Background

Court

Features used

model

Related work

SOME TITLE THAT SUMMARIZES THIS PAPER

The article ‘Empirical Studies of the Webs of International Case Law: A new Research Agenda’ maps out a variety of methodologies employed by political scientists and legal scholars to examine the ***developme***nt of case law, especially in international courts, both qualitative and quantitative. The authors propose an approach that combines citation network analysis, computational linguistics and legal analysis as a new direction for international case law scholarship.

The article sets two branches of case-law scholarship in opposition. One branch represents the political scientist’s approach, which operates empirically and quantitatively to focus on the behavior of judges or the outcome of cases, often ignoring the content of the judgment itself. What’s in focus is perhaps the judge’s professional or ideological background, or the court’s motivation for the number of citations used. The other branch is represented by the legal scholar. The latter can be situated in the ‘text book tradition’, updating a record of the important cases pertaining to an area of the law. A legal scholar may also study an important decision in order to clarify the precedent that has been set. The authors point critically to the manner in which many legal scholars are too eager to find congruence between previous practice and legal categories, and what might arise of new case law.

A brief explanation of citation network analysis, its pertinence to the study of the law, and various importance measures is followed by an overview of how citation network analysis has been applied to domestic and international court jurisprudence. A legal theoretical framework is set out in which the interpretation of the law by practitioners and scholars requires a mapping of the case-law that must be comprehensive in order to be correct. The tools outlined in this article provide support for a theoretical framework highlighting the importance of case-law over policy in the application and interpretation of the law.

Despite acknowledging the weaknesses of network analysis (a static picture of the law is posited, and the reason for citation is ignored) the authors advance an approach that builds on previous attempts to combine quantitative methods such as network analysis with qualitative examinations of legal language and argument. More specifically, they perform modularity maximization on a citation network of the Court of Justice of the European Union. This technique, which groups cases into communities according to shared connections, provides the authors with a finding that a central node to one community, ranking high in authority, has gone largely ignored in textbooks. Having made note of the oversight detected through network analysis, an analysis of the case’s language is anchored on the presence or absence of a passage involving the word ‘effectiveness’. They employ corpus linguistics tools to show that different collocations involving ‘effectiveness’ can be juxtaposed with case-law network communities, presenting a systematic study of precedent combining network structure and language.

SOME TITLE THAT SUMMARIZES THIS PAPER

The article ‘Who Should I cite? Learning literature Search models from Citation behavior’ is heavily used as inspiration for my thesis and comes from the field of Natural Language Processing and Information Retrieval. The authors combine information from a citation network, the text content and metadata to produce a retrieval model for the sake of scientific literature search. Their features include:

-A score for the similarity of terms in the query (an article abstract) and the documents (article reference canditates), relying on tf-idf values of the vocabulary.

-The PageRank score of a document, and its citation count, as well as the citation count of its venue, its author, and its author’s H-index.

-The number of years between the article abstract (the query) and the potential reference.

-A similarity score for query terms and document + citing documents terms.

-PMI, EXPLAIN IF I USE IT

-A cosine similarity score for topic distributions vectors of query and document , and a score for query and citing documents

-A series of features based on the query author’s previous citations, and the names of the author(s) of the query and documents.

They compare different classifiers for learning the weights of these features, using an iterative approach where sample articles are added to the training data at every iteration, and being compared the training abstract’s actual reference list. The classifier iteratively learns what is relevant and what is not and updates the feature weights accordingly, before resampling and retraining. And SVM classifier outperformed a L-BFGS classifier evaluated with mean average precision. Using all features also produced the best results, with citation-count, and topic similarity between abstract and reference candidates(+ reference’s citing documents) AND TERMS PMI??? Proved most important.

SOME TITLE THAT SUMMARIZES THIS PAPER

The article ‘The Role of Precedent at the European Court of Human Rights: A Network Analysis of Case Citations’ examines the development of precedent at the ECHR and the various political, strategic and legal motivations that factor into the courts decisions.

Beyond ensuring a consistency in judgments, a court can use precedent in order to legitimize its decision-making and convince parties to follow its judgments and principles. In the case of an international court such as the ECHR, examined by the authors, affected parties are state actors. The court has no mechanism for enforcing its judgments so a domestic court needs to be persuaded by resorting to the legitimacy of existing precedent. A third type of citation practice is said to be influenced by a careful consideration of a nation’s interest and how differing morals play against individual rights. This caution is also strategic; the display of sensitivity is necessary to influence the affected government. These three perspectives (consistency, as well as ‘strategic’ and ‘relativist’ legitimation) present varying expectations as to why judges cite precedent, how important the cited cases are, and how many cases are cited. Hypotheses as to use of precedent are formulated to test the validity of either perspective.

The article goes on to describe the data (an ECHR citation network ending at 2006), the distribution of inward and outward citations and their totals over time, how this compares to the US Supreme Court and other figures, i.e. the number of citations and cases, and the distribution of cases across convention articles.

The authors arrive at a series of findings that support a perspective in which citation is done to communicate with the involved parties. Hypotheses based on a view in which the purpose of citation is to persuade domestic courts and governments are validated by their research. By examining the characteristics of the cases in the network along with number and importance of cases they cite (hub score), they find that 1) violations of convention articles associated to more fundamental issues such as life and torture are backed up by more authoritative cases, 2) these judgments also go on to be more authoritative, i.e. cited heavily 3) the court relies on precedent more when ruling against a government.

Modularity maximization is performed on the citation network to create communities of citations. They find that the communities are formed around common legal issues rather than being grouped by nation, confirming that 4) the court cites previous cases based on legal substance rather than being concerned with legal culture of a domestic court system and other country-specific characteristics.

SOME TITLE THAT SUMMARIZES THIS PAPER

DEFINITELY WRITE ABOUT THIS ONE!!!!

The article ‘Predicting a Scientific Community’s Response to an Article’

DATA:

The entirety of the data used for my experiments consist of a corpus of judgments texts from the European Court of Human Rights and a network representing these judgments and the citations between them. These two datasets were generously provided by the iCourts research centre at the University of Copenhagen. Additional datasets employed in my analyses are a corpus of publications covering the case law related to particular articles of the European Convention of Human Rights, as well as a list corresponding to ‘controversial’ issues from Wikipedia. ADD IF ADDED

The judgment texts total \_\_\_ documents, mostly in HTML format. A look at how these judgments are distributed across time reveals a much more active court in recent years: \_\_\_ percent of the data consists of judgments from after 2010, \_\_\_ percent from after 2000 while the first 20 years of the court’s history produced \_\_\_\_ judgments, representing \_\_\_\_ percent of the corpus.

An important preprocessing step involved the removal of texts corresponding to the court’s decisions rather than judgments. The former purely examines the admissibility of the case and will consequently not go at length into the merits of the case. Every judgment text is made of several parts, each varying in degrees of standardized language. The opening of a judgment lays out the respondent, applicant and which judges are sitting for this particular case. A ‘Procedure’ section establishes when the application originated and the article of the convention that may have been violated. Under the ‘Facts’ heading, the circumstances of the case are explained, often including the trajectory of the case along various national courts. Under the same header, the court will sometimes look at relevant law from the respondent state as well as international law beyond the European Convention on Human Rights. A section entitled ‘The Law’ examines the purported violation of the applicant’s right, (as it is argued by the applicant and the respondent Government) how the article convention applies to case at hand and how the court has dealt with similar cases in the past. If the case is submitted to the court under many article violations, each violation and article is considered in turn. The court also considers the extent and nature of the compensation awarded to the applicant if a violation of a convention has been found. Any judge(s) disagreeing with majority opinion of the court will lay out their argument in a dissenting opinion, and possibly in a separate opinion, if a judge is dissenting for reasons that differ from other dissenting judges. Judges can also express their view in ‘partly concurring and partly dissenting’ opinions.

Although it is possible that judgments may cite previous cases due to these dissenting opinions, I have filtered out these sections from the texts when modeling the court’s citations. The sections of a judgment that approach legal boilerplate were kept in with the assumption that a proper model would ignore such linguistic patterns on account of their pervasiveness.

Having removed from this corpus the dissenting opinions, decisions, documents in other languages or formats other than HTML, the data is converted to plain text and any reference to a case number or case title is also filtered out so as to avoid predictors that simply result from the citing and cited documents having in their shared vocabulary the title or number of the cited case. In some instances, a case number may correspond to multiple documents, if the court revisits the case in order to handle compensation for a violation, which has not been dealt with. For the sake of my experiments, these resubmissions are ignored. When modeling the texts and their citations I am also limiting my data to judgments that appear in both the citation network and the corpus of judgment texts. After these preprocessing steps, the total count of judgments used is \_\_\_\_\_\_.

The citations between judgments are represented by a directed graph, upon which the modeling of citation is performed. This network has been constructed at iCourts by identifying a variety of citation patterns in the texts: a case number or title is referenced, or paragraphs from the cited text might be mentioned, often in the form of “(see paragraph…)”. The graph is composed of \_\_\_\_ nodes and \_\_\_\_ directed edges, where source nodes are the citing case and targets the cited. Node attributes consist of various metadata, some of which is used as features for prediction. The language of the judgment (English or French), the type of document (judgment or grand chamber judgment) and a short string identifying the conclusion reached by the court are all ignored in my feature candidates, as are title and originating body. My experiments do make use of the case’s date, respondent country and the article(s) under which the application is submitted.

The distribution of cases across countries and across articles is far from uniform. A total of \_\_\_\_\_ cases (\_\_\_ percent of the data) fall under article 6, protecting the right to a fair trial. Italy is the respondent country in \_\_\_ percent of the cases, \_\_\_ of which deal with article 6. Russia and Turkey are also overrepresented (\_\_\_ percent and \_\_\_percent). The court has classified cases according to their level of importance (1,2, 3 or 4), corresponding to the case’s contribution to the case-law. The relative frequency of certain respondent countries or purported article violations can be indicative of applications to the court where the circumstances of the case closely resemble those of existing case-law, and consequently provide little opportunity for the development of the case-law. Their estimated importance is thus also used in this project’s experiments.

I have created article–specific subgraphs for articles that appear at least \_\_\_ times, both as a result of the observing this distribution across articles, and in order to return ranked predictors that would hopefully be more specific to the case-law for each article, while taking into consideration the sufficiency of case numbers and graph edges for adequate modeling. The majority of the cases fall under more than one article, and an analysis of the most common article pairings may be of interest, but I have limited my experiments to articles:\_\_\_\_\_\_. The number of edges, representing citations, in these subgraphs range from \_\_\_ to \_\_\_\_\_.

I have collected a corpus of relevant publications that are affiliated to the court without being binding on its judges. One series, Guides on the Convention, is prepared by the Research and Library Division within the Directorate of the Jurisconsult and is made available on the ECHR website. A second series, Human Rights Handbooks, is made available by the Directorate General Human Rights Council of Europe. Both series cover important cases, discuss how the convention article has been interpreted, and provide insights on how the article ought to be implemented. I have collected \_\_\_ of these PDF documents, which I have converted to plain text files. The average publication contains \_\_\_ pages, \_\_\_ words.

Article-specific publications from the series often contain in the title ‘A guide to the implementation of Article \_\_\_’.

Wikipedia provides a list of topics that it deems controversial on account of the number of revisions Wikipedia contributors carry out and how edits often are performed in a circular manner: a piece of information is changed only to eventually be re-edited back in, in a circular manner. This list of controversial issues is divided into \_\_\_ sections (such as Philosophy, Religion and Sexuality) and at the time of writing totals \_\_\_\_\_ issues.

Methodology

If yo do l1 penalization you end up with some components of your estimator parameter vector being 0 so you can do this automatic simultaneous variable selection and parameter estimation. Stability selection is not in itself a variable selection method but it’s one that youo can bolt onto any existing variable seleciotn technique to improve its performance. Can be used in conjucuton with any base variable selction procedure and any underlying data generating mechanism. The idea is simple: you apply your favorite base selection procedures to subsamples of the data then you aggregate. Do with amny splits of the data , subsamples then youe eventaulyl select the variables that keep getting chosen over and over agin in these subssampels.

In paper introduce idea of feature of method where they prove a finite sample upper bound on the expected umber of falsely selected samples you end up with through this procedure. You can use this upper bound to determine the threshold of the proportion of times a variable has to occur in the asubsample for you to choose it eventually

The model im assuming is very straighotofrwatd, have some data. I.i.d . some are signal varibales some are noise vairbales. Not worry too muc by what I mean with signal and noise in this talk. Imagine each random vector can be artitioned as a corvariate and reposne pair with a covariate I p dimensional space and a real valued response. Imagine you ave som log likelihood of the form. Signal variable are component are non zero noise are zero coefficients. I

The limitation of L1-based sparse models is that faced with a group of very correlated features, they will select only one. To mitigate this problem, it is possible to use randomization techniques, reestimating the sparse model many times perturbing the design matrix or sub-sampling data and counting how many times a given regressor is selected.

[Stability selection](http://stat.ethz.ch/~nicolai/stability.pdf) is a relatively novel method for feature selection, based on subsampling in combination with selection algorithms (which could be regression, SVMs or other similar method). The high level idea is to apply a feature selection algorithm on different subsets of data and with different subsets of features. After repeating the process a number of times, the selection results can be aggregated, for example by checking how many times a feature ended up being selected as important when it was in an inspected feature subset. We can expect strong features to have scores close to 100%, since they are always selected when possible. Weaker, but still relevant features will also have non-zero scores, since they would be selected when stronger features are not present in the currently selected subset, while irrelevant features would have scores (close to) zero, since they would never be among selected features.

Randomized Logistic Regression We use randomized logistic regression to assess the importance of features. Our model uses 27 features to model stopout. In order to best fit a training set, a logistic regression model optimizes weights for each feature. To assess the importance of the features randomized logistic regression repeatedly models a perturbed data set (sub-sample) with regularization, and works as follows:

Step 1: Sample without replacement 75% of the training data each time (the variables are normalized ahead of training).

Step 2: Train a logistic regression model on the subsampled data, with randomized regularization coefficient for each variable.

The randomized coefficient βj is sampled from uniform distribution [λ, λ α ], where α ∈ (0, 1] and λ is the regularization coefficient usually used in standard regularized regression approaches. This randomization places different selection pressure for different variables.

Step 3: For every covariate evaluate b j s = µ(wj , th) where µ is a unit step function and wj is the

coefficients for covariate i and th is the threshold we set to deem the feature important.

This is set at 0.25.

Thus this results in a binary vector, that represents the selection of the covariate.

This binary vector is (lag × |features|) long where 1 at a location j implies feature i = j mod 27 was present in this model.

Step 4: Repeat Steps 1, 2 and 3 a total of 200 times.

Step 5: Estimate the importance of the covariate j by calculating the selection probabilities P s b j s 200 .

introduces a new method called stability selection whose goal is to provide an algorithm for performing model selection in a structure learning problem while controlling the number of false discoveries.

It works in the high-dimensional data setting (p n), which is currently a very active area of research in statistics and machine learning. 2. It provides control on the family-wise error rate in the finite sample setting, which is more practical than an asymptotic guarantee.

Let’s assume that we have a generic structure estimation algorithm that takes a dataset Z = Z1, . . . , Zn and a regularization parameter λ and returns a selection set Sˆλ . We can think of this algorithm as a black box. The stability selection algorithm then runs as follows:

1. Define a candidate set of regularization parameters Λ and a subsample number N.

2. For each value of λ ∈ Λ, do:

(a) Start with the full dataset Z(full) = Z1, . . . , Zn

(b) For each i in 1, . . . , N, do:

i. Subsample from Z(full) without replacement to generate a smaller dataset of size bn/2c, given by Z(i) .

ii. Run the selection algorithm on dataset Z(i) with parameter λ to obtain a selection set Sˆλ (i) .

(c) Given the selection sets from each subsample, calculate the empirical selection probability for each model component:

Πˆ λ k = P{k ∈ Sˆλ } = 1 N X N i=1 I{k ∈ Sˆλ i }.

The selection probability for component k is its probability of being selected by the algorithm.

3. Given the selection probabilities for each component and for each value of λ, construct the stable set according to the following definition:

Sˆstable = {k : max λ∈Λ Πˆ λ k ≥ πthr}

where πthr is a predefined threshold.

Note that this procedure doesn’t simply find the best value of λ ∈ Λ and then use it in the algorithm, but actually identifies a set of “stable” variables that are selected with high probability. The authors state that the empirical results vary little for threshold values in the range (0.6, 0.9), and are also not sensitive to choices of Λ.

The Stability Selection theory, recently proposed by [[21](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576737/#R21)] is a general approach to address problems related to variable selection or discrete structure estimation (as graphs or clusters). The properties of this approach are particularly beneficial for applications involving high dimensional data, specially in cases where the number of variables or covariates *p* largely exceeds the number of examples *n* (i.e. the *p* >> *n* case).

In the stability selection framework, data are perturbed several times (for example by iterative sub-sampling the examples). For each perturbation, a method that produces sparse coefficients is applied to a sub-sample of the data. After a large number of iterations, all features that were selected in a large fraction of the perturbations are chosen. Finally a cutoff threshold (0 < *thr* < 1) is applied in order to select the most stable features.

According to the stability selection theory, for every set *K* ⊆ 1, ⋯ , *p*, the probability of *K* being in the selected set Sˆλ(I) is defined as

ΠˆλK=P∗(K⊆Sˆλ(I))

(2)

where *I* is a random subsample of 1, ⋯ , *n* of size n2 drawn without replacement. According to [[21](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576737/#R21)], the probability *P*\* in the definition 2 regards both the random sub-sampling and other sources of randomness.

It is important to emphasize that according to stability selection theory, any regression method which produces sparse results can be used to select the features, as one is interested in the frequency of selections and not in the sparsity inherent to specific methods.

In [[21](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576737/#R21)], the authors used the LASSO to demonstrate the properties of the stability selection framework in an application to select relevant features in a vitamin gene expression data set. The data set consisted of 115 examples and 4088 features. The authors permuted 4082 features and applied stability selection to find the remaining six relevant features.

The original formulation of the stability selection theory is based on sub-sampling of examples (as in bootstrapping procedures). However, the authors also proposed a modified version of the original framework, which they called *Randomised LASSO*. In this approach, instead of penalizing the absolute value *βk* of every component with a penalty term proportional to *λ* (as in [equation 1](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576737/#FD1)), the Randomised LASSO changes the penalty *λ* to a randomly chosen value in a predefined range, according to the following equation:

βλˆ,W=argminβ∈Rp‖Y−Xβ‖22+λ∑k=1p∣βk∣Wk

(3)

The re-weighting is not based on any previous estimate, but is simply chosen randomly. According to [[21](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576737/#R21)], applying this random re-weighting several times and looking for variables that are chosen often will turn out to be a very powerful procedure. They showed the superiority of Randomized LASSO in relation to the original stability selection formulation in the vitamin data set. Using the Randomized LASSO the six non-permuted features were selected and much less permuted features were included in the selected set (i.e. the number of false positive selections was lower than in the original formulation).

Stability selection is a general framework to combine variable selection methods such as penalized regression models with subsampling strategies. Variable selection probabilities are estimated by applying variable selection methods to subsamples of the data, drawn without replacement, and estimating the proportion of subsamples where the variable was included in the fitted model. These selection probabilities are used to define a set of stable variables. [Meinshausen and Bühlmann (2010)](http://bioinformatics.oxfordjournals.org/content/early/2015/04/28/bioinformatics.btv197.full#ref-17) provide a theoretical framework for controlling Type I error rates of falsely assigning variables to the set of stable variables. Here we suggest to apply the subsampling scheme of stability selection to the lassoestimator involved in the RSPCA algorithm to estimate selection probabilities which are then used to identify the truly relevant variables contributing to a PC. As the lasso selects true variables with high probability the corresponding selection probabilities estimated with stability selection are expected to dominate those of irrelevant variables. Applying a classical forward model selection to the features ranked by these selection probabilities, sparse loadings vectors that are parameter estimation consistent as well as model selection consistent can be identified.

In each experiment 200 logistic regression models are formed, thus adding up to a total of approximately 72,000 logistic regression models. For each experiment, randomized logistic regression resulted in a vector of covariates selection probabilities. Each of these probabilities ranged from 0 to 1. 7 Randomized logistic regression analysis gave us fascinating covariate selection probability vectors for all 91 experiments and all cohorts. For each experiment the randomized logistic regression gives us these selection probability vectors for all the covariates which are learner features for different weeks. In order to gain a more quantitative grasp of which features matter for different prediction problem, we aggregate these probabilities.

Stability selection,Logistic regression

Features

The estimator upon which stability selection is performed in order to return ranked predictors is limited to training on features whose interpretation can help explain the court’s case-law. Thus, features pertaining to textual similarity between judgment pairs, citation network metadata and node importance measures are left out. However, in attempting to gauge the performance of a predictor trained on binary ngram feature vectors, judgment citations are modeled by including these additional features.

One such feature involves the textual similarity between citing and cited judgment, based on the cosine similarity of either document’s tf-idf weighted vector representation.

**Tf-idf term weighting**

Tf-idf term weighting is a method for producing vector representations of documents whereby the terms that are important to that document are given a stronger weight at the index corresponding to that term (instead of a 1 for presence and a 0 for absence). ‘Tf’ refers to the frequency of the term in the document, which will be multiplied by the ‘idf’, the inverse of that term’s frequency in the entire corpus. This controls for words whose high frequency in a document should be unsurprising: common words such as determiners or words such as ‘court’ and ‘applicant’, which should be found in every document in the corpus. The weight should be higher if a term appears frequently in a document and if it is specific to it.

EQUATION LOG?

A document is represented as a tf-idf vector, with a length corresponding to the total number of tokens in the corpus, and zeroes where a specific token is absent from the document. The calculated tf-idf value is stored at the index for every term present in the document.

The similarity between the vectors for a cited and citing pair is measured by calculating the cosine of the angle between either vector.

EQUATION

The numerator of this function is the dot product of either vector, and the denominator is the product of the vectors’ Euclidean norms. The cosine similarity measure will be 0 if no words between both documents overlap and 1 if two documents are identical.

**Network analytics**

The authority score of the cited judgment is included as a feature as well. This is returned by running a Hyperlink-Induced Topic Search (HITS) algorithm over the citation network, producing both authority scores and hub scores. A node with a high authority score is cited by more nodes with higher hub scores. A node with a higher hub score cites more cases with high authority scores. The eigenvector centrality score of the cited judgment is also included. The eigenvector centrality of a node is derived by the eigenvector centrality of its neighboring nodes. A case has a higher eigenvector centrality score than one with the same amount of MAKES SURE I GOT The RIGHT SCORES. The indegree centrality score included is simply a count of the node’s incoming edges.

METADATA:

Four more features included in the full feature set make use of metadata associated to the nodes of the citation network. I’ve included a feature for the number of days in between the cited and citing judgment, assuming that predicting a link between two judgments ought to be informed by whether the judgment have happened too close together in time. Whether or not the respondent country in the case being cited is the same as the citing judgment was included as well, under the assumption that case law pertaining to the same country might be highly relevant to future judgments. While my modeling is performed over article-specific subgraphs and their associated documents, I am including as an additional feature the number of articles under which the judgment falls that both cases have in common. A case about the right to privacy as well as protection from discrimination should be more likely to cite cases that are submitted under the same articles. Finally the court provides a rating for almost every case, from one to four, indicating its importance to the case-law. The court’s importance rating for the cited case in a pair of judgments is used as a feature as well.

Evaluation survey?

Experimental set up

For every article of the convention and its associated subgraph of the citation network and judgment texts, stability selection is performed in order to retrieve a ranked list of terms (unigrams, bigrams or trigrams) corresponding to predictors or citation. In order to validate the scores attributed to text features, this ranking is compared to the appearance of predictors in the literature on the case law of the court. To get an idea of why citation happens, the ranking is also compared their appearance in either the ‘Fact’ section or ‘Law’ section of the judgments. ***And in a list of ‘controversial topics’***. The performance of the predictors trained on ngrams only is also compared to predictors trained on textual content, metadata and network structure features.

I am using a module for feature selection provided by SciKit-Learn named RandomizedLogisticRegression, which performs stability selection and returns a list of all feature scores, between 0 and 1. This is performed on features corresponding to a ‘Bag of Common Words’ between two judgment texts. Rather than vectorizing each judgment text from the article-specific subcorpus, every feature vector used for modeling describes word occurrences (1 for presence, 0 for absence) shared by two documents, where the length of the vector represents the length of the corpus’ vocabulary, with stopwords removed. The intersection of the vocabulary of all texts associated to all connected nodes in the article-specific subgraph makes up half of the data, while the second half is an equal number of vocabulary intersections from pairs of judgment texts sampled at random without replacement, whose nodes in the subgraph are not connected. The prediction task is thus a simple binary classification, where a judgment is predicted to cite another based on the intersection of their vocabulary. Performing feature selection on this prediction task to improve the accuracy of the estimator is what returns terms that are stronger predictors or citation. Whether these feature scores correspond to relevance and importance in the case-law is the purpose of the following experiment.

Ranked predictor lists are returned for text features as unigrams, bigrams and trigrams and the following validation step is conducted with every ngram representation.

WHAT SHOULD I DO

Judgment texts are then split by section: the court’s description of the circumstances surrounding the case (‘The Facts’) and it’s interpretation and overview of legal issues arising from the case (‘The Law’). The score attributed to every text feature is compared to the count of that ngram in either section of the judgment. Similarly, CONTROVERSIAL

In order to gauge the performance of the predictor used to return these feature scores, an assortment of features related to the data’s textual content, the citation network’s metadata and its structure is used for training and comparison. The \_\_\_\_\_\_ error metric !! is returned on all possible combinations of these features. The ngram representation of the judgment pairs vocabulary intersection is used as a baseline, onto which all other combinations of features are added.

Prediction accuracy, what kind of error measure. Differnet features

Evaluation survey?