Search Engine for Shoah Foundation

Outline
- Shoah Foundation
- Our project
- What we need

Shoah Foundation
- The USC Shoah Foundation Institute for Visual History and Education, with an archive of nearly 52,000 videotaped testimonies from Holocaust survivors and other witnesses, is part of the College of Letters, Arts & Sciences at the University of Southern California.
- The USC Shoah Foundation Institute’s Visual History Archive (VHA) is a software tool that allows users to search data and view digital video in the USC Shoah Foundation Institute’s archive.

Shoah Foundation
- A segment is a one-minute unit of a testimony in the VHA. Testimonies are divided into one-minute segments which can be retrieved by the end user through keyword searches.
- Not every segment has keywords attached.
- Keywords are attached to one-minute segments when a topic is discussed or described in some detail.
  - If the discussion or description spans several segments, the relevant keywords are usually applied once.

Shoah Foundation
- Testimonies can be searched based on keywords
  - You can choose to perform an AND or OR search
    - The AND search retrieves segments that include all of your Selected Keywords (up to 35 keywords).
    - The AND search also permits you to choose a Segment Range. It is possible to search for all Selected Keywords appearing in the same segment (i.e. 1 segment), within 5 (consecutive) segments, within 10 (consecutive) segments, within 15 (consecutive) segments or within the Entire testimony.

Shoah Foundation
- Two types of search
  - Quick search
    - Regular keyword search
  - Global keyword search
    - search for segments of testimonies that discuss specific topics
    - Topics are predefined
    - using more than 50,000 geographic and experiential keywords
Our project

- Robust, efficient and interactive search engine ranking testimonies based on combination of
  - Textual (regular) keywords
  - Spatial keywords
  - Temporal keywords
- This search engine finds and ranks the most textually, spatially and temporally relevant testimonies (segments) according to
  - query keywords
  - query location
  - query time interval

Input

- Query Keywords
  - Set of keywords inputted as text
- Query Location
  - A region drawn on the map OR
  - A spatial keyword inputted as text
- Query time interval
  - An interval specified by a time slider OR
  - An interval inputted as two numbers

Output

Tasks

1- Data tier
  - Data Cleansing
    - Understand / format / standardize the data
  - Geocoding
    - Find missing lat/long information for some spatial keywords
  - Index Construction
    - Create inverted files for regular keywords
    - Create inverted files for spatial keywords
Tasks

2- Middle tier
   - Intelligent web-services
     • Talk to interface
       – Receive input (query parameters)
       – Send output (query result)
     • Talk to data tier
       – Get data
       – Access index
     • Perform necessary operations
       – Process data
       – Calculates scores
       – Format the results

3- Interface (GUI)
   - User friendly interface to receive input from the user
     • Textbox for textual keywords
     • Map interface to draw/show query location
       – A textbox can be used to input a location's name
     • Time slider to specify time interval
       – A textbox can be used to input time interval
   - Displays the result dynamically and interactively
     • Results should be changed on-the-fly based on map location and time slider
     • Provides mechanism to show the testimonies from the interface
       • Show testimonies on the same page
       • Link to a new page for showing the testimonies

4- Research/Algorithm
   - Hybrid index structure
     • captures spatial and textual keywords (probably using inverted files) as well as temporal keywords
   - Relevance ranking function
     • Formulas for spatial, textual and temporal scores
     • A combined scoring function with different weights for different features

Expertise

- Database
  - Sybase
  - SQL
- Web-services
  - ASP.NET
  - servlets / jsp
- Interface/GUI
  - Ajax
  - Google maps API
  - XHTML / CSS
- Research
  - Information retrieval
  - Spatial keyword queries