System Summary and Timing
Organization Name: City University
List of Run ID's: citya1 citya2 cityr1 cityr2 cityi1

- Construction of Indices, Knowledge Bases, and other Data Structures
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list:
      - 247 (citya1 citya2)
    - Length (in words) of the stopword list:
      - 247 + 226 semi-stopwords (cityr1 cityr2 cityi1)
    - Stemming Algorithm: Modified Porter with spelling normalization
  - Statistics on Data Structures built from TREC Text
    - Inverted index
      - Run ID : citya1 citya2
      - Total Storage (in MB): 1200
      - Total Computer Time to Build (in hours): about 25 hours
      - Automatic Process? (If not, number of manual hours): Yes
      - Use of Term Positions? : Yes
      - Only Single Terms Used? : Some prespecified phrases are mapped to tokens
    - Inverted index
      - Run ID : cityr1 cityr2 cityi1
      - Total Storage (in MB): 530
      - Total Computer Time to Build (in hours): about 14 hours
      - Automatic Process? (If not, number of manual hours): Yes
      - Use of Term Positions? : Yes
      - Only Single Terms Used? : Some prespecified phrases are mapped to tokens
  - Data Built from Sources Other than the Input Text
    - Internally-built Auxiliary File
      - Domain (independent or specific): Somewhat angled towards American data
      - Type of File (thesaurus, knowledge base, lexicon, etc.): Contains synonym classes, go phrases, stop- and semi-stopwords, prefixes
      - Total Storage (in MB): ¡¡ 1
      - Number of Concepts Represented: About 1500
      - Type of Representation: Lookup table
      - Total Manual Time to Build (in hours): Not recorded,done haphazardly. Much of it is common to many databases
      - Total Manual Time to Modify for TREC (if already built): Not known
      - Use of Manual Labor

        I. Other: Yes
- Query construction

- Automatically Built Queries (Ad-Hoc)
  * Run Id: citya1
  * Topic Fields Used: Title, narrative, description
  * Average Computer Time to Build Query (in cpu seconds): About 500
  * Method used in Query Construction
    · Term Weighting (weights based on terms in topics)? : See below
    · Other: Addition of terms from top documents retrieved by trial search using topic fields. Terms weighted based on term frequency in topics and in documents, collection frequency, and frequency in documents retrieved by trial search.
- Automatically Built Queries (Ad-Hoc)
  * Run Id: citya2
  * Topic Fields Used: Title, narrative, description
  * Average Computer Time to Build Query (in cpu seconds): About 2
  * Method used in Query Construction
    · Term Weighting (weights based on terms in topics)? : See below
    · Other: Terms weighted based on term frequency in topics and in documents, and collection frequency.
- Automatically Built Queries (Routing)
  * Run Id: cityr1 cityr2
  * Topic Fields Used: None
  * Average Computer Time to Build Query (in cpu seconds): Thousands
  * Method used in Query Construction
    · Terms Selected From
      I. Only Documents with Relevance Judgments: Yes (only from relevant documents)
    · Term Weighting with Weights Based on terms in
      I. Only documents with Relevance Judgments: Yes (only relevant documents).
- Interactive Queries
  * Initial Query Built Automatically or Manually: Manually
  * Type of Person doing Interaction
    · Other: Information scientists and graduate students.
  * Average Time to do Complete Interaction
    · CPU Time (Total CPU Seconds for all Iterations): Unknown
    · Clock Time from Initial Construction of Query to Completion of Final Query (in minutes): 39
  * Average Number of Iterations: 1.5
  * Average Number of Documents Examined per Iteration: 18.5
  * Minimum Number of Iterations: 0
  * Maximum Number of Iterations: 3
  * What Determines the End of an Iteration: searcher decision
  * Methods used in Interaction
    · Automatic Term Reweighting from Relevant Documents? : Yes

        - Automatic Query Expansion from Relevant Documents? : Yes
          I. Only Top X Terms Added (what is X): Yes; 20
          II. User Selected Terms Added: Yes
        - Manual Methods
          I. Using Individual Judgment (No Set Algorithm)? : Yes

- Searching
  - Search Times
    - Run ID : citya1 citya2 cityr1 cityr2
    - Computer Time to Search (Average per Query, in CPU seconds): Varies enormously, from about 10 seconds with 3 or 4 terms up to a few minutes. Would guess overall average about 60 on SS10
    - Run ID : cityi1
    - Computer Time to Search (Average per Query, in CPU seconds): Unknown
  - Machine Searching Methods
    - Machine Searching Methods
      - Probabilistic Model? : Yes
  - Factors in Ranking
    - Factors in Ranking
      - Term Frequency? : Yes
      - Inverse Document Frequency? : Yes
      - Other Term Weights? : From relevance information when available
      - Document Length? : Yes
  - Machine Information
    - Machine Type for TREC Experiment: Suns. SS10, 4/330, IPX
    - Was the Machine Dedicated or Shared: SS10 dedicated, others shared.
    - Amount of Hard Disk Storage (in MB): About 11 GB for TREC
    - Amount of RAM (in MB): 64, 40, 16 resp.
    - Clock Rate of CPU (in MHz): N/K
  - System Comparisons
    - Amount of "Software Engineering" which went into the Development of the System: A lot, but very little which is TREC-specific.
    - Given appropriate resources
      - Could your system run faster? : Of course.
      - By how much (estimate)? : Order of magnitude?
    - Features the System is Missing that would be beneficial: Factors X and Y.
  - Significant Areas of System
    - Brief Description of features in your system which you feel impact the system and are not answered by above questions:
      - Both the databases had enough within-document information to allow searching sub-documents consisting of any run of algorithmically-determined paragraphs; this facility was used in the citya1 run.