System Summary and Timing
Organization Name: NEC
List of Run ID's: virtu1, virtu2

- Construction of Indices, Knowledge Bases, and other Data Structures
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list: 430
    - Stemming Algorithm: Yes
      - Morphological Analysis: Yes
    - Term Weighting: Yes
  - Statistics on Data Structures built from TREC Text
    - Inverted index : Yes
      - Run ID : virtu1, virtu2
      - Total Storage (in MB): 2056
      - Total Computer Time to Build (in hours): 214
      - Automatic Process? (If not, number of manual hours): Yes
      - Use of Term Positions? : Yes
      - Only Single Terms Used? : Yes
  - Data Built from Sources Other than the Input Text
    - Internally-built Auxiliary File : Yes
      - Domain (independent or specific): independent
      - Type of File (thesaurus, knowledge base, lexicon, etc.): lexicon, inflection table
      - Total Storage (in MB): 4
      - Number of Concepts Represented: 143360
      - Type of Representation: look up table
      - Total Computer Time to Build (in hours): Not recorded
      - Total Computer Time to Modify for TREC (if already built): 0.5
- Query construction
  - Automatically Built Queries (Ad-Hoc) : Yes
    - Topic Fields Used: All
    - Average Computer Time to Build Query (in cpu seconds): ¿ 10 sec per query
    - Method used in Query Construction
      - Term Weighting (weights based on terms in topics)? : Yes
      - Phrase Extraction from Topics? : Yes
      - Syntactic Parsing of Topics? : Yes
      - Proper Noun Identification Algorithm? : Yes
      - Tokenizer? : Yes
        I. Patterns which are Tokenized: part of noun phrase identification
      - Automatic Addition of Boolean Connectors or Proximity Operators? : Yes
  - Manually Constructed Queries (Routing) : Yes

    * Topic Fields Used: All
    * Average Time to Build Query (in Minutes): 30 min per query
    * Type of Query Builder
        · Domain Expert: No
        · Computer System Expert: Yes
    * Data Used for Building Query from
        · Training Topics: Yes
        · All Training Documents: Yes
        · Only Documents with Relevance Judgments: Yes
    * Method used in Query Construction
        · Term Weighting? : Yes
        · Boolean Connectors (AND, OR, NOT)? : Yes
        · Proximity Operators? : Adjacency operator used
        · Addition of Terms not Included in Topic? : Yes
        I. Source of Terms: relevant document

- Searching
  - Search Times
    * Run ID : virtu1, virtu2
    * Computer Time to Search (Average per Query, in CPU seconds): 1200
  - Machine Searching Methods
    * Vector Space Model? : Yes
    * Probabilistic Model? : Yes
    * Boolean Matching? : Yes partially
  - Factors in Ranking
    * Factors in Ranking
        · Term Frequency? : Yes
        · Inverse Document Frequency? : Yes
        · Proximity of Terms? : Adjacency used
        · Document Length? : Yes
  - Machine Information
    * Machine Type for TREC Experiment: sparc10
    * Was the Machine Dedicated or Shared: shared
    * Amount of Hard Disk Storage (in MB): 8000
    * Amount of RAM (in MB): 90
    * Clock Rate of CPU (in MHz): 40
  - System Comparisons
    * Amount of "Software Engineering" which went into the Development of the System: It was a first prototype. Three person month were spent.
    * Given appropriate resources
        · Could your system run faster? : Yes
        · By how much (estimate)? : Main part of CPU time is disk I/O. If separate inverted files are used, 20-40of CPU time may be decreased.

    * Features the System is Missing that would be beneficial: Document Clustering, General Proximity Operation

- Significant Areas of System
    * Brief Description of features in your system which you feel impact the system and are not answered by above questions:
        · Morphological analysis on making inverted files, Noun phrase detection on automatic query Construction (ad-hoc), Adjacent word detection