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.

Tuesday 7th May, 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, Tuesday 7th May, 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

Chapter 1

Introduction

1.1 Chapter Overview

To open the thesis, this chapter gives a brief background of why the project is being pursued in the law domain. It continues by explaining the project's aims with regards to the given background and then finishes with how the aims intended to add value to the law domain once they had been met.

1.2 Project Context

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.

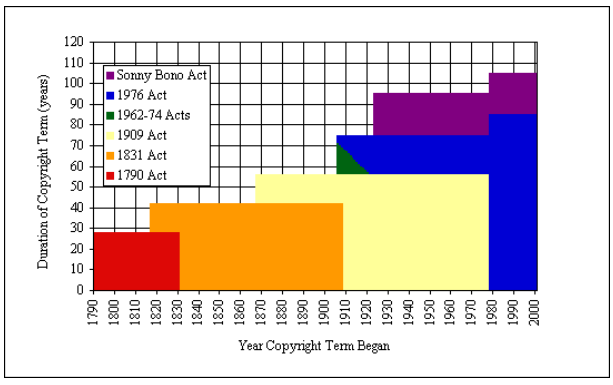

Figure 1.1: Expansion of U.S. copyright term lengths[4].

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. However, recently scholarship and legislation has progrssively begun to incorporate each other's linguistics[5].

Dr. Megan Rae Blakely of Lancaster University is looking for assistance in analysing journals and legal instruments for overlap in languages. As well as the human rights and intellectual property, Dr. Blakely hypothesised that it may be interesting to see how the language in these fields benefit the user and the creator of intellectual property[6]. 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.

1.3 Ground Truths

Dr. Blakely supplied a set of ground truths that she wished to analyse the language of. This consisted of the following:

- Four journals, shown in Table A.1: two that are on the topic of human rights and two that are on the topic of intellectual property;
- Four international treaties, shown in Table A.2: two that are on the topic of human rights and two that are on the topic of intellectual property;
- Four lists of words, shown in Table A.3: one that Dr. Blakely expects to see in documents about human rights, one that Dr. Blakely expects to see in documents about intellectual property, one that Dr. Blakely thinks indicates that a segment indicates benefits to the user and one that Dr. Blakely thinks indicates that a segment indicates benefits to the creator.

1.4 Requirements

In the early stages of the project, I worked with Dr. Blakely to establish a set of requirements for the project. These requirements can be split into three categories: natural language processing, visualisation, and usability.

Each of the following subsections discusses the requirements of a different category. Each subsection starts by specifying the deliverable elements for the category, followed by high-level criteria of how they will be evaluated. All goals needed to be completed by 16th April 2019, when Dr. Blakely presents to the British and Irish Law Education and Technology Association (BILETA) Conference.

1.4.1 Natural Language Processing

The natural language processing in the project will take the form of two models. Each model will take the same corpus of training PDF documents as input but will classify different characteristics of the corpus.

- The first model will use topic classification to identify whether a document is related more to human rights or intellectual property;
- The second model will use sentiment analysis to identify whether the document indicates that the current legal climate more strongly benefits the user of intellectual property of the creator of it.

It is notable that Dr. Blakely expressed that the former is the more important model in terms of the project's success as there is more scholarly work done on this relationship. The latter models' relationship is less widely covered and was requested by Dr. Blakely as a personal preference.

These models will be successful if they meet the following criteria:

- The model accurately represents a large proportion of the documents in the corpus with regards to its classification characteristics;
- The model deduces any trends in the corpus with regards to its classification characteristics, if those trends exist;

The model must only consider the language of the content of the article.

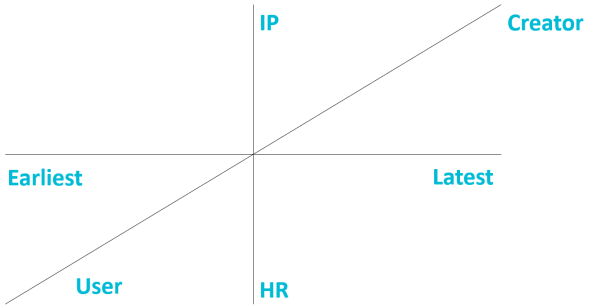

Figure 1.2: An adaptation of Eboch's tone graph for this domain.

1.4.2 Visualisation

The previously mentioned models will then have to be visualised. Dr. Blakely stated her belief that a modification of Eboch's tone graph[8] shown in Figure 1.2 would be a suitable visualisation for the models. This would involve an x-axis representing time, a y-axis representing human rights-intellectual property and a z-axis representing user-creator.

The visualisation will be successful if it meets the following criteria:

- The visualisation and its axes' meanings are self-explanatory;
- It is clear where a point lies on the visualisation's axes;
- The visualisation has all the possible information that Dr. Blakely requires, on it;
- The visualisation clearly indicates any trends deduced by the models;
- The visualisation is aesthetically pleasing.

1.4.3 Usability

There will be two key types of stakeholders in this project. Both their circumstances will have to be considered in order to deliver a suitable final product.

The first type of stakeholder are the users of the project's end product. This will primarily be Dr. Blakely but is likely to extend to other law academics. It is important to note that these will likely be people with standard information technology skills and no computer science skills. Therefore, the end product will need to be in the form of a tool that is accessible to people of this calibre.

The tool must have the following features:

- Allow for input of new documents;
- Display visualisations based on models with given documents as input;
- Display information about success of model.

The tool will be successful if it meets the following criteria:

- The tool does not require any prior Computer Science knowledge to setup or use;
- The tool has a graphical user interface which is intuitive;

The second type of stakeholder is the future developers furthering Dr. Blakely's research. These developers will be experts in natural language processing and computer science and so will be able to both use and change the tool to their liking. However, changes to the tool will only be easily changeable if the codebase for it is well written.

The codebase will be successful if it meets the following criteria:

- It is easy to understand the purpose of each section of code;
- It is easy to expand upon and adapt the overall codebase.

1.5 Added Value

This project is original in its application of natural language processing methods to find the relationship between intellectual property and human rights language. None of the individual computing methods used were original, but the added value appears through the iterative process which found the most appropriate visualisation of results for the domain. The outcome of the project will help add value to its domain by illustrating how historical changes in technology have impacted the tone of intellectual property 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.

Chapter 2

Literature Review

2.1 Chapter Overview

This chapter details the background research that was carried out in order to make informed decisions during the project.

2.2 Natural Language Processing

2.2.1 Evaluation of Binary Classification Algorithms

A binary classification model is an algorithm that sorts a set into two categories. One set is often referred to as 'positive' and the other 'negative'. It is important to validate a model in order to see whether its results can be trusted. Upon validating, if an item in the set has been correctly classified, it is considered a 'true' classification, otherwise, it is considered a 'false' classification.

The holdout method of validation involves training the model on a large proportion of the data, known as training data, and then classifying the rest of the data, known as test data, and giving the model a score based on how the test data was classified based on their actual class. This is limited as the test data set may be overfit to the model, making the model perform better for that test data than it would for new incoming data. Two common methods of solving this problem are cross-validation and bootstrapping.

Cross-validation is where the dataset is split into k segments. Each segment is treated a training data $k-1$ times and treated as test data once. This will give k different scores for the model. Meanwhile, bootstrapping is where n items in the dataset are picked uniformly at random. Each item is put back into the dataset once picked. This can be done several times to produce different scores for the model. In both cases, if the scores have a high average and low variance, the model is successful and not overfit to the data.

Kohavi compares these methods and recommends the use of cross-validation because bootstrap can be prone to large biases[9]. Kohavi found that the optimum number of folds depends on the stability of the dataset. He also notes that cross-validation's evaluation is most accurate when each fold is stratified to have a similar number of each class in it.

The simplest score that can be used for a model is accuracy. Accuracy is the ratio of correctly classified test documents to the total number of test documents. This is insufficient for imbalanced datasets[10].

A common solution to this in topic classification is to use precision, which is the ratio of true positives to the total number of positive classifications shown in Equation 2.1, and recall, which is the ratio of true positives to the total number of classifications made in total, shown in Equation 2.2. Precision and recall can be combined in two ways. The F1 score finds the harmonious mean between the two as shown in Equation 2.4. The precision/recall breakeven point is obtained by modifying the classifying threshold until the precision and recall values are equal[11]. These methods are most appropriate for information retrieval when it is important to evaluate true positives and true negatives are in excess[12].

$$precision = \frac{tp}{tp + fp} \tag{2.1}$$

$$recall = \frac{tp}{tp + fn} \tag{2.2}$$

type	score
accuracy	0.95
f_1 score	0
BEP	0
BACC	0.48

Table 2.1: Scores with a ground truth of 5 positives and 95 negatives and a prediction of 0 positives and 100 negatives.

type	score
accuracy	0.95
f_1 score	0.97
BEP	0.95
BACC	0.48

Table 2.2: Scores with a ground truth of 95 positives and 5 negatives and a prediction of 100 positives and 0 negatives.

$$tnr = \frac{tn}{tn + fp} \tag{2.3}$$

Velez suggests balanced accuracy, shown in Equation 2.6, as the solution. This is the average of the true positive rate, which is another name for recall shown in Equation 2.2, and the true negative rate, which is the ratio of true negatives to the total number of classifications made shown in Equation 2.3. This method is more appropriate for cases where it is important to evaluate both the true positives and the true negatives.

$$f_1 = 2 \times \frac{precision \cdot recall}{precision + recall} \tag{2.4}$$

$$BEP = precision = recall \tag{2.5}$$

$$balanced\text{-}accuracy = \frac{recall + tnr}{2} \tag{2.6}$$

Tables 2.1 and 2.2 show how the different scoring systems rate the performance of imbalanced datasets and algorithms. The algorithms both classify a given test item into a predetermined class, regardless of the item's contents. Balanced accuracy is the only scoring system to rank both algorithms poorly.

2.2.2 Topic Classification

Document topic classification is the automated assignment of natural language texts to predefined categories based on their content[11].

Support vector machines are a common feature-based approach to classification. They work by considering features, which are to be discussed in Section 2.2.4, 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, as exemplified by Figure 2.1. This led to an 0.86 precision/recall breakeven point when Joachims classified the Reuters 'acq' dataset[13].

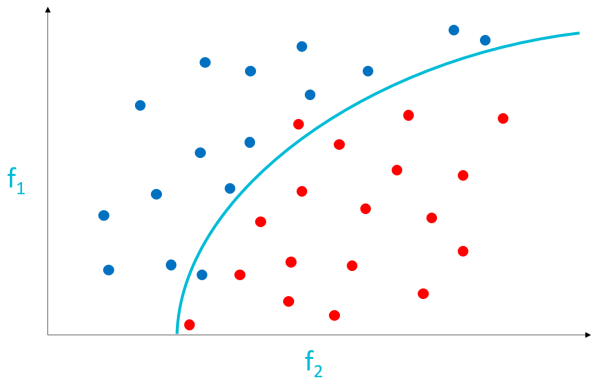

Figure 2.1: A simplified example of how Support Vector Machines work.

In contrast, Naive Bayes is a probabilistic approach based on Bayes' rule. Bayes' rule, shown in Equation 2.7, is a formula for finding the probability of class C_k, given a set of features x. There are many variations of this type of classification where the likelihood, representing the probability of a document given a class, diers. 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 0.89 precision/recall breakeven point when classifying the Reuters 'acq' dataset, slightly better than Support Vector Machines 0.86.

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})} \tag{2.7}$$

Wang and Manning improved the performance of both naive Bayes and support vector machines, for datasets with average wordcounts over 100, by adding elements of naive Bayes to support vector machines[14].

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[15]. This relationship between topics can be captured in a topic-topic distribution and used to enhance results. 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 of F1-score across two datasets.

2.2.3 Sentiment Analysis

The same classification techniques used in topic classification in Section 2.2.2 can be used for sentiment analysis as demonstrated by Moraes et al who experimented with SVM, naive Bayes and artificial neural networks. They classified reviews as positive or negative with a 0.87 accuracy score using neural networks on a movie reviews dataset[16]. Using these machine learning techniques, however, needs significant data to train on which is not always available.

An alternative method for sentiment analysis is a lexicon based approach. As a preliminary investigation, Pang et al obtained a list of positive keywords and a list of negative keywords for a different set through analysis of the term frequencies of a set of training data[17]. They then classified the test data based on whether they had more positive or negative keywords. This achieved an accuracy of 0.64, performing significantly worse than Moares et al at the same task.

Mandal and Gupta propose a more advanced lexicon based approach[18]. This involves a detecting comparative and superlative words and assigning a higher weighting when these precede a keyword. Most notably, it involves negating rating should negation words such as 'not' appear before a keyword. This increased Mandal and Gupta's accuracy from 0.87 to 0.97 on their manually created online review dataset.

2.2.4 Language Features

In natural language processing, each document is represented as a set of features where each feature is given a value. The simplest set of features a document can have is the set of unique words that appear in the document. This is known as the bag of words model. It has the benefit of having dense feature spaces, meaning that there is a low number of features which reoccur many times in the document. Dense feature spaces are more computationally efficient than sparse feature spaces. However, bag of words is limited in that these features are less capable of conveying the document's meaning. For example, Wallach points out that the phrase the department chair couches oers has a dierent topic to the chair department oers couches, yet a bag of words model will represent them as the same[19]. This sort of ambiguity would clearly give inaccurate 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 feature context. For example, a 5-gram would have the capability to distinguish between the above examples. However, the higher n is the less computationally efficient the processing will be and Tan et al found that the pay-off in computational efficiency was not worth it for n=3 but was for n=2[TAN]. There are methods to improve computational efficiency, such as filtering out all features that do not meet some threshold. Tan et al's found that filtering out all features that appeared in less than three documents or less than 0.5% of documents helped sufficiently. Implementing bigrams and combining them with the documents' unigrams, resulted in an average F1-score improvement of 0.03 on the Reuters datasets[20].

Alongside changing the type of features, the way that features are counted can be changed. The simplest form of this is counting the features as binary, representing a feature as 1 if it occurs in a document and 0 if it does not. Another simple count is to represent a feature by the number of times it

occurs in a document. This method gives more influence to longer documents. Term frequency, shown in Equation 2.8, solves this by dividing the term count by the total number of words in the document. Inverse document frequency, shown in Equation 2.9, gives a higher weight to words that appear in low amounts of words. Term frequency-inverse document frequency (tf-idf) is a combination of its namesake methods, shown in Equation 2.10. Lan et al investigate these methods and variants of them and find that term frequency performs better than tf-idf frequency for the Reuters dataset despite tf-idf being the standard.

$$tf = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2.8}$$

$$idf = \log \frac{N}{n_j} \tag{2.9}$$

$$tf\text{-}idf = tf * idf \tag{2.10}$$

Liu and Yang proposed a simple improvement to tf-idf named term frequency-inverse document frequency-class frequency (tf-idf-cf), shown in Equation 2.11 this adds weight to a term if it commonly occurs in the document's class. This gave a 0.08 accuracy improvement over tf-idf on the Reuters dataset.

$$tf\text{-}idf\text{-}cf = tfidf * \frac{n_{cij}}{N_{ci}} \tag{2.11}$$

2.2.5 Detecting Trends

Regression analysis is the construction of mathematical models which explain relationships between variables. Ordinary least squares is the simplest regression technique intuitively. A line $f(x, \beta)$ is constructed from the parameters **beta** that minimise the total squared distance, $S(\beta)$, of the line to each predicted input y_i[21].

$$f(x, \beta) = \sum \beta_j \phi_j(x) \tag{2.12}$$

$$r_i(\beta) = y_i - f(x_i, \beta_\mathbf{i}) \tag{2.13}$$

$$S(\beta) = \sum r_i^2(\beta) \tag{2.14}$$

The least squares method can produce lines that are not straight using polynomial linear regression where β remains linear but the x inputs are polynomials. Alternatively, a nonlinear method can be used which involves the β parameters taking the form of functions[22].

Ordinary least squares works on the assumption that all distances are errors that are independent and identically distributed normally. However, a single grossly out-lying observation can spoil the estimate[23]. The solution to this is to modify least squares in order to make the estimate less sensitive to extreme values of residuals. This is called robust regression.

2.2.6 Statistical Significance

If a linear regression line has a gradient not equal to zero, there is a relationship between the variables. It is very likely that this will be the case but it is not necessarily the case that the gradient will be similar every time. Statistical significance is the idea that a claimed trend has not happened by chance and will consistently occur with other data. Fisher arbitrarily dictated that a trend is statistically significant if its p-value is less than 0.05[24].

A p-value is the probability that a statistical summary of the data would be equal to or more extreme than its observed value[25]. It is worked out based on the gradient's t-value. A t-value is the ratio of the difference between the estimated value of a parameter, from its hypothesised value and the standard error. For finding the statistical significance of a gradient, the estimated value of the parameter, is the gradient worked out in linear regression and the hypothesised value is zero.

However, the importance of p-values has come into question since many scientific claims have turned out to be false despite being backed by p-values. Benjamin et al recommend reducing the statistical significance threshold to 0.005 to ensure it is very likely to be statistically significant before making any bold claims[26]. The American Statistical Association, meanwhile, just warn that statistical significance should not be taken as gospel and any decisions came to based on p-values should be assisted by thorough

human analysis[25]. As well as this, Boos and Stefanski question the reproducibility of p-values and encourage cross-validation-like methods to ensure that a claim is statistically significant[27].

2.2.7 Pre-processing

Pre-processing text before classification is where text is formatted and changed before classification. It primarily reduces the feature set which improves the efficiency of classification but it can also improve classification accuracy. Three of the main pre-processing methods are stemming, removing stopwords, lowercasing.

Stemming is where the suffix of a word is removed in order to collate all variants of the same word. This would change the words 'property' and 'properties' to 'properti'. Lemmatising is a similar method performing same the task which removes the suffix of a word but then replaces it with a standard suffix so it is still a valid word[28]. This would change both 'property' and 'properties' to 'property'. Toman et al found that both stemming and lemmatising either made no significant improvement on classification or gave significantly worse classification results than using neither method[28]. Meanwhile, Uysal and Gunal found that whether there was performance improvement with stemming was dependent on the domain but the proportion of words that were stemmed were up to 0.45[29], therefore, increasing the efficiency of the classification.

Stopwords are words that are so common in all classes that they make no difference to classification and therefore, there may be no point in including them. This includes words such as 'the', 'and' and 'i'. Interestingly, Toman et al found that stopwords should be taken out of text as that improves classification but Uysal and Gunal found that stopwords should be left in as they can inadvertently indicate class. This reaffirms Uysal and Gunal's hypothesis that there is no set of pre-processing methods that can be applied to all domains and they should instead be experimented with.

Lowercasing is replacing all capital letters with lowercase letters so a normal word is treated as the same word even when it is at the start of the sentence. Uysal and Gunal found this made a significant improvement to classification in all domains.

2.3 Visualisation

2.3.1 Displaying Trends

Gestalt set out principles of visual perception. These are widely used in information visualisation because they aid understanding of how people perceive patterns[30]. The following list is a subset of Gestalt principles that are most useful for emphasising trends and anomalies:

- Proximity, illustrated in Figure 2.2, is the idea that groups of objects that are close together are quickly identified as a group;
- Similarity, illustrated in Figure 2.3, is the idea that objects that have similar attributes, such as colour or shape, are quickly identified as a group;
- Enclosure, illustrated in Figure 2.4, is the idea that objects that are encapsulated by another object are quickly identified as a group;
- Figure & ground, illustrated in Figure 2.5, is the idea that a perceived foreground and background are quickly identified as two separate groups.

Healey extends upon this by investigating preattentive features of objects[31]. Preattentive features are features which are identified immediately as an image is first scanned. These included length, width, volume, orientation, number, density, curvature, closure, terminators, colour, luminance and more. The idea is that a human can identify any inconsistencies in these features without having to think, such as if there is one red dot among many blue dots. It is important to note that some features are assymetric, for example, a curved line among vertical lines can be identified preattentively but a vertical line among curved lines cannot be. Healey also investigated how visual features compete for attention. The purpose of this is so the most important information can be encoded to be processed earliest. He identified the order of processing of visual features as follows:

1. Determine the 3D layout of a scene;
2. Determine surface structures and volumes;

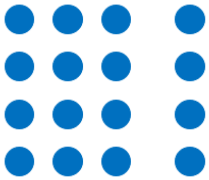

Figure 2.2: An illustration of proximity.

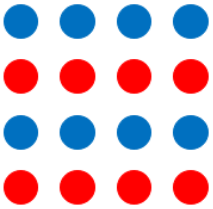

Figure 2.3: An illustration of similarity.

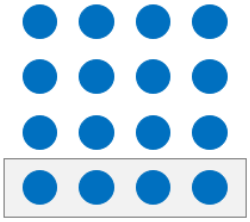

Figure 2.4: An illustration of enclosure.

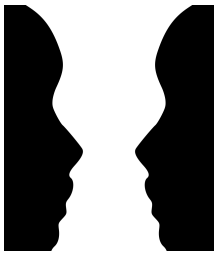

Figure 2.5: An illustration of figure & ground.

3. Establish object movement;
4. Interpret luminance gradients across surfaces;
5. Use colour to fine tune these interpretations.

Kosslyn lays out eight principles of graph design which emphasise the importance of the graph's context[32]. The following is a list of the principles that do not overlap with Gestalt principles:

- Relevance - communication is most effective when the appropriate amount of information is displayed. All decisions should be made based on whether the outcome more effectively conveys the desired message;
- Appropriate knowledge - communication requires prior knowledge of relevant concepts, jargon and symbols. The amount of explanation behind the graph should be tailored to the prior knowledge of the audience.
- Salience - attention is drawn to large perceptible differences. It is important to highlight the significant information in the graph with striking changes in the visualisation;
- Compatibility - a message is easiest to understand if its form is compatible with its meaning. This can be fundamental such as using line graphs for a continuous x-axis or to illustrate a message;
- Informative changes - a change in properties should carry a change in information. Any changes in the properties of the visualisation must have a reason behind it, otherwise it may confuse the viewer;
- Capacity limitations - people have a limited capacity to retain and process informations so too much information can cause important information to be lost. This is why it's important either to limit information in a visualisation or to prioritise the most important information according to Healey's processing order, for example.

2.4 Usability

2.4.1 User Interface Heuristics

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 problems that occur in real systems[33]. 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 consistent 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.

2.4.2 Code Usability

Spinellis outlines the best practices for readable code[34]. 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 code should be somewhat self-explanatory and therefore, bad code cannot be rectified with lengthy documentation and the code should be rewritten instead. There are many textbooks on the principles of writing readable and refactorable code. Most suggest using object oriented techniques for this[35].

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[36].

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, CC-BY-SA 3.0, 2008.

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

[6] Megan R. Blakely, *BILETA 2019 Abstract*, 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] Doug Eboch, *Some Thoughts About Tone*, letsschmooze.blogspot.com, 2015.

[9] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence*, vol. 14, 1995, pp. 1137–1145.

[10] Digna R. Velez, Bll C. WHite, Alison A. Motsinger et al, "A Balanced Accuracy Function for Epistasis Modelling in Imabalanced Datasets Using Multifactor Dimensionality Reduction," *Genetic Epidemiology*, vol. 31, no. 4, pp. 306–315, 2007.

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

[12] David M. W. Powers, "Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

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

[14] 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.

[15] 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.

[16] Rodrigo Moares, Joao F. Valiati, Wilsion P.G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[17] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Empirical Methods in Natural Language Processing*, vol. 10, 2002, pp. 79–86.

[18] Santanu Mandal, Sumit Gupta, "A lexicon-based text classification model to analyse and predict sentiments from online reviews," in *International Conference on Computer, Electrical & Communication Engineering*, 2016.

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

[20] 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.

[21] George A.F. Seber, Alan J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2012, ISBN: 978-0471415404.

[22] Harvey J. Motulsky, Lennart A. Ransnas, "Fitting Curves to Data Using Nonlinear Regression: A Practical and Nonmathematical Review," *The Federation of American Societies for Experimental Biology Journal*, vol. 1, no. 5, 1987.

[23] Peter J. Huber, "Robust Regression: Asymptotics, Conjenctures and Monte Carlo," *The Annals of Statistics*, vol. 1, no. 5, pp. 799–821, 1973.

[24] Ronald A. Fisher, *Statistical Methods for Research*, 14th ed. Oliver & Boyd, 1970, ISBN: 978-0050021705.

[25] Ronald L. Wasserstein, Nicole A. Lazar, "The ASA's Statement on p-values, Conext, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016.

[26] Daniel J. Benjamin et al, "Redefine Statistical Significance," *Nature Human Behaviour*, vol. 2, pp. 6–10, 2018.

[27] Daniel D. Boos, Leonard A. Stefanski, "P-value precision and preproducibility," *The American Statistician*, vol. 65, no. 4, pp. 213–221, 2011.

[28] Michal Toman, Roman Tesar, Karel Jezek, "Influence of Word Normalization on Text Classification," in *The Association for Information Science and Technology*, 2006.

[29] Alper K. Uysal, Serkan Gunal, "The Impact of Preprocessing on Text Classification," *Information Processing and Management*, vol. 50, pp. 104–112, 2012.

[30] Stephen G. Kobourov, Tamara Mchedlidze, Laura Vonessen, "Gestalt Principles in Graph Drawing," in *Graph Drawing and Network Visualization*, vol. 9411, 2015, pp. 558–560.

[31] Christopher G. Healey, James T. Enns, "Attention and Visual Memory in Visualization and Computer Graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 7, 2012.

[32] Stephen M. Kosslyn, *Graph Design for the Eye and Mind*. OUP USA, 2006, ISBN: 978-0195311846.

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

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

[35] Robert C. Martin, *Clean Code: A Handbook of Agile Software Craftsmanship*, 1st ed. Prentice Hall, 2008, ISBN: 978-0132350884.

[36] Guido van Rossum et al, *Index of Python Enhancement Proposals*, python.org/dev/peps, 2019.

Appendix A

Ground Truths

name	class
The International Journal of Cultural Property	Human Rights
The International Journal of Heritage Studies	Human Rights
The Journal of Intellectual Property Law	Intellectual Property
The Journal of World Intellectual Property	Intellectual Property

Table A.1: A list of journals for each class supplied by Dr. Megan Rae Blakely.

name	year of publication	class
Basic Texts of the 1972 World Heritage Convention	1972	Human Rights
Convention for the Safeguaring of the Intangible Cultural Heritage	2003	Human Rights
The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS)	1995	Intellectual Property
Berne Convention for the Protection of Literary and Artistic Works	1979	Intellectual Property

Table A.2: A list of journals for each class supplied by Dr. Megan Rae Blakely.

human rights	intellectual property	user	creator
heitage	protection	access	right
cultural	property	open	protection
intangible	author	right	control
safeguarding	licence	progress	exclusive
natural	intellectual	compulsory	
protection	trademark	licence	licence
international	authority	free	royalty
property	patent	collaborative	creative
educational	monopoly	reuse	control
conservation	scientific	acknowledgement	exclusionary
universal	artistic	credit	limited
identity	literary	noncommercial	author
generation	intangible	fan fiction	individual
inherent	idea	adaptation	literary
generation	expression	parody	artistic
transmission	discovery	research	invention
	copyright	exception	economic
	useful	limitation	moral
	derivative	education	compensation
	transformative	personal	incentive
	limited		
	infringe		
	fixed		
	incentive		

Table A.3: A list of keyword for each class supplied by Dr. Megan Rae Blakely.