

Automating Document Review

CS224n Final Project

Nathaniel Love

June 9, 2006

Abstract

Law firms engaged in litigation expend significant time and resources on document review, a process requiring brief examination of thousands to hundreds of thousands of client documents. Often performed by first-year associates, the document review task is essentially a classification problem: documents need to be sorted into a number of categories, based on their content, their author(s), and a variety of other features. Due to the high volume of documents to be examined, and the often simple rules needed to classify documents, this process is ripe for assistance by a natural language processing system. This paper explores the desired features of an NLP document review system and demonstrates some results using a prototype.

1 Discovery and Document Review

In large litigation cases or government investigations, law firms expend significant effort in the discovery process, during which litigants must produce internal documents relevant to the matter in question, and make them available either to the opposing counsel or in response to a government subpoena. The documents in question may include memoranda, financial statements, internal work papers, and a great deal of email. The widespread use of email by employees of large corporations has significantly increased the volume of documents that need to be reviewed as part of the discovery process. For the purposes of this exploration, we will focus on the task of classifying emails; many of the other documents in question are included as attachments to emails, and references to these documents in the email body may be used for classifying both types of documents.

1.1 Categorization

When performing document review (often abbreviated as “doc review”), law firm associates are generally looking to sort documents among two pairs of categories. These categories are defined by the parameters of the particular case being litigated— documents may be *responsive* (relevant to one or more of the specific requests in the subpoena) or *non-responsive* (not requested by the subpoena), and either *privileged* or *non-privileged*. Privileged documents are attorney-client communications; in the corporate context, this

can include in-house counsel as well as outside attorneys. As noted above, these categories are fluid—the responsive character of an email may hinge on the sender, the recipient, the topic, the date it was sent, and so on. The *responsive* category is often expressed as the union of a large group of subcategories—the umbrella term *responsive* is used because each of these subcategories is effectively defined as a response to a particular request made by the opposing counsel or government subpoena.

1.2 Costs

This initial document review process alone is extraordinarily time-consuming and costly. A major litigation case or government investigation can produce on the order of 500,000 documents needing review. As some attorneys involved in this process have described their work, a rate of 100 documents/hour is on the quick side of typical. This means that reviewing documents for a major case could easily require 5000 hours of work; when billed at the current standard first-year associate’s rate of around \$275/hour, that means a staggering cost of \$1.375 million to the client.

1.3 Current Process

Associates are assigned sets of documents—the email *in* and *out* boxes of a “custodian” (employee), consisting of thousands of messages—and must pass through and classify all of them. The emails are in electronic format, and may be text-searchable, but the semantic data (sender, recipients, date) may not have been preserved, and may need to be recovered by reading the headers. Associates looking for a particular term can search for it, but using this search term will likely return significant numbers of documents in which this term is used in a non-responsive context. Furthermore, documents for a given custodian are not necessarily presented in any particular order—they are just drawn from the mailboxes one at a time and presented for review. This slows down the process because associates looking for particular topics, senders, or date ranges (based on the subpoena) are presented with essentially randomly drawn emails. Jumping from one topic (an NCAA basketball score update email) to another (a critical internal memo detailing misconduct) in this manner also increases the likelihood of errors.

1.4 Existing Doc Review Systems

One of the prominent doc review systems in use today is LexisNexis’ Applied Discovery. The strong selling point of this electronic doc review system is the move to PDF representations of all documents, from the scanned TIFF files that previous systems used¹. This means that documents are now text-searchable, and Applied Discovery can now extract senders, recipients, and dates from emails, and track email threads². Many of Applied Discovery’s competitors include systems like that still use TIFF files and do not offer text search capabilities.

These doc review systems have overcome the non-trivial task of acquiring all the relevant documents (from various computer systems, hard copies, etc.) and getting them all in the same electronic format—this

¹See, for example, http://www.lexisnexis.com/applieddiscovery/lawlibrary/whitePapers/ADI_PDFTrumpsTIFF.pdf

²Features described at <http://www.lexisnexis.com/applieddiscovery/electronicDiscovery/litigation.asp>

alone has made the document review process much more efficient. However, even the most advanced doc review systems are taking only minimal advantage of the opportunities offered by the transition to electronic text.

2 NLP Doc Review

The document review and classification performed by law firm associates is subject to review by other attorneys working on the case— the classifications they perform essentially flag documents that require closer examination. The discovery process is highly sensitive, so several layers of expert oversight are required. For this reason, it is inappropriate to attempt to build and field an NLP doc review system intended to supplant the entire process, or even the initial associate review. Rather, the intention is to apply natural language processing techniques to augment and increase the efficiency of the process. While associates still need to examine documents in person, categorization by an NLP doc review system can greatly reduce both the amount of time spent on this task as well as the error rate. An NLP doc review system might ideally perform the following:

1. Allow users to initialize system for a particular case with externally provided feature definitions (for responsive and privileged documents).
2. Initial classification of documents based on raw features.
3. Online learning and reclassification as associates categorize documents themselves.
4. Organize pre-classified documents for more efficient presentation to reviewers.

Because the doc review process may involve dozens of associates, each classifying tens of thousands of documents over months of time, the classifiers have high potential to take advantage of the length of this process. The classifiers, using online learning, will quickly acquire sufficient training data to present highly accurate classification suggestions to users, and at the end of the doc review process, the fully-trained classifiers can re-classify the documents that users have already processed, helping catch any errors. Attorneys who perform document review have indicated that presenting pre-classified documents organized by category would greatly increase the speed and accuracy of the process. In this system, the documents arrive in meaningful groups, alerting the users as to which classification is likely appropriate and why. Following the figures given in section 1.2, if using this NLP doc review system could improve the rate by just 10%, this would save \$138,000 in costs; in practice, especially toward the end of the process, the well-trained classifiers may be able to improve the review rate by a much more significant margin.

2.1 Project Scope

This project explores building an NLP doc review system that performs the 4-step task outlined above, with a particular focus on the third step. This interaction with the attorney reviewing documents provides the best opportunity for the system to improve its own performance, and the implementation of this interaction is key to improving the efficiency of the doc review process. The following sections detail the architecture

of a prototype NLP doc review system, using the Enron Corpus as a source of documents; accuracy results are presented as well, using a hand-categorized subset of the Enron Corpus.

2.2 Enron Corpus

The vast majority of documents reviewed by attorneys in many current litigation cases are employee emails. The Enron Corpus, produced in response to subpoenas from the Federal Energy Regulatory Commission (FERC) investigation into Enron's activities, is a set of approximately 500,000 documents (from 150 employee email mailboxes) exactly of the type typically managed by attorneys in the document review process. Critically, the Enron Corpus includes both documents that are essentially universally non-responsive (emails of jokes, personal emails, etc.), as well as documents that are highly responsive to the FERC subpoena. Additionally, the Enron Corpus also contains documents that can be considered privileged— emails between Enron officials and their attorneys, and emails to and from Enron's internal general counsel. While these documents ultimately were produced in the course of the investigation, it would have been a reviewing attorney's job to flag such emails as privileged so as to protect their release; other considerations (outside the scope of this investigation) may have compelled their release. In any event, the Enron Corpus contains documents on both sides of each category.

2.3 Hand-tagged Enron Documents

Classified documents are required in order to perform training and testing of the NLP doc review system using the Enron Corpus. A team at the School of Information Management and Systems at UC-Berkeley has hand-tagged a portion of the Enron Corpus (consisting of approximately 1700 documents) with a set of several dozen categories, including 20 categories of business and non-business topics, 13 categories for included information (attachments, urls, etc.), and 19 categories of emotional tone; each email is tagged with one or more categories³. Clearly, these are not the precise categories required for attorneys engaged in doc review, which requires each email to be classified along the axes of responsiveness and privilege. Fortunately, given the subpoena in question in the Enron investigation, many of the hand-tagged categories used can be directly mapped into the required doc review categories. Some notes on this mapping appear in Appendix A.

3 The Doc Review System

3.1 Feature Selection

By default, the system builds features using the following document attributes: date (month and year), length of email (in 1000s of characters), sender, recipient(s), sender domains, recipient domains, presence of attachments, and unigrams and bigrams (words, not letters) in the subject line. In general, based on the subpoenas in question, a user would have some additional terms to add which would be relevant to the task at hand. For example, the system might support adding feature types like string matching in

³Available at http://bailando.sims.berkeley.edu/enron_email.html

email bodies, before/after date, and defining sender groups. In the Enron case, examples might include the string "California regulations," "market manipulation," or the date January 1, 2000: these are features likely associated with documents responsive to the subpoena[2]⁴. Enron (like all large companies) had outside attorneys as well as in-house counsel (attorneys who worked at Enron)– emails from these law firms' domains as well as these particular employees are very likely privileged. Appendix B gives more detail on the hand-built features.

In the current system, these features are hard-coded for the purposes of testing, and a user interface for entering the features has not been implemented. These hand-added features are the only ones that examine strings in the bodies of emails. Building features based on all words appearing in email bodies would result in an explosion in the number of features, especially since the Enron Corpus contains some very large emails (many thousands of words, consisting of several news articles pasted into the body). With the small amount of training data available, it seems unlikely that any benefit could be gained from this approach. With more data available (as there would be in an implemented system, running over thousands of documents as they are classified) it may be beneficial to perform some limited extraction of strings from email bodies, based on other NLP techniques. Information extraction techniques, particularly named-entity recognition, could assist the system in extracting a smaller set of important terms from documents.

3.2 Initialization

The system builds two maximum-entropy classifiers: one to handle responsiveness, and the other to handle privilege. The classifiers are based on my `MaximumEntropyClassifier` code from assignment 2, with modifications made to support online learning and feature extraction appropriate to this domain. When the system starts, both classifiers are untrained. An unimplemented feature (to be added in subsequent work) would actually initialize the weights of both classifiers given the initial features added by users– in this way, the system could at least make raw, highly uneducated guesses until it had obtained enough data from users to make better classifications.

3.3 Online Learning

After any initial feature definition, the user is presented with the first email to be processed, displayed in a window with a panel with buttons for selecting the proper categorizations for this email. After determining both the responsive/non-responsive and privileged/non-privileged nature of the email, the user continues to the next email. Each classification creates two new datums (getting all features, then adding them to the lists of responsive and privileged training data). The system can be tuned to wait for different numbers of new documents to be classified before retraining; the default is 500 documents. When this threshold is reached, the system pauses to retrain each classifier, using all data obtained thus far. In order to decrease the number of iterations needed for convergence, the classifiers are initialized with the weights determined by previous

⁴One of the most relevant passages from the FERC subpoena: "On February 13, 2002, FERC ordered a staff investigation into Enron and other sellers. Staff was told to gather information on 'whether any entity, including Enron Corporation (through its affiliates or subsidiaries), manipulated short-term prices in the electric energy or natural gas markets in the West, for the period January 1, 2000, forward.'"

training. Because the new datums will introduce new features, feature weights from the old classifiers need to be mapped into the new linear index of features. Figure 1 shows a screenshot of the live system; the selection of the “Not Privileged” and “Responsive” indicates the classifiers’ proposed classifications for the current email; the user can change these selections, if appropriate, before clicking “Done” to continue to the next email.

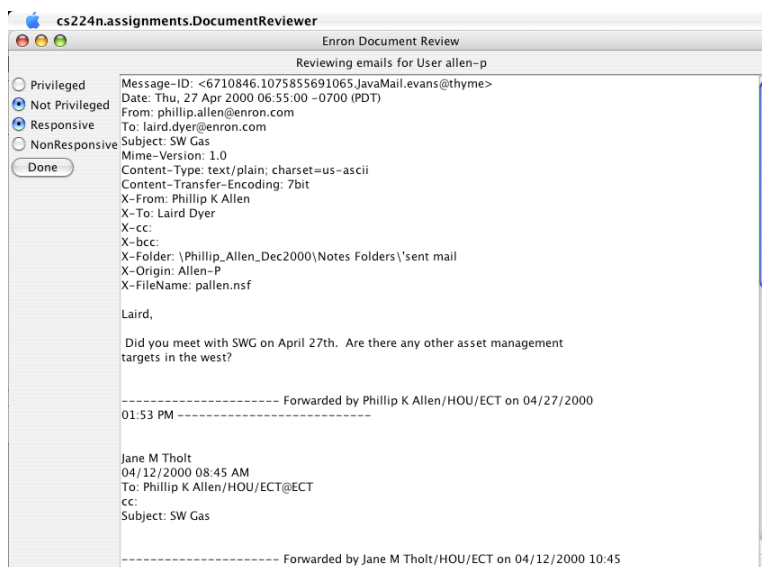


Figure 1: Screenshot of Doc Review User Interface

3.4 Reorganization

Because each individual (owner of an email mailbox) has a particular relationship to the legal case, it seems reasonable to preserve the standard doc review procedure of processing email by user. However, in order to make the review process more efficient, the system orders the documents after classifying them, using the current trained classifiers. Documents are presented to the user with responsive and privileged documents first, then responsive and non-privileged, then non-responsive and privileged, and finally non-responsive and non-privileged. Because the feature sets are dependent on the content of the emails themselves, each unseen document is highly likely to introduce new features. However, the use of encoded datums within the MaxEnt classifier code guarantees that any new features will be ignored in computing the classifications on these new datums.

Other reorganization principles could be added to further sort these grouped emails, following some capabilities already being explored by existing doc review systems mentioned above. Some of these, like threading detection, are the subject of other NLP research that could be applied to document review: it is likely that if one email in a thread is responsive to the subpoena, other emails in the thread will get the same categorization.

One further enhancement to the user interface would be to visually highlight any hand-added features that appear in the current email and that contributed significantly to its classification. This would help users focus immediately on the reason(s) why the system is proposing to classify the document as responsive or non-responsive, and allow them to either quickly confirm that classification or correct an error the system is making. With the high number of features generated, it is likely undesirable to highlight all features, but the hand-built ones are especially relevant to the case and will likely be the most meaningful to the user.

4 Results

The hand-tagged Enron Corpus is rather small, consisting of only 1700 documents. However, since the deployed NLP system would be expected to start classifying documents immediately, training for the first time when only 500 documents had been classified, this data set does provide enough for a demonstration for the system’s in-the-field capability. The 1700 tagged documents were randomly separated into 1500 training emails and 200 test emails. 664 of the training emails were responsive, as were 94 of the test emails. The number of privileged documents was significantly lower, with only 78 privileged training emails and 13 privileged test emails. Training the system on this data yields an accuracy of 72.5% for the responsive classifier, and an accuracy of 93.5% for the privileged classifier. The maximum iterations for both classifiers were set at 40.

The system was then tested as designed to be used in practice, training on the first 500 documents, then on the first 1000, then all 1500. The following table shows the results, including the types of errors being made by the system:

Documents	Responsive Accuracy	False +	False -	Privileged Accuracy	False +	False -
500	69.0%	26	36	94.0%	2	10
1000	74.5%	20	31	92.5%	5	10
1500	75.5%	19	30	93.0%	4	10

4.1 Features

By analyzing the feature weights in the trained responsive classifier, we can see that in spite of a small amount of training data, useful features are receiving appropriate weights. For example, the feature corresponding to emails with less than 1000 characters (a typical single-recipient email header has approximately 500 characters, so this means less than 500 characters in the body) received negative weight. Many of the shortest emails correspond to quick confirmations of meetings or other day-to-day issues unconcerned with matters relevant to the case. In contrast, the longer emails in the 4-5,000 character range received positive weight.

The classifier learned to recognize the major players in the Enron case, assigning positive weights to the email addresses and names of those who were discussing the issues responsive to the subpoena—the “Ken Lay” subject bigram, and `To:andrew.fastow@enron.com` (Enron’s CFO) both corresponded to likely responsive emails. Additionally, the classifier also learned the roles of much more minor play-

ers. For example, I noted that `To:david.parquet@enron.com` had received high positive weight, while `To:nicholas.o'day@enron.com` had received high negative weight. As it turns out, David Parquet was Enron's Vice President for project development in the western U.S.—very likely to be involved in Enron's activities in the California energy markets⁵. In contrast, Nicholas O'Day was Vice President at Enron Japan, certainly an important position at Enron, but unlikely to be involved in the issues investigated by the FERC⁶.

The hand-added features, searched for in the bodies of the emails, also proved to be useful, with both “California” and “markets” receiving positive weight. The subject-line features corresponding to these words also had positive weights, as one would expect, but for these important terms, it is critical to also look in the body—more than half of the training emails mentioning California do so only in the body, and not in the subject line.

One surprising feature was the subject unigram `http://www.stanford.edu/~wolak/`, which received a strong positive weight; this is the web site for Stanford Economics professor Frank A. Wolak. As it turns out, Professor Wolak is an expert on California energy markets and the California energy crisis; he was in communication with Enron employees during the relevant time span.

4.2 Analysis

One major deficiency in the evaluation of the system was the poor training data for the privileged classifier. Most employees would never have cause to contact an attorney, so getting positive examples for this category is very unlikely for many of the emails being reviewed. Enron's general counsel's email folder (`derrick-j`) would have a much higher percentage of privileged emails, but there are only 11 emails in the hand-tagged training data in which he appears, and then only as one of many recipients of an email—making all of these documents unlikely to be privileged attorney-client communications. As a result, the privilege classifier performs well mostly because it labels almost all emails as non-privileged, and 93% of the test emails are non-privileged. Nevertheless, the proper features (both hand-built and automatic) are in place in the system to improve performance of this classifier based on further training as associates review documents, and a much greater number of privileged documents are seen.

A second issue with the system concerns the initial hand-added features. While users provide critical case-specific features at the initialization stage—for example, the name of the illegal trading strategy “Death Star”—associates may review thousands of documents before ever coming across an email containing this term. In the interim, while the system is training, this feature will have received no weight in the classifier. Therefore, when emails with this term do appear, the system is likely to fail to predict the proper classification for the document. However, by defining and extracting these features, they will be present in the classifier, and when users begin classifying emails containing these features, their weights will move in the appropriate direction.

While the hand-categorized portion of the Enron Corpus was useful for testing the NLP doc review system, this investigation clearly could have benefited from a larger amount of training and test data. Unfortunately, such data is extremely expensive to come by—as noted above, expert classification of documents costs

⁵Source: <http://www.sacbee.com/static/archive/news/special/power/080501cleaner.html>

⁶Source: <http://www.gasandoil.com/goc/company/cns12585.htm>

\$275/hour. Law firms certainly have large sets of reviewed documents available, but such data is completely unavailable, privileged and confidential client material. Given time, I would be interested in expanding this prototype into a more full-featured system (implementing the feature-input portion; adding a web interface), and pitching the system to a law firm for testing. My discussions with lawyers who spend significant time in document review indicate that a system with these capabilities would be well-received.

5 Related Work

A group at the University of Massachusetts has investigated using a variety of statistical NLP techniques (including MaxEnt, Naive Bayes, and SVM classifiers) to perform email categorization into folders, training and testing on the Enron Corpus[1]. Here, the goal is to learn and reproduce the process by which humans sort email into folders. This task presents many special problems that are not found in the doc review domain, arising from the different organizational techniques used by different individuals to produce folders and the creation and abandonment of folders over time. The investigators here used only bag-of-words features for the emails, and online learning is not discussed; in our investigation, the features arise from the particular characteristics of the document review task and from specifics of the case at hand, and online learning and reclassification are of critical importance.

Much email classification work focuses on Spam filtering, where spammers essentially tamper with the email content to defeat feature-based classification techniques. This adversarial scenario presents a primary difference from doc review because emails are not intentionally tampered with to obscure their topic or relevance. Spam filters, therefore, tend to be more concerned with non-natural language features of emails, including URLs, images, and spoofed headers. However, the most significant amount of work on email classification has been in this area, and further work on document review could benefit from some of the feature extraction techniques employed in spam filtering.

A Hand-Tagged Categories

The most useful UC-Berkeley categories for determining whether the emails were responsive to the Enron subpoena were the following subject-based categories: *California energy crisis/California politics*, *internal company operations*, *alliances/partnerships*, and *regulations and regulators (includes price caps)*. Emails tagged with these subjects were considered responsive. Emails without any of these category tags were considered non-responsive. For determining privileged emails, the categories *legal documents (complaints, lawsuits, advice)* and *legal advice* were used. The “emotional tone” tag *secrecy/confidentiality* was actually a poor fit, since it did not always indicate attorney-client confidentiality: a large number of personal emails were labeled with this tag, and others were concerned with keeping misconduct under wraps, but were not communications falling under the requirements for privilege. Certain other tags, including *newsletters*, *purely personal*, and *jokes, humor (unrelated to business)* were useful in classifying emails as both non-responsive and non-privileged.

B Initial Feature Selection

The FERC documents[2] also indicated several other areas of investigative interest, including “whether any FERC rules were violated by Enrons dealings with El Paso Electric Company, Portland General Electric Company and Avista Corporation.” Additionally, the FERC was concerned with particular Enron trading strategies, with the nicknames *Ricochet*, *Get Shorty*, *Death Star*, and *Fat-Boy*. These names were added as features to search for in email bodies. For privilege classification, special features were added for the outside attorneys Christian Yoder, Stephen Hall, Gary Fergus and Peter Meringolo, and the law firm of Brobeck, Phleger & Harrison (these names were obtained from FERC documents), as well as James Derrick, Richard Sanders, and Robert Walls, all from Enron’s office of general counsel. These names were relevant as features not only as senders/recipients of emails, but when mentioned in the body of the message (for example, forwarded information from the corporation’s attorney).

References

- [1] Ron Bekkerman, Andrew McCallum, and Gary Huang. Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora. UMass CIIR Technical Report IR-418. <http://www.cs.umass.edu/~ronb/papers/email.pdf>.
- [2] Federal Energy Regulatory Commission. The western energy crisis, the Enron bankruptcy, and FERC’s response. <http://www.ferc.gov/industries/electric/indus-act/wec/chron/chronology.pdf>.