Run Descriptions: TREC 2009 Legal Track, Batch Task

Run EmcRun1 (EMC - CMA - R&D)
Task: batch
Run type: manual
Description of manual approach:
This is a manual run because bugs were found in the query evaluation software, and as a result the system received one hot-fix, and some queries were adapted to avoid hitting bugs I could not address in time. I did not review any documents, or made any form of direct evaluation of the results.
Fields: request text, final negotiated Boolean query, complaint
Document fields indexed: OCR, metadata
Boolean reference run used? no
Past relevance judgments used? no
Number of past relevant documents used? no
Judging order: 1
Description of this run:
For different reasons, the time available for this experiment was /much/ shorter than desirable. Therefore the experiment had to be kept as simple as possible. A list of noteworthy points about our run: - relevance calculation was based upon TF/IDF; - a list of 500 stop-words was filtered out of indexing; - indexing and searching was case-insensitive; - the final boolean queries were expanded manually after consulting the request and the complaint.

Run ucedlsi (Ursinus College)
Task: batch
Run type: manual
Description of manual approach:
No.
Fields: request text
Document fields indexed: OCR
Boolean reference run used? no
Past relevance judgments used? yes
Number of past relevant documents used? no
Judging order: 2
Description of this run:
This run is distributed EDLSI. The indexed dataset is divided into 81 pieces, and Essential Dimensions of Latent Semantic Indexing (EDLSI), a weighted sum of the result from Latent Sematic Indexing (LSI) and vector space information retreival (IR) is applied to each. The scores are then compiled and sorted. The K values are chosen to represent approximately the number of documents at which the number common documents between the three runs is maximized. K_h is chosen at the approximate point where the document scores drop off.

Run uclsi (Ursinus College)
Task: batch
Run type: manual
Description of manual approach:
No.
Fields: request text
Document fields indexed: OCR
Boolean reference run used? no
Past relevance judgments used? yes
Number of past relevant documents used? no
Judging order: 1
Description of this run:
This run is LSI with folding-in. The indexed dataset is divided into 81 pieces, and Latent Sematic Indexing (LSI) is performed on the first piece. The remaining 80 pieces are folded in, and the scores computed and sorted. The K values are chosen to represent approximately the number of documents at which the number common documents between the three runs is maximized. K_h is chosen at the approximate point where the document scores drop off.

Run ucscra (Ursinus College)
Task: batch
Run type: manual
Description of manual approach:
No.
Fields: request text
Document fields indexed: OCR
Boolean reference run used? no
Past relevance judgments used? yes
Number of past relevant documents used? no
Judging order: 3
Description of this run:
This run is SCRA-based distributed LSI. The indexed dataset is divided into 81 pieces, and each piece is further divided into 40 pieces. Latent Sematic Indexing (LSI) is applied to each piece, but using the Sparse Column-Row Approximation (SCRA) instead of the Singular Value Decomposition (SVD) which is traditionally used. The scores are then compiled and sorted. The K values are chosen to represent approximately the number of documents at which the number common documents between the three runs is maximized. K_h is chosen at the approximate point where the document scores drop off.

Run watlogistic (University of Waterloo)
Task: batch
Run type: automatic
Fields:
Document fields indexed: OCR
Boolean reference run used? no
Past relevance judgments used? yes
Number of past relevant documents used? no
Judging order: 3
Description of this run:
Logistic regression using all training examples as described for watrrf. All training examples were used; no cross-validation. K values were estimated from a separate run (not submitted) using the same method with 2-fold cross-validation.

Run watrrf (University of Waterloo)
Task: batch
Run type: automatic
Fields:
Document fields indexed: OCR
Boolean reference run used? no
Past relevance judgments used? yes
Number of past relevant documents used? no
Judging order: 1
Description of this run:
Reciprocal rank fusion of several ranking methods: BM25 relevance feedback, Naive Bayes, online logistic regression, batch logistic regression. 2-fold cross validation (splitting examples into equal test and validation sets) was used to determine K. Topic was not used at all. Training examples and features as per watlogistic.

Run watstack (University of Waterloo)
Task: batch
Run type: automatic
Fields:
Document fields indexed: OCR
Boolean reference run used? no
Past relevance judgments used? yes
Number of past relevant documents used? no
Judging order: 2
Description of this run:
Same as watrrf, but classifiers were stacked using logistic regression and 2-fold cross validation.

Run otL09F (Open Text Corporation)
Task: batch
Run type: manual
Description of manual approach:
0
Fields:
Document fields indexed: OCR, metadata
Boolean reference run used? no
Past relevance judgments used? yes
Number of past relevant documents used? yes
Judging order: 2
Description of this run:
pure relevance feedback run based on forming a query from a random sample of the known relevant documents of size less than 10000 bytes; no topic fields were used; the K value was set to the greater of the retrospective optimal K value for F1 and the estRelL09.append K value, plus 10 percent; the Kh values were just taken from estRelL09.append

Run otL09frwF (Open Text Corporation)
Task: batch
Run type: manual
Description of manual approach:
0
Fields: request text, final negotiated Boolean query
Document fields indexed: OCR, metadata
Boolean reference run used? yes

Past relevance judgments used? yes
Number of past relevant documents used? yes
Judging order: 1
Description of this run:
rrf-based fusion of feedback (otL09F weight 3), ranked final Boolean (weight 3), request text vector (otL09rvl weight 2), and vector of final Boolean terms (weight 1); the K value was set to the greater of the retrospective optimal K value for F1 and the estRelL09.append K value, plus 10 percent; the Kh values were just taken from estRelL09.append

Run otL09rvl (Open Text Corporation)
Task: batch
Run type: manual
Description of manual approach:
0
Fields: request text
Document fields indexed: OCR, metadata
Boolean reference run used? no
Past relevance judgments used? no
Number of past relevant documents used? no
Judging order: 3
Description of this run:
baseline run (no feedback); vector run based on request text terms; English inflections were matched; common instruction words (e.g. "please", "produce", "documents") were manually removed; K includes rsv of 200 or more, Kh includes rsv of 225 or more