Challenges in Legal Electronic Discovery

CISML 2012
Dr Jeremy Pickens
Sr Applied Research Scientist
Likes: Information Retrieval, collaborative search, photography, and dancing.

Larry Barela
Product Manager
Likes: cool software, hockey, outdoors
Catalyst provides secure, scalable multi-lingual document repositories for electronic discovery, litigation support, and other complex regulatory matters. For over a decade, corporations and their counsel have relied on Catalyst to control litigation costs and make review teams more effective. Our systems and supporting services cover the heart of the litigation lifecycle—from processing and search, to analytics, review, production and trial.
1. Short Course in eDiscovery
2. Working in eDiscovery
3. Open Research Challenges in eDiscovery
1 Short Course in eDiscovery
Plaintiff: Daniel
Suing ACME corporation for suspected patent violations.

Defendant: Anne
Corporate counsel for ACME Corporation
Daniel learns about ACME corporation through a process called Discovery. Thirty years ago, Discovery was a cordial affair exchanging a handful of documents and a professional vow of thoroughness. Today, Discovery has become eDiscovery and the scale and scope of the process continues to grow.
ACME hired Anne after realizing that lawsuits were increasing and the process of protecting the company from risk had to be balanced with the proportionality of the case. Anne manages all eDiscovery cases by understanding the "Custodians", collecting data, reviewing data, and sharing data with outside law firms for human review before producing to Daniel.
Although possible, Daniel and Anne will not likely use the same tools or repositories to store the documents. Each will likely choose expert help from law firms. Those experts have favorite tools and vendors for eDiscovery. Catalyst is one of the tools that can be used. Our software can be used by both sides with very different goals.
Anne is interested in making sure key information is not produced to the other side and she will use basic search terms to look for product names or people names. Anne may be interested in the quality of the human review and use a supervised learning system to classify documents for quality control.
Examples:
Search: "Anne Smith" or "Ann Smith" or "E.Anne Smith" or "Elizabeth Anne Smith"

A Support Vector Machine is trained to look for privileged information from a statistical sample. Anne uses the inferred function to compare the human reviewers work to the results of the classifier as a form of QC.
Daniel may be interested in unsupervised feature extraction to summarize batches of documents. Daniel may use supervised learning to reduce the amount of data for human review through any number of classifiers. He will use simple search for finding important people and dates.
Working in eDiscovery
Time is an elusive commodity
Documents are plentiful and messy
(and mostly electronic and multi-language unlike this photo)
Pressure is stressful
Gamesmanship is always in play
Deadlines are real and costly
1.6 custodians per upload
1.14 GB per upload
3300 docs per upload

Batches are common
Collections are incomplete

300 Uploads
45 Days
• Time is not a lawyer's friend
• Documents are numerous
• Pressure is high
• Gamesmanship
• Real deadlines
• Constantly changing corpus
Open Research Challenges in eDiscovery
Rolling Batches
(Identically Distributed)
These are really different custodians, with new specialities and different lexicons.
Relevance
Non-Ergodic?
Poincaré recurrence theorem: Almost all points in any subset of the phase space eventually revisit the set.
Machine Learning vs Search?
Machine Learning
Learn a generalized function

Search
Find everything.
Only once.
Multiple Matters: Transfer Learning?
Biggest Open Problem: Estimating Recall