DEPARTMENT OF COMPUTER SCIENCE

Natural Language Processing in the Law Domain

Oliver Ryan-George

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering.

Friday 22nd February, 2019

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of BSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Oliver Ryan-George, Friday 22nd February, 2019

Acknowledgements

Executive Summary

This project is a contribution to a law academic's research into the growing relationship between intellectual property law and human rights law; in particular, the extent to which intellectual property laws involve human rights considerations, and their balance between consideration of creators and users.

The first technical objective for the project is to use natural language processing to estimate measures of these two aspects for input law journals and statutes. This will involve using Python to explore a variety of machine learning models, starting with support vector machines, to optimise results. Classifications will be evaluated using F1-scores and compared against current research's achievements.

The second is to collaborate with the law acadmic iteratively to find the most appropriate way to visualise the results for the project's domain using 'matplotlib' for plotting and tKinter for the user interface. Success in this objective will be based upon typical human computer interaction practices.

Currently, the relationship between intellectual property and human rights is analysed on a case-by-case basis due to a lack of systematic means of analysis. The project will give that systematic method and therefore allow a more comprehensive argument to be made about the relationship.

The project is likely to be successful because there have previously been successfully projects that involved similar topic classification tasks. It will, however, be made individual from many of these by my iterative approach to the project, collaborating with a law academic to find the most appropriate way to classify and visualise the results.

Should it be successful, the project will further understanding of how past events have impacted intellectual property laws.

Contents

1 Background **1**
1.1 Introduction 1
1.2 Literature Review 2
1.3 Preliminary Investigation 4
1.4 Project Plan 4
1.5 Conclusion 5

Chapter 1

Background

1.1 Introduction

1.1.1 The Problem

Intellectual Property law is a term typically used to describe the areas of law which establish property protection over intangibles such as ideas, signs and information. This protection is in order to make the advancement of ideas profitable which therefore incentivises this act[1].

There is, therefore, a balance to be struck between limited exclusive rights and benefits to the public. While limiting exclusive rights may facilitate progress, benefiting the public, the overprotection of the exclusive property may restrict their access[2]. One example of this is the expansion of copyright terms such as the controversial Copyright Term Extension Act of 1998 which was heavily lobbied for by Disney just years before Mickey Mouse's copyright ran out[3]. The trend in extension of copyright terms, illustrated by Figure 1.1, illustrates the appearance that the balance is tipping towards exclusivity rights.

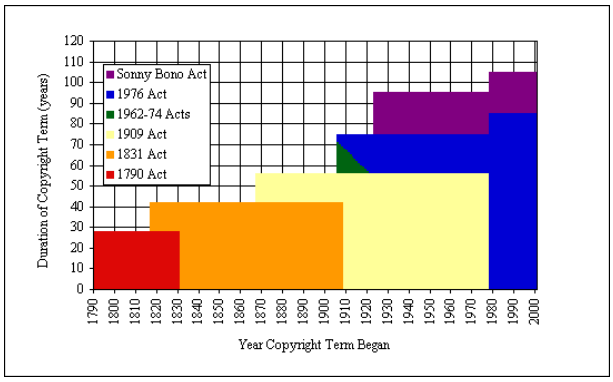

Figure 1.1: Expansion of U.S. copyright term lengths[4].

This implementation of intellectual property law brings in the question of human rights of access to culture, education and other social and economic rights, but traditionally this has not been included in discussion of intellectual property law[5]. However, recently scholarship and legislation has progrssively begun to incorporate each other's linguistics[6].

Dr. Megan Rae Blakely of Lancaster University is looking for assistance in analysing journals and legal instruments for overlap in languages. In this analysis, I will map the dynamics of this change in legal and social perspective, to make evident moments of significant shifts in language.

Previously, analysis of the intersection between human rights and intellectual property law, such as Helfer's[7], has been limited to more manual case-by-case methods. Supplying a more systematic method using natural language processing that can cope with large amounts of data would give concrete evidence of the relationship between human rights and intellectual proeprty.

1.1.2 Aims and Objectives

In the early stages of the project, the following requirements for the project were established:

- A model that takes a corpus of training PDF documents as input and assess the typical language characteristics of intellectual property documents and human rights documents and which indicates the extent to which each document suggests the law protects the user or the creator/owner;
- A visualisation of this model with the x-axis as time; the y-axis as the intent of the document toward human rights to intellectual property; and the z-axis as the measure of the extent to which the document suggests the protection of the creator/owner or the user;
- An easy to use Graphical User Interface which allows for input of law journals and treaties in PDF form and then outputs the updated visualisation;
- A code base written and documented well enough for any future researcher to easily understand and extend on.

Over the course of the literature review, section 1.2, I will review past literature in order to find the best methods to achieve each of these requirements.

1.1.3 Added Value

The originality of the project stems from its application of natural language processing methods in this domain, rather than the natural language processing methods used. The added value will be the unique way in which the findings are best illustrated for this application through the visualisation.

The outcome of the project will help add value to its domain. As the first systematic, easily scalable technology in this domain, it will help illustrate how historical changes in technology have impacted the tone of intellectual law. Blakely points out that this, in turn, will allow legal professionals to consider how future technological changes will impact their work and adapt accordingly.

1.2 Literature Review

1.2.1 Machine Learning Classifiers

Document topic classification is the automated assignment of natural language texts to predefined categories based on their content[8].

The precision/recall break-even point is a typical measure of success in topic classification[9]. It is obtained by modifying parameters until the precision and recall values are equal. However, this breakeven point does not always exist[10]. An alternative to this is F1-score which is calculated by multiplying two by the precision and recall and then deviding this by the sum of the precision and recall.

Support vector machines are a common feature-based approach to classification. They work by considering words occuring in the training set (alternatively, see section 1.2.2) as a feature and plotting each document in the feature space by the number of times each feature occurs in it. It then plots a hyperplane that best splits each document class. This led to an 86% precision/recall breakeven point when they classified the Reuters 'acq' dataset[11].

In contrast to support vector machines, Naive Bayes is a probabilistic approach based on Bayes' rule. Bayes' rule is a formula for finding the class given a document. There are many variations of this type of classification where the likelihood, representing the probability of a document given a class, differs. The Bernoulli multivariate model only considers whether a word occurs in a document while the multinomial model considers how many times a word appears in each document. Predictably, McCallum and Nigam found that the multinomial model performs consistently stronger over any significant vocabulary size. This led to a 89% precision/recall breakeven point when classifying the Reuters 'acq' dataset, slightly better than Support Vector Machine's 86%.

Wang and Manning improved the performance of both naive Bayes and support vector machines, for datasets with average wordcounts over 100, by adding naive Bayes features to support vector machines[12].

All these methods assume that each document can only be a member of one class. This is often not the case. Godbole and Sarawagi use the example that if a document is classed as being about wheat, it is also likely to be about grain[13]. This relationship between topics should be captured in a topic-topic distribution and took advantage of. Godbole and Sarawagi implement a few variations of this, all based around the standard support vector machines with one added feature per class representing how similar the given class is to all other classes. It leads to a slight improvement on F1-score across two datasets.

1.2.2 Language Features

The models from Section 1.2.1 mostly used single words as features. This is known as a bag of words model. This has the benefit of having dense feature spaces and therefore, higher computational efficiency but it is not representative of the actual meaning of text. Wallach points out that the phrases "the department chair couches offers" has a vastly different topic to "the chair department offers couches", yet a bag of words model will represent them as the same[14]. This sort of ambiguity would clearly give innacurate results so it is worth using a feature a bit more resource heavy that provides more accuracy.

The most common feature used is an n-gram model. This is where n words are taken as the feature, giving each word some ordered context. Tan et al use bigrams that pass a certain threshold to avoid the feature space being too sparce, combined with unigrams resulting in a 3% average F1-score improvement on the Reuters datasets[10].

1.2.3 Graphical Visualisation

What visualisations have been used for topic classification before?

What visualisations have been used in the law domain before?

1.2.4 Iterative Process

HCI - how to review in an iterative process

1.2.5 User Interface

Nielsen collated his heuristics from existing sets of heuristics with the aim of creating a general set which is as good as possible at explaining the usability prblems hat occur in real systems[15]. These have been highly regarded since he finalised them in 1994. They are as follows:

- The user must be kept informed of the system status;
- The system must match the real world with concepts familiar to the user;
- The user must have control over the system, especially with an easy emergency exit;
- The system should be consisent with its looks and actions;
- The system should do everything within its power to prevent errors;
- The system should make the user's options visible so they do not have to memorise them;
- The system should be flexible so novice users can use with ease but expert users can use hidden accelerators for more efficient use;
- The design should be aesthetic and minimalist since every piece of information is competing for the user's attention;
- The system should give error messages that are expressed in plain language, that precisely indicate the problem and that constructively suggest a solution;
- The system should provide help and documentation for the user in case anything about the system is unclear.

Nielsen also presents how to measure the severity of usability problems. The severity of a problem is split into three categories: the frequency with which a problem occurs which is how commonly it occurs; the impact of when it occurs which is how easy it is to work around; and the persistence of the problem which is how frequently a user will be bothered by the problem.

1.2.6 Documentation

Spinellis outlines the best practices for readable code[16]. Notably, he explains that the 'Don't Repeat Yourself' principle means to write what the code is doing and not how in this context as the code itself shows how an action is done. He also highlights how bad code cannot be rectified with lengthy documentation and the code should be rewritten instead.

Python documentation is standardised by the Python Enhancement Proposals. PEP 8 and PEP 257 are useful for the readability of Python code as they supply a style guide for the code and documentation respectively.

1.3 Preliminary Investigation

1.3.1 Investigation

word counts

word counts in sentences with words indicative of classes

1.3.2 Conclusions

learnt key information about dataset to guide decisions - 64 documents (27IP;37HR - vocabulary of over 20,000 words in 37 HR documents - average wordcount 3965

1.4 Project Plan

1.4.1 Timeline

Figure 1.2 is a plan of what actions should be carried out each week in order for the timely completion of the final deliverable. It is split into three columns, each representing a different key part of the project's source code that can be worked on somewhat independently.

Week	Model	Visualisation	User Interface
11/02	Establish specific requirements	Establish specific requirements	Establish specific requirements
18/02	SVM, bag of words model for IP/HR	Initial sketches and feedback	Initial sketches and feedback
25/02	Evaluation tools - algorithm to determine number of successful test cases	2D visualisation of HR-IP against time	Nothing
04/03	SVM, bag of words model for User/Creator	2D visualisation of HR-IP against User-Cretor	Display results as part of UI
11/03	Evaluation tools - cross validation algorithm	3D visualisation with time added as an axis	Add PDF selection from user interface
18/03	SVM, ngram models	Line of best fit	Add tabs for different visualisations
25/03	SVM/naive Bayes model	Futher tweaks based on feedback	Futher tweaks based on feedback
25/03	Multi-label models	Futher tweaks based on feedback	Futher tweaks based on feedback
01/04	Final checks and cleanup	Final checks and cleanup	Final checks and cleanup
08/04	Review of results	Review of results	Review of results

Figure 1.2: Week by week timeline of actions for project

1.4.2 Evaluation

Classification Results

Test samples appear in correct classification - aim for similar to paper that was used Cross validation produces consistent results

Visualisation and Usability

HCI process

1.5 Conclusion

References

[1] Lionel Bently, Bred Sherman, *Intellectual Property Law*, 4th ed. OUP Oxford, 2014, ISBN: 978-0199645558.

[2] Christophe Geiger, *Research Handbook on Human Rights and Intellectual Property*. Edward Elgar Publishing, 2016, ISBN: 978-1786433411.

[3] Victoria A. Grzelak, "Mickey Mouse & Sonny Bono Go To Court: The Copyright Term Extension Act and Its Effect On Current and Future Rights," *John Marshall Review of Intellectual Property Law*, vol. 2, no. 1, pp. 95–115, 2002.

[4] Tom Bell, *Trend of Maximum U.S. General Copyright Term*, tomwbell.com, 2008.

[5] Laurence R. Helfer, *Mapping the Interface Between Human Rights and Intellectual Property*. Cambridge University Press, 2011, ISBN: 978-0521711258.

[6] Megan R. Blakely, *BILETA 2019 Proposal*, Not published - available upon request, 2019.

[7] Laurence R. Helfer, "Human rights and intellectual property: Conflict or coexistence?" *Minnesota Intellectual Property Review*, vol. 5, no. 1, pp. 47–62, 2003.

[8] Febrizio Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[9] Claude Sammut, Geoffrey I. Webb, *Encyclopedia of Machine Learning*. Springer, Boston, MA, 2010, ISBN: 978-0-387-30164-8.

[10] Chade-Meng Tan, Yuan-Fang Wang, Chan-Do Lee, "The use of bigrams to enhance text categorisation," *Information Processing & Management*, vol. 38, no. 4, pp. 529–546, 2002.

[11] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European Conference on Machine Learning*, 1998.

[12] Sida Wang, Christopher D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Annual Meeting of the Association of Computational Linguistics*, ser. 50, vol. 2, 2012, pp. 90–94.

[13] Shantanu Godbole, Sunita Sarawagi, "Discriminative methods for multi-labeled classification," in *Advances in Knowledge Discovery and Data Mining*, ser. 8, vol. 1, 2004, pp. 22–30.

[14] Hannah M. Wallach, "Topic Modelling: Beyond Bag-of-Words," in *Journal for Language Technology and Computational Linguistics*, vol. 23, 2006, pp. 977–984.

[15] Jakob Nielsen, "Enhancing the Explanatory Power of Usability Heuristics," in *SIGCHI Conference on Human Factors in Computing Systems*, vol. 1, 1994, pp. 152–158.

[16] Diomidis Spinellis, "Code documentation," *IEEE Software*, vol. 27, no. 4, pp. 18–19, 2010.