Legal Knowledge and Information Systems
A. Wyner and G. Casini (Eds.)

doi:10.3233/978-1-61499-838-9-165

Automatic Detection of Significant Updates in Regulatory Documents

Kartik Asooja, Oscar Ó Foghlú, Breiffni Ó Domhnaill, George Marchin, Sean McGrath
firstname.lastname@propylon.com
Propylon Ltd., Dublin 14, Ireland

Abstract. Regulations and legislations are regularly updated, which significantly burdens up the lawyers and compliance officers with a firehose of changes. However, not all changes are significant, and only a percentage of them are of legal importance. This percentage can certainly vary in different types of regulations. This paper focuses on automatic detection or ranking of meaningful legal changes, and presents a preliminary approach based on machine learning for the same, in the domain of Internal Revenue Code (IRC) related regulatory documents. Such system would provide the users with a means to quickly identify significant legal changes.

Keywords. Change detection, Version, Regulation, Regulatory Change Management, Machine Learning

1. Introduction

Lawyers, tax professionals, and compliance officers need to efficiently research and understand constantly changing regulations in order to competently respond to the updates and understand what their client/industry needs to comply with. The volume and velocity of changes and updates of laws and regulations are growing dramatically, which makes it even more difficult for professionals [5]. Following this, the legal publishers operate in a competitive, constantly changing environment where technology is supplying the unique proposition to many content products. Legal research often requires the study of the change timeline of the regulatory documents, especially in the cases of litigations, where one might need to study point-in-time changes in the regulatory framework. However, picking up the significant material or legal changes in a version history can be really expensive and cumbersome, as there can be good number of versions just accounting for changes in text formats, spellings, etc.

In this paper, we present a machine learning based approach to automatically detect the versions with significant material changes. Environmental Data and Governance Initiative (EDGI)[1] monitors government webpages to track environment related regulatory changes, and they are also doing a highly relevant project which aims at automatically identifying and prioritizing those changes[2].

[1] https://envirodatagov.org/website-monitoring/
[2] https://github.com/edgi-govdata-archiving/web-monitoring

2. TimeArc

Propylon's[3] TimeArc[4] platform is a legal research software solution that enables easy understanding of changes in legislation and regulation with the help of version timeline, with redlining comparison capabilities and point-in-time hyperlinking. It enables you to see the most up-to-date information available as well as a historical view, with access to a full revision history of every change ever made to any given document. This allows the user to find and compare historical changes, discover intent, and filter by commentary, context, and editorial overview. All of the information is available in an easy-to-use, efficient, browser-based tool that gives new and enhanced insights, as shown in Figure 1.

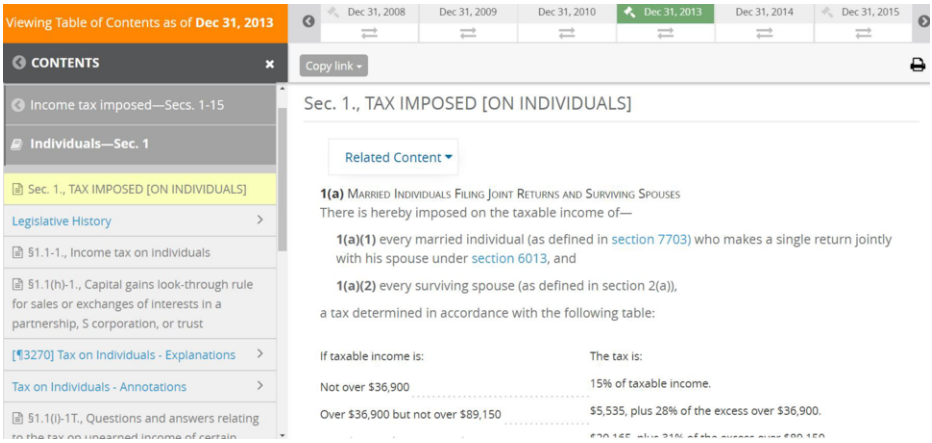

Figure 1. Sample Document Versions on TimeArc Platform. Versions with *gavel* show the significant ones.

However, TimeArc is designed to pick up even the smallest of changes in documentation and present them on the timeline in order to ensure that strict audit trails are maintained. This can result in the plotting of potentially insignificant changes on document timelines. Therefore, in this work, we focus on highlighting the significant changes in the document version timeline.

.01 Amended by P.L. 113-295 (deadwood amendment), P.L. 112-240, P.L. 110-185 (conforming amendment), P.L. 110-28, P.L. 109-222, P.L. 108-357 (conforming amendments), P.L. 108-311, P.L. 108-27, P.L. 107-16, P.L. 106-554 (conforming amendment), P.L. 105-277 (technical corrections), P.L. 105-206, P.L. 105-34, P.L. 104-188, P.L. 103-66, P.L. 101-508, P.L. 101-239, P.L. 100-647, P.L. 99-514, P.L. 97-448 (clarification, not an amendment), P.L. 97-34, P.L. 95-600, P.L. 95-30, P.L. 91-172, P.L. 89-809 and P.L. 88-272. For details, see the Code Volumes.

Figure 2. Sample 1: Human commentary section for a document

[3]https://www.propylon.com/
[4]https://www.propylon.com/legal-research/

.01 Historical Comment: Proposed 7/13/55. Adopted 2/3/56 by T.D. 6161. Amended 5/24/71 by T.D. 7117, 12/20/74 by T.D. 7332, and 4/4/2008 by T.D. 9391. [Reg. §1.1-1 does not reflect P.L. 95-600 (1978), P.L. 97-34 (1981), P.L. 97-488 (1983), P.L. 99-514 (1986), P.L. 100-647 (1988), P.L. 103-66 (1993), P.L. 107-16 (2001), P.L. 108-27 (2003), P.L.108-311 (2003) or P.L. 112-240 (2013). See ¶3260.045 et seq. and ¶3270.01].

Figure 3. Sample 2: Human commentary section for a document

3. Data

Definition of a significant change in a document can be highly contextual depending on the user and domain. Our use case deals with tax professionals and documents related to the US IRC and Treasury Regulations, provided by a legal publisher. The documents have a parallel human commentary, as shown in the figures 2 and 3. It summarizes the changes in a version for a section of data by giving citations to related Public Laws (P.L.) and Treasury Decisions (T.D.). Based on the suggestions from subject matter experts, we consider a change in a document as significant if there is a change in the citations within this commentary, especially to the ones related to Public Laws and Treasury Decisions. This implies that if there is a relevant regulatory change in P.L./T.D.s, it requires the publisher to make major amendments in related regulatory documents. However, still there can be some potential significant updates to the regulatory documents, which are not dependent on the changes in the P.L./T.D.s. For the experiments reported here, we do not consider these versions as significant, as it would require expensive tagging by subject matter experts.

We use the documents from 2005 to 2015 for our experiments. Total number of versions over all the documents in the data is 41,965, out of which only 24,839 ($\sim$ 59%) are significant, as per the above definition.

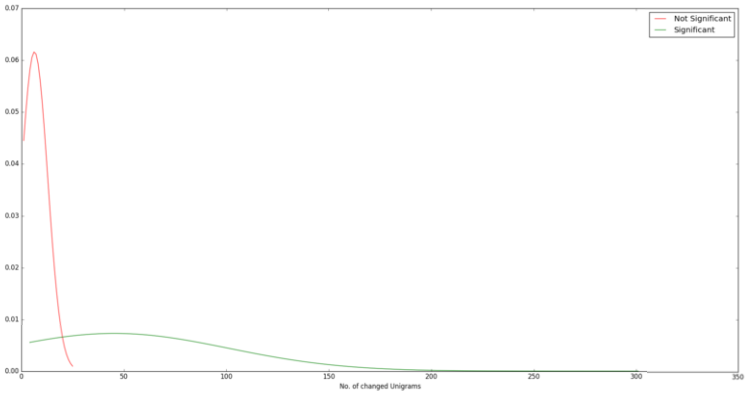

Figure 4. Data distribution (assuming Normal) against the number of unigrams in the symmetric difference, after removing 5% tail outliers.

4. Learning to Identify Significant Changes

As we have human provided commentary on the versions for only a section of our overall data from the client, the aim of this work is to learn and evaluate a machine learning based approach to automatically identify significant legal material changes for other documents in the same domain. This can be considered as a binary text classification problem, where the classes are significant change (positive class) and insignificant change (negative class).

We assume D_t^i and D_{t+1}^i represent the bag of words (BoW) sets present in the i-th document at time t and t+1 respectively. We use the symmetric difference between the BoW sets of the documents at consecutive time steps to define a single data instance, thus considering the correlation between the added and deleted words to the significance of a document revision. Therefore, the training data instance takes the following form: $(D_t^i \triangle D_{t+1}^i, y)$, where, $D_t^i \triangle D_{t+1}^i$ represents the symmetric difference between the versions, and y represents a boolean label for the class.

We consider unigrams and their counts as the features, and use Support Vector Machines (SVM) classification algorithm. Two methods were employed to evaluate our classifier: 10 fold cross validation, and 70:30 split of the total data as train and test datasets. We employ the LibSVM library [1] for SVM classifier using Weka machine learning toolkit [2]. The parameters for SVM classifier are as follows: SVM type = C-SVC, kernel function = Radial Basis Function. Features are ranked using the feature selection algorithm InfoGain [3]. Table 1 summarizes the results for the identification of the positive class. We can see that just using the count of unigrams as features can model a good predictor of the significant changes. This follows the data distribution graph shown in the figure 4, implying that if there are many added or deleted unigrams in a version, it leads to a significant version. Moreover, with unigram features, the classification improves significantly in comparison to just using the counts.

Feature	**Precision**		**Recall**		**F-measure**	
	10-fold	Train-test	10-fold	Train-test	10-fold	Train-test
Unigrams	0.925	0.912	0.910	0.890	0.917	0.901
Unigrams changed count	0.738	0.734	0.729	0.725	0.731	0.727

Table 1. Classification performance (weighted avg. metrics)

5. Conclusion

In this work, we present a preliminary approach to automatically mine the significant versions in a document timeline. Initial results from the classifier show a good performance, which can clearly enable the user to easily focus on the significant changes. The significant versions in the document timeline are highlighted in the TimeArc platform allowing the users to quickly navigate between meaningful changes in the law without seeing editorial, typographical or stylistic changes to content, as shown in demo in the figure 1. As future work, we would work on improving the classification performance by identifying more features for this problem, and by using deep learning algorithms.

References

[1] Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), p.27.

[2] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp.10-18.

[3] Mitchell, T.M., 1997. Machine learning. WCB.

[4] Akbani, R., Kwek, S. and Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. Machine learning: ECML 2004, pp.39-50.

[5] Asooja, K., Bordea, G., Vulcu, G., OBrien, L., Espinoza, A., Abi-Lahoud, E., Buitelaar, P. and Butler, T., 2015. Semantic annotation of finance regulatory text using multilabel classification. LeDA-SWAn (to appear, 2015).