System Summary and Timing
Organization Name: Siemens Corporate Research, Inc.
List of Run ID's: siems1 siems2

- Construction of Indices, Knowledge Bases, and other Data Structures
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list: 571
    - Stemming Algorithm: standard SMART
    - Term Weighting: yes (cosine-normalized tf)
    - Phrase Discovery? : 011
  - Statistics on Data Structures built from TREC Text
    - Inverted index
      - Run ID : siems1, siems2
      - Total Storage (in MB): 201
      - Total Computer Time to Build (in hours): 1.8
      - Automatic Process? (If not, number of manual hours): yes
      - Use of Term Positions? : no
      - Only Single Terms Used? : yes
    - Inverted index
      - Run ID : siems1, siems2
      - Total Storage (in MB): 95
      - Total Computer Time to Build (in hours): .75
      - Automatic Process? (If not, number of manual hours): yes
      - Use of Term Positions? : no
      - Only Single Terms Used? : yes
    - Inverted index
      - Run ID : siems1, siems2
      - Total Storage (in MB): 72
      - Total Computer Time to Build (in hours): 1.5
      - Automatic Process? (If not, number of manual hours): yes
      - Use of Term Positions? : no
      - Only Single Terms Used? : yes
    - Inverted index
      - Run ID : siems1, siems2
      - Total Storage (in MB): 194
      - Total Computer Time to Build (in hours): 1.4
      - Automatic Process? (If not, number of manual hours): yes
      - Use of Term Positions? : no
      - Only Single Terms Used? : yes
    - Inverted index
      - Run ID : siems1, siems2
      - Total Storage (in MB): 130
      - Total Computer Time to Build (in hours): 2.2

            - Automatic Process? (If not, number of manual hours): yes
            - Use of Term Positions? : no
            - Only Single Terms Used? : yes
        * Knowledge Bases
            - Automatic Process? (If not, number of manual hours): yes
    - Data Built from Sources Other than the Input Text
        * Internally-built Auxiliary File
            - Domain (independent or specific): collection specific
            - Type of File (thesaurus, knowledge base, lexicon, etc.): siems1: ranks of relevant docs in training queries siems2: training query clusters
            - Total Storage (in MB): siems1: .1 per collection siems2: .025 per collection
            - Total Computer Time to Build (in hours): .5 per collection including doing retrieval

- Query construction
    - Automatically Built Queries (Ad-Hoc)
        * Topic Fields Used: title, nat, narr, fac, desc, con
        * Average Computer Time to Build Query (in cpu seconds): .03 per collection
        * Method used in Query Construction
            - Term Weighting (weights based on terms in topics)? : yes, (cosine-normalized tf*idf, within each collection)

- Searching
    - Search Times
        * Run ID : siems1
        * Computer Time to Search (Average per Query, in CPU seconds): 42 (assuming parallel execution of searches)
        * Component Times : index query in fusion's query collection – .03 (once) index query in 5 subcollections – .03 (per collection) retrieve 1000 docs on each subcollection – 11 (slowest single collection) fuse –30 (once)
        * Run ID : siems2
        * Computer Time to Search (Average per Query, in CPU seconds): 12 (assuming parallel execution of searches)
        * Component Times : index query in 5 subcollections – .03 (per collection) retrieve 1000 docs on each subcollection – 11 (slowest single) fuse – .5
    - Machine Searching Methods
        * Machine Searching Methods
            - Vector Space Model? : yes
            - Cluster Searching? : query clusters in siems2
    - Factors in Ranking
        * Factors in Ranking
            - Term Frequency? : yes
            - Inverse Document Frequency? : yes

        - Document Length? : yes

        - Other: in siems1, expected usefulness of subcollection

- Machine Information
    * Machine Type for TREC Experiment: SPARC-10/41
    * Was the Machine Dedicated or Shared: mostly dedicated
    * Amount of Hard Disk Storage (in MB): 13,000
    * Amount of RAM (in MB): 128
    * Clock Rate of CPU (in MHz): 40 MHz
- System Comparisons
    * Amount of "Software Engineering" which went into the Development of the System: individual subcollection retrieval done by SMART, which is a well-tuned research prototype; fusion code experimental with almost no tuning
    * Given appropriate resources
        - Could your system run faster? : yes
        - By how much (estimate)? : with some effort, fusion time for method in siems1 could be halved