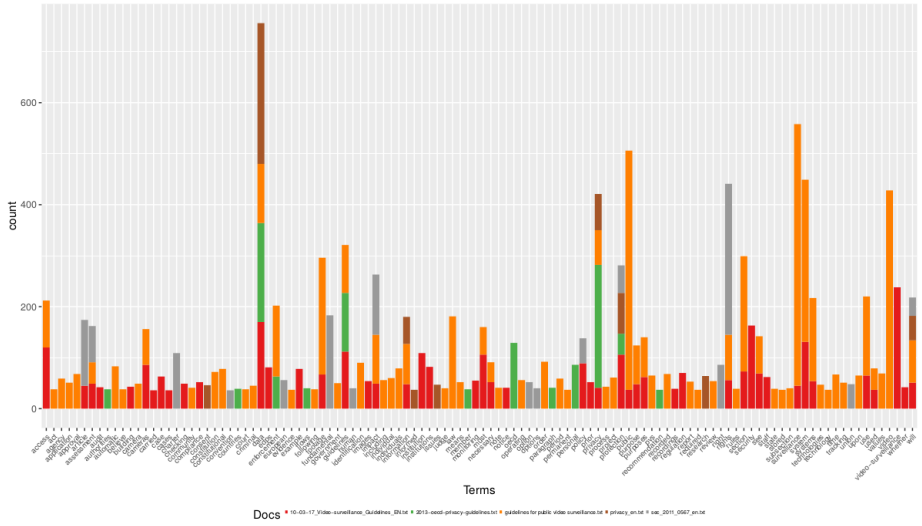


Fig. 2: Histogram for *privacyAll*, weightTf

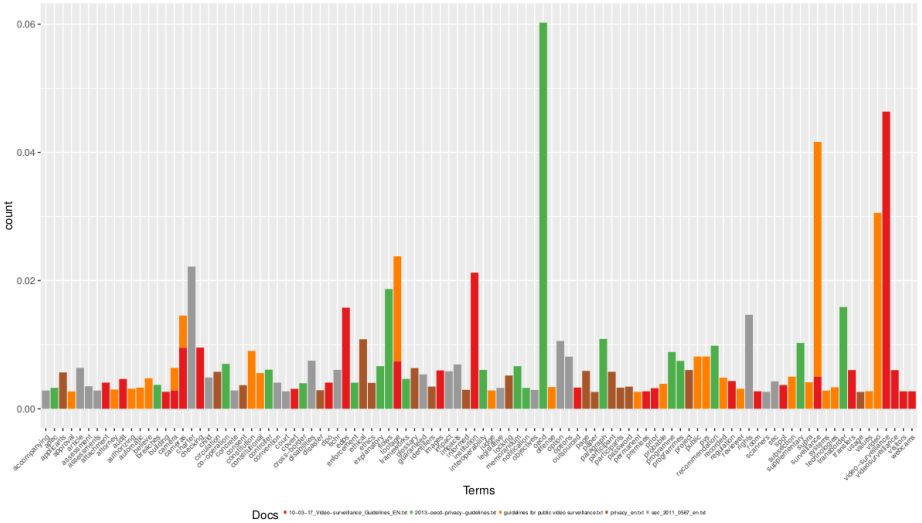


Fig. 3: Histogram for *privacyAll*, weightTfIdf

Following the methodology (**data preparation**) all documents are fragmented resulting in 48 partial documents (contained in the corpus *privacySections*). The fragmentation level was chosen as outlined in the paper (cf. Sect. 2), e.g., the EDPS was fragmented based on its section structure.

The **model building** bases on the elbow plot for *privacySections*, cf. Figure 4. According to this plot and the previous results emerging from the histograms, $k = 6$ was selected for k-means. Significant sentences are determined for each cluster by using unigram and bigram wordlists (**wordlist generation**, **sentence extraction**). Like in the first case study, one cluster is randomly picked to show the feasibility of the approach.

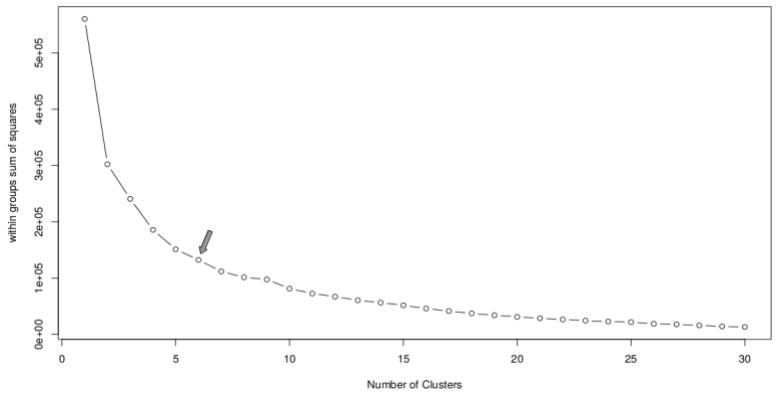


Fig. 4: Elbow plot for *privacySections*

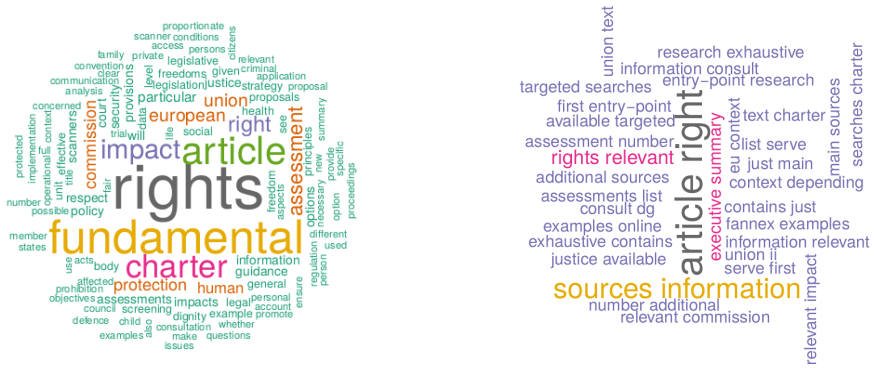
Example Cluster (Privacy) ($\mathcal{ECP}$): This cluster contains 6 fragments. In Figure 5 two out of four associated word clouds (**2nd analysis**) are depicted.

According to these plots the terms `rights, fundamental, article, impact, human, protection, sources, executive, fundamental rights, impact assessment, european union` are most frequent in the fragments of cluster $\mathcal{ECP}$. So these fragments seem to treat legal aspects of privacy and their impact on the society and companies.

All fragments, i.e., `1.key context and background, 2.operational guidance how to address fundamental rights step-by-step in commission impact assessments, 3.executive summary of the impact assessment, annex i, annex ii, introduction`, stem from the Guide document. This is basically the whole document which confirms the observation emerging from the histograms in Figure 2 and 3 that the topic of this document is different than the others. The derived subjects correspond to the document description in Sect. 1.

For this cluster, the following **wordlists** were created:

– Unigrams: `article, assessment, charter, fundamental, impact, right, rights, sources`



(a) Unigram word cloud for $\mathcal{ECP}$, weightTf

(b) Bigram word cloud for $\mathcal{ECP}$, weightTfIdf

Fig. 5: Word clouds for $\mathcal{ECP}$

– Bigrams: `fundamental rights`, `impact assessment`, `article right`, `executive summary`, `rights relevant`, `sources information`.

Extracted **significant sentences** are e.g.

– "While recognising the principle of transparency, the court considered that the contested provisions disproportionally interfered with the fundamental right to protection of personal data and to private life as provided for by articles 7 and 8 of the charter."
– "A possible negative impact deriving from the increased role of the victim in criminal proceedings could accrue, if this strengthened role were to endanger the defendants procedural rights, in particular the right to a fair trial (article 47 eu charter) and the right of defence (article 48 of eu charter)."
– "An initial screening of fundamental rights aspects should first check whether absolute rights are likely to be affected, as any objectives or options that violate such rights should be avoided from the very beginning (see 2.2.b and 2.4.a)."

These results were compared to the ones stemming from a (manual) content analysis (cf. [1]). In this paper the codes for process elements, i.e., the important parts for implementation, are structured as the sentences discovered by our methodology. Almost each sentence contains a "should + verb part" as well as actors. Since this structure can also be observed for the security documents in the paper we see this as an indicator that the approach assists in understanding and implementing guidelines.

References

1. Rinderle-Ma, S., Ma, Z., Madlmayr, B.: Using content analysis for privacy requirement extraction and policy formalization. In: 6th International Workshop on Enterprise Modelling and Information Systems Architectures (2015)