

# Emerging AI & Law Approaches to Automating Analysis and Retrieval of Electronically Stored Information in Discovery Proceedings

ICAIL 2009 Global E-Discovery/E-Disclosure Workshop  
(DESI III)

Kevin D. Ashley  
Professor of Law and Intelligent Systems  
University of Pittsburgh  
School of Law  
Pittsburgh, PA 15260  
([Ashley@pitt.edu](mailto:Ashley@pitt.edu))

Will Bridewell  
Cognitive Systems Laboratory  
Center for the Study of Language and Information  
Stanford University  
Stanford, CA 94305  
([willb@csl.stanford.edu](mailto:willb@csl.stanford.edu))

## *Abstract*

This paper focuses on three emerging techniques for improving the process of automating analysis and retrieval of electronically stored information in discovery proceedings: (1) machine learning to extend and apply users' hypotheses (theories) of document relevance; (2) a hypothesis ontology to generalize user modeling regarding relevance theories; and (3) social network analysis to supplement and apply user modeling regarding relevance theories. Since all three pertain to representing and reasoning with litigators' hypotheses about relevance, and since a central theme of AI & Law involves computationally modeling legal knowledge, reasoning and decision making, all three techniques can be described as characteristic of that field's potentially unique contribution to e-Discovery.

## **Introduction**

The fact that two of the three DESI workshops have been offered in conjunction with International Conferences on Artificial Intelligence (AI) and Law (ICAIL) is, perhaps, a manifestation of an intuition that the field of AI & Law should offer a special approach or contribution to the process of automating analysis and retrieval of electronically stored information in discovery proceedings.

This paper is an attempt to assess that intuition, to describe emerging contributions that reflect an "AI & Law" approach, and to elaborate the reasons why. The purpose is not normative; any approach that applies AI to address this practical legal problem could with justification be called an AI & Law approach. Nor are AI approaches to this legal problem preferable to other approaches more typical of information retrieval (except, perhaps, to an ICAIL audience). Rather, the purpose is descriptive and thematic. The field of AI & Law has a history of research from which certain themes have emerged. It is of some interest to see if the emerging approaches to the e-Discovery problem relate to those themes and how the demands of e-Discovery stretch the themes and methods.

Notably, the problems presented by e-Discovery differ from those usually tackled by AI & Law researchers. While research in AI & Law has addressed information extraction from and automatic classification of legal texts, the work has focused on relatively homogeneous documents, such as legal opinions dealing with a single type of claim and sharing structural features (see, for example, Brüninghaus and Ashley, 2005; Daniels and Rissland, 1997; Gonçalves and Quaresma, 2005; Grover et al., 2003; Hachey and Grover, 2006; Jackson et al., 2003; McCarty, 2007; Thompson, 2001; Uyttendale et al., 1998;

Weber, 1998). The principle difference is the extreme heterogeneity of documents produced in litigation, including not only corporate memoranda and agreements, but also the full panoply of email and other internet-based communications. As a result, techniques developed in AI & Law that rely on either explicit or implicit structure in comparatively homogenous legal documents may be of little avail.

What is a characteristically “AI & Law” approach? The aims and scope of the *Journal of Artificial Intelligence and Law* provide a working definition of the core of the field AI & Law as encompassing “theoretical or empirical studies in artificial intelligence (AI), cognitive psychology, jurisprudence, linguistics, or philosophy which address the development of formal or computational models of legal knowledge, reasoning, and decision making.” In view of this definition, a special approach or contribution of the field of AI & Law would be one that involves computational models of legal knowledge, reasoning and decision making as it pertains to document production and analysis.

It may be premature to identify a characteristically AI & Law approach to automating the analysis and retrieval of electronically stored, heterogeneous documents in discovery proceedings. Since the first DESI Workshop, however, a number of threads have emerged that may coalesce into such an approach. In the absence of predictable document structure, these techniques coalesce around the theme of representing and reasoning with litigators’ hypotheses about relevance.

Traditionally, the area of AI & Law focuses on computational models of legal knowledge, reasoning, and decision making. In the context of e-Discovery these concerns map to the representation of litigators’ hypotheses (or theories) about relevance and their reasoning as they review and make decisions about documents.

## **Definition of Relevance Hypothesis and Examples**

A *relevance hypothesis*, also referred to as a theory of relevance, is a more-or-less abstract description of subject matter that, if found in a document, would make that document relevant (Hogan et al., 2009). In their view, a “theory of relevance” is constructed in an iterative process of modeling the user. Here the users are the senior litigators, the persons engaged in “sensemaking”, a “process of collecting, organizing and creating representations of complex information sets, all centered around some problem they need to understand.” (Bauer et al., 2008 [citation omitted]). The authors state,

The “distinction between senior and junior analyses... as information processing ‘driven by bottom-up processes (from data to theory) or top-down (from theory to data)’ (Pirolli, 2005)... is central to efforts to automate the sensemaking task.... [I]t dictates that replicating senior litigator sensemaking must be rooted not in ‘data mining’ approaches but in systems that reason from a set of “hypotheses.” As noted for the legal domain, these hypothetical constructs characterize the expected intentions of individuals that are involved in the topics of interest in the case,” that is, “the ‘hypotheses’ for a senior litigator [are] the context for and intentionality of the author of documents comprising the ESI [Electronically Stored Information].”

Similarly, Ashley and Bridewell (2007) proposed a hypothesis-based, legal information retrieval approach for e-Discovery that lets litigators express their criteria of document relevance in terms of hypotheses about who communicated what to whom, when, and, to the extent possible, why. With such hypotheses, attorneys can evaluate

their theories about the factual and legal elements of a case, but such hypotheses are entirely inexpressible with advanced Boolean queries.

Examples and their context can bring to light the significance of litigators' hypotheses about document relevance and how to represent them. In terms of legal practice, litigators' hypotheses about document relevance relate to the formal requests for documents that the litigators filed in the case. These requests, in turn, relate to the details of the plaintiff's complaint that formally state the legal claims with which he commenced a lawsuit or the defendant's answer that formally states his defenses. The parties' litigators' prepare and file these statements based upon their best information about the facts that occurred and how the facts connect with relevant legal rules. As new facts, claims, and defenses emerge, the litigators sometimes file modifications. Although Bauer et al. (2008) and Hogan et al. (2009) do not provide examples of litigators' relevance hypotheses or theories, according to Ashley and Bridewell (2007), examples of relevance hypotheses (expressed here as natural language statements) include:

- "There are documents showing that the Vice President of Marketing knew that cigarette advertisements were targeted to children by 1989," or
- "There exist documents to or from employees of a tobacco company or tobacco organization in which a tobacco company officer refers to illegal payments to foreign officials," or
- "There are documents that are communications between Alice and her lawyer Bob between 1985 and 1989," or more generally,
- "There are documents *of a particular kind*, satisfying *particular time constraints*, satisfying *particular social interaction constraints*, that refer to *particular concepts or phrases of interest*."

By analyzing the complaints and document requests, one may infer information about the types of hypotheses that motivated the requests and the corresponding Boolean queries. These examples are based upon the TREC Legal Track (<http://trec.nist.gov/call09.html>) materials including the University of Illinois's Complex Document Information Processing (CDIP) Test Collection. The collection comprises some 6.9 million documents released by tobacco companies in connection with the Master Settlement Agreement mentioned above. As materials for the competition, experts associated with the Sedona Conference (<http://www.thesedonaconference.org/>)—a continuing forum of jurists, attorneys and technologists addressing issues of complex litigation—have created nine fictitious but realistic complaints, and 53 retrieval topics cast as document requests each with a corresponding Boolean query created through a realistic process of negotiation (Baron and Thompson, 2007). For each topic, the National Institute of Standards and Technology generated judgment pools for relevance assessment based on searches of an expert manual searcher and the participating teams and manual assessments by 35 volunteer assessors (Baron and Thompson, 2007). The TREC Legal Track competition has also made public another similarly prepared repository of discovery requests and materials in connection with email documents generated by the Enron litigation (Baron and Thompson, 2007).

## **User Modeling to Generate / Represent Litigators' Relevance Hypotheses**

As noted, Bauer et al. (2008), assert that litigators' relevance hypotheses can be elicited in an iterative user modeling procedure, a task analysis “replicating the cognitive sensemaking task of a senior litigator with an automated, computational platform.” This is the knowledge acquisition procedure for the “Socio-Technical Information Retrieval” or “STIR” approach, which the authors describe as “a knowledge-based system in the classic AI sense.”

Their user-modeling process develops four types of information about relevance:

- (1) the “use case”, namely the user’s objectives for production given the Request for Production, including his objectives to avoid producing too much or too little;
- (2) the scope of the conceptual boundaries of interest (e.g., the legal and other concepts that are relevant to the case);
- (3) the nuance, or the level of specificity of the relevant concepts of interest;
- (4) the linguistic variability—“the variety of ways a concept can be expressed, whether lexically or syntactically.”

Together, “the resultant representation [of these four types of relevance information] is a description of subject matter, that, if found in a document, would make that document relevant (Hogan et al., 2009).”

Exactly how the relevance hypotheses are represented is somewhat unclear. Bauer et al. (2008) emphasizes that “in e-Discovery, senior litigators are NOT reviewing the literal content of text (i.e., bottom-up), but rather the overarching aspects of the situation and the author’s intent (i.e., top-down).” Along with Hogan et al. (2009), these authors caution that this practice and the fact that the documents were created for other purposes mean that “a direct mapping of the documents to the legal topics [is] very difficult.” (As described above, these legal topics are, at least, closely related to the relevance hypotheses.) The iterative cognitive task analysis results in “computational constructs into which the linguistic variations can be expressed and the operations (e.g., Boolean expressions) that enable phrasing of query alternations which capture the meaning being sought in IR computation.” Presumably, the representation of relevance hypotheses captures the linguistic variations and operations and uses them to elaborate upon the Boolean query alternatives.

## **Three Emerging AI & Law Techniques for e-Discovery**

Significantly for AI & Law, the user modeling procedure developed by Bauer, Hogan, and their colleagues is an example of a kind of cognitive task analysis that Buchanan and Headrick (1970) long ago recommended as a prerequisite for progress in applying Artificial Intelligence to law and legal reasoning. Unfortunately, in the annals of the field, this kind of cognitive task analysis has occurred seldom if at all. It would be a welcome return to first principles, and a contribution to AI & Law, simply to provide a detailed description of how to conduct the modeling process and to represent the relevance knowledge obtained, illustrated with some examples of litigators' relevance hypotheses, and demonstrating empirically the utility of the relevance hypotheses.

It would also make important contributions to the field of AI & Law if the process of acquiring problem-solving knowledge from legal experts could be conducted or supplemented, or its results applied, in an automated way. Hogan et al. (2009) raise, and appear to have implemented, a number of such automated techniques, and they hint at others.

The remainder of this paper discusses three such techniques that use (1) machine learning to extend and apply users' theories of relevance; (2) a hypothesis ontology to generalize user modeling regarding relevance theories; and (3) social network analysis to supplement user models of and to apply relevance theories in document analysis and retrieval.

### ***Machine learning to extend and apply theories of relevance***

Machine learning is one means for automating the process of applying and modifying the theory of relevance. Examples of documents that have been classified according to the theory of relevance can be used to train an automated classifier to identify and classify other similar documents.

For instance, in the user-modeling process described by Hogan et al. (2009), a knowledge engineer prepares guidelines, based on the litigator's theory of relevance, for human readers to follow in assessing whether documents are relevant or not. In addition, a substantial number of documents are assessed as relevant or not; these become training data for an automated classifier in a supervised machine learning process. The classification system is regularly run over all of the documents in the corpus and its effectiveness assessed based on random sampling of classified documents. The data thus generated are used to tune the system and modify the theory of relevance (Hogan et al., 2009). Unfortunately, the authors do not elaborate on the particular machine learning techniques applied.

Applying machine learning to extend, modify, and apply relevance hypotheses to legal documents makes an important contribution to AI & Law, especially to heterogeneous legal documents. The approach of learning an automated classifier from manually classified, textual examples was applied in the SMILE+IBP program involving case-based evaluation of hypotheses about the outcomes of cases expressed as texts (Brüninghaus and Ashley, 2005). There, the classifier was trained to identify factors, stereotypical fact patterns that strengthen or weaken a side's claim, in much more homogeneous textual descriptions of case facts, but the goal was similar: to automatically classify the problems and cases in order to evaluate hypotheses (about who should win particular issues) based on the most relevant prior examples. In that work, a number of machine learning algorithms were applied and evaluated, including rule induction programs (e.g., C4.5, Ripper, RL), case-based methods (e.g., nearest neighbor) and statistical weighting (i.e., Naïve Bayes).

### ***Generalizing relevance theories with a hypothesis ontology***

Representing knowledge associated with users' relevance hypotheses so that it can be used by an automated system is of primary importance; Hogan et al. (2009) recognize that the knowledge should be "formalized into the system via query development, vocabularies, etc." although they provide no examples. Clearly, techniques for

representing such legal knowledge and making it more broadly applicable are of great interest to AI & Law research.

As proposed by Ashley and Bridewell (2007), a “hypothesis ontology” and an “event-based framework” similar to those developed for bioinformatics (Racunas et al., 2004) could represent, in a broadly applicable form, knowledge associated with users’ relevance hypotheses. Based on Racunas’s definition, a hypothesis ontology comprises “a vocabulary of objects (agents) and processes”, associated with recurrent areas of interest in e-Discovery such as knowledge transmission in corporate or commercial settings “and define[s] the relationships in which these entities can participate.” For instance, an ontology will include key knowledge transmission concepts, such as “*communicated-with*” and “*during-interval*”, that expect arguments of a particular type and invoke an established function. An event-based framework supports specifying hypotheses about “knowledge transmission” events “by naming the agents from the ontology ... and the [knowledge transmission] processes... that connect them (Racunas et al., 2004).”

Implemented in a suitable user interface, a hypothesis ontology and event-based framework could enable users to represent their own hypotheses in a semi-formal language that specifies the time and context-dependent relationships among the objects and processes in the corpus (Racunas et al., 2004). Framed in terms from a hypothesis ontology, a rough example of a relevance hypothesis might be, “there exist documents in which Alice *communicated-with* Bob *during-interval* 1976 to 1978 that *contain-keywords* {tobacco, children, advertising}”.

The solicitation of information for building and modifying the hypothesis ontology, event-based framework, and the relevance hypotheses themselves, would be an iterative task of user modeling. Ultimately, however, the same interface could serve as the user’s interface with the retrieval system, and the core ontology, including the more general concepts of knowledge transmission in corporate or commercial settings, would be available to serve other users in other cases. The ontology and framework could also support reasoning about communications and documents, especially if supplemented with information about the social network, addressed next.

### ***Social network analysis to apply relevance theories***

Social networks, graphs representing probable social interactions among a group of people, could enable a system to infer which individuals most likely communicated information or had knowledge relevant to a query (Schwartz and Wood, 1993; Agrawal et al., 2003). Since senders, recipients, and owners of documents identify themselves through email records and the contents of their workspaces, one conceivably can build a model of the network of knowledge and, from that structure, infer something about the likely content and target of the communications. That is, by determining how each item moves through a social organization, one can map the general flow of knowledge and infer something about the relevance of the content of documents based on who has them, who sent them, and who likely read them.

As an automated technique for supplementing, representing, and reasoning about the information targeted by the process of modeling the litigators’ relevance hypotheses, social networks could make an important contribution to AI & Law. Essentially, the social network reduces one’s reliance on individual words and phrases while increasing one’s ability to infer relevance based on a history of contact. Given a social network, one

can expand the set of responsive documents by including all the documents associated with an author of relevant text or with an individual in contact with that person. For instance, suppose one knows of correspondence between Alice and Bob about nicotine addiction and similar communications between Charlie and Dana. If Bob and Dana are frequent correspondents, then they may have discussed nicotine addiction as well.

To some extent, the social networks could be constructed automatically from the documents themselves. As a first approximation to the social structure, a system can use document metadata, such as the *to* and *from* fields in email headers, to build a network of every identifiable individual and create a link only if supported by the metadata or extracted named-entities (Kossinets and Watts, 2006; see also, Eichmann and Chin, 2007). A document's unstructured text may mention names of people, companies, and other entities that belong in the network, information that could be captured with named entity recognition techniques, which identify meaningful words and phrases that fall into informative social categories (Curran and Clark, 2003). In the AI & Law community, for example, Hachey and Grover (2007) present a potential rule-based solution that recognizes named entities specific to the legal domain such as references to acts, sections, and judgments. Tools from graph mining and statistical relational learning could help to deal with problems of multiple people with the same name, individuals using multiple names and email addresses, and communications that do not imply a strong social link (e.g., interoffice memos).

In the user modeling process, litigators would edit this first order social network and supplement it with social role information. Such roles can arise through formal organizational structure (e.g., Chief Counsel) or through the organics of interoffice communication (e.g., social hub), and would aid the system in assessing the importance of a social link.

## **Two Research Challenges re Relevance Hypotheses**

If a social network could help to inform the process of relating documents to users' relevance hypotheses, one research challenge is to model dynamic change in the social network. Typically, social networks present a flat, static view of a rich, dynamic organization. Over time, communication links form and break, individuals appear and disappear, and these changes directly affect the flow of information and, in turn, the relevance of documents to a query.

Ashley and Bridewell (2007) proposed to record the distribution of correspondences over time for each social link, where the time stamps are available either from document metadata (as in the Tobacco Litigation corpus) or from email headers (as in the Enron email collection). In the above example, if the four individuals wrote about nicotine addiction within the same or overlapping time frames, then the documents between Bob and Dana deserve more attention than if Alice and Bob wrote to each other in 1981 and Bob and Dana were corresponding in 1995. A rational consideration of information dynamics could refine the system's decision to include documents via the social network and rule out documents that were created within an unrelated time frame or were communicated between irrelevant parties.

Another challenge is how to explain to users why retrieved documents are relevant in terms of their research hypotheses. Since the justifications for a document's inclusion or

exclusion from a set of results may be fairly complex, the system would need to explain the reasoning behind its decision, that is, why it infers that a set of documents bears upon a user's relevance hypothesis. A sample justification might be, "The Vice President of Marketing communicated frequently with Sara between 1985 and 1989. Sara communicated with a third party, Tom, several times about cigarettes and children. This document between Sara and the Vice President of Marketing mentions children." Explanation schemas (Thagard, 1999; Schank, 1986) could relate the support for a document's inclusion. Since documents will need to be clustered to avoid information overload, the schemas will need to be adapted to identify and integrate the clusters of documents relevant to particular aspects or portions of an explanation.

## Conclusions

The paper has focused on three emerging techniques for improving the process of automating analysis and retrieval of electronically stored information in discovery proceedings: (1) machine learning to extend and apply users' hypotheses (theories) of relevance; (2) a hypothesis ontology to generalize user modeling regarding relevance theories; and (3) social network analysis to supplement and apply user modeling regarding relevance theories. Since all three pertain to representing and reasoning with litigators' relevance hypotheses, and since a central theme of AI & Law involves computationally modeling legal knowledge, reasoning and decision making, all three techniques can be described as characteristic of that field's potentially unique contribution to e-Discovery.

As observed in (Ashley, 2007), other areas of research in AI & Law or more generally in AI may have implications for the e-Discovery problem such as:

- Textual Case-Based Reasoning (TCBR) methods to analyze electronic corporate textual documents and capture implicit structural or ordering information, for instance, using Hidden Markov Models to identify anomalous inspection reports (Mustafaraj et al., 2007).
- Natural Language Processing (NLP) techniques to recognize subjectivity and sentiment in text to identify textual documents where subjective opinions or emotionally-charged recommendations are made (e.g., Wilson, 2008).
- Techniques for visualizing evidential reasoning in argument diagrams to help litigators collaboratively construct their legal arguments *and* integrate the documents into them. (Tillers, 2007)
- Detailed legal ontologies as a basis for indexing schemes required to make legal sense of the electronic documents discovered; such ontologies can seed the hypothesis ontologies to be fleshed out from the electronic documents in the corpus and in user modeling. (See, e.g., Breuker et al. 2004).

It is apparent that applications of these techniques to the e-Discovery problem could readily be related to the theme of representing and reasoning about litigators' relevance hypotheses. (And, of course as noted, any research approach to the problem of e-Discovery that includes AI will be duly accepted as an "AI & Law" approach despite the lack of connection to that theme.) Indeed, appropriate extensions of any of the work reported in these papers could form the nucleus of a special issue of the *Journal of*

*Artificial Intelligence and Law* devoted to the e-Discovery problem, and it is hoped, this paper provides an example of a thematic rationale for such an issue.

## References

- Ashley, K. (2007) Can AI & Law Contribute to Managing Electronically Stored Information in Discovery Proceedings? Some Points of Tangency; Position Paper ICAIL 2007 Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings (DESI). Stanford.
- Ashley, K. and Bridewell, W. (2007) III-CXT-Medium: Tracing Information Flow to Improve Legal e-Discovery. NSF Proposal No. 0803049. Declined.
- Bauer, R., Jade, T., Hedin, B., and Hogan, C. (2008) Automated Legal Sensemaking: The Centrality of Relevance and Intentionality. In Proceedings 2nd Intl. Workshop on Supporting Search & Sensemaking for Electronically Stored Information in Discovery, London. June.
- Breuker, J., Valente, A., and Winkels, R. (2004) Legal Ontologies in Knowledge Engineering and Information Management. *Artificial Intelligence and Law*, V.12, No. 4, pp. 241-277. Springer.
- Brüninghaus, S., & Ashley K.D. (2005). Generating legal arguments and predictions from case texts. *Proceedings of the Tenth International Conference of AI and Law* (pp. 65–74). Bologna, Italy: ACM Press.
- Buchanan, B. and Headrick, T. (1970) Some Speculation about Artificial Intelligence and Legal Reasoning. *Stanford Law Review*, 23: 40-62.
- Daniels, J., & Rissland, E. (1997). Finding legally relevant passages in case opinions. *Proceedings of the Sixth International Conference on Artificial Intelligence and Law* (pp. 39–46). Melbourne, Australia: ACM Press.
- Gonçalves, T., & Quresma, P. (2005). Is linguistic information relevant for the text legal classification problem? *Proceedings of the Tenth International Conference on Artificial Intelligence and Law* (pp. 168–176). Bologna, Italy: ACM Press.
- Grover, C., Hachey, B., Hughson, I., & Korycinski, C. (2004). Automatic summarisation of legal documents. *Proceedings of Ninth International Conference on Artificial Intelligence and Law* (pp. 243–251). Edinburgh, Scotland: ACM Press.
- Hachey, B, & Grover, C. (2006). Extractive summarization of legal texts. *Artificial Intelligence and Law*, 14, 305–345.
- Hogan, C., Bauer, R., and Brassil, D. (2009) Human-Aided Computer Cognition for e-Discovery. in Proceedings of 12<sup>th</sup> International Conference on Artificial Intelligence and Law (ICAAIL-09). Barcelona. June.
- Jackson, P., Al-Kofahi, K., Tyrrell, A., & Vacher, A. (2003). Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150, 239–290.
- McCarty, L.T. (2007). Deep semantic interpretations of legal texts. Proceedings of the Eleventh *International Conference on Artificial Intelligence and Law* (pp. 217–224). Stanford, CA: ACM Press.
- Mustafaraj, E., Hoof, M., and Freisleben, B. (2007) Knowledge Extraction and Summarization for an Application of Textual Case-Based Interpretation. In R.O. Weber and M.M. Richter (Eds.) *Proceedings of the 6th ICCBR Conference*, ICCBR 2007, LNAI 4626, pp. 517–531.

- Pirolli, P. (2005) The Cognitive Structure of User Sense Making. In 7th International Workshop of the EU Network of Excellence: DELOS on Audio-Visual Content and Information Visualization in Digital Libraries (AVIVDiLib'05,) Cortona, Italy.
- Racunas, S.A., Shah, N.H., Albert, I., & Fedoroff, N.V. (2004). HyBrow: A prototype system for computer-aided hypothesis evaluation. *Bioinformatics*, 20, pp. i257–i264.
- Thompson, P. (2001). Automatic categorization of case law. *Proceedings of the Eighth International Conference on Artificial Intelligence and Law* (pp. 70–77). St. Louis, MO: ACM Press.
- Tillers, P., ed. (2007) Special Issue on Graphic and Visual Representations of Evidence and Inference in Legal Settings, *Law, Probability, and Risk*. 6:1-4.
- Uyttendaele, C., Moens, M., & Dumortier, J. (1998). SALOMON: Automatic abstracting of legal cases for effective access to court decisions. *Artificial Intelligence and Law*, 6, 59–79.
- Weber, R. (1998). *Intelligent Jurisprudence Research*. Federal University of Santa Catarina, Florianópolis, Brazil, unpublished doctoral dissertation.
- Wilson, T. (2008) Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States. Ph. D. Dissertation. University of Pittsburgh Intelligent Systems Program.  
<http://homepages.inf.ed.ac.uk/twilson/dissertation.html>