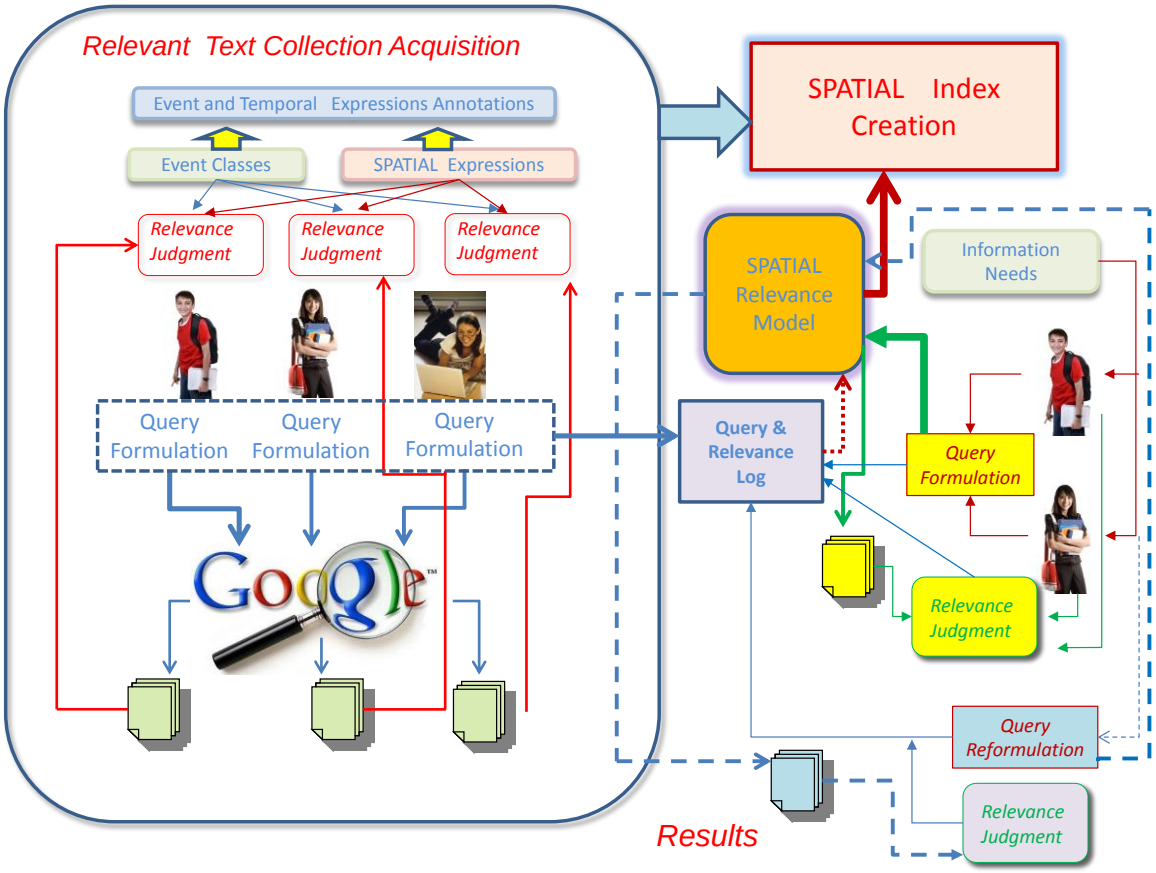
University of Texas at Dallas
CS 6322 : Information Retrieval
Fall 2014
Instructor: Dr. Sanda Harabagiu
Grader: Travis Goodwin

Project 2: Spatial Information Retrieval

In this project, an information retrieval architecture involving several modules has to be built. The main characteristic of this project is that it builds a modern retrieval framework that takes into account spatial information expressed in documents. For this reason, spatial expressions need to be identified and events that are spatially grounded be included in a model of relevance.

Each team that shall develop this project shall be provided with (a) a class of events that shall be of interest (e.g. related to the Ebola outbreak) as well as (b) a pointer to open-source software capable of automatically recognizing and normalizing temporal expressions. The architecture of the Temporal IR project is:



First a large set of relevant paragraphs needs to be acquired by taking into account: the information needs stated for the project as well as the relevance factors given for this project. For this purpose, you will formulate queries and use Google to retrieve and judge paragraphs. For each information need you will create a separate log of (a) the queries you have formulated; (b) the paragraphs you have selected and the judgments you have performed when taking into account the relevance factors that you have identified in the paragraph.
An incremental index is produced that considers (1) paragraphs and (2) the relevance factors that were identified. Paragraph indexing will operate on (1) the documents that are produced by the data acquisition task; and (2) by acquiring additional paragraphs using the same or similar queries and automatically identifying relevance factors based on quasi-pattern matching with the factors recognized by students that perform the data acquisition.
A spatial relevance model is developed such that it allows spatial expressions or operators (e.g. AROUND, BEHIND) to be used in the query as well as the other relevance factors.

When the search engine is built and the index and the relevance model are operational, the queries that were used to acquire data shall be used against the search engine and the results shall be judged. An additional set of 25 queries, provided by the team that generates the relevance model shall be also judged.

During the demo, a select 5 queries from the acquired set shall be presented as well as 5 additional queries. The report shall include all materials about the entire judged set. The judgments shall be produced in a format provided to the project.

The Information Needs for this project are:
1/ Find where the first cases of Ebola occurred in 2014
2/ Find which human organs are affected by Ebola ? Human organs are also considered spatial information.
3/ Find where are symptoms monitored for Ebola, Malaria, Typhus and Cholera?
4/ Where were US patients were the first successfully treated for Ebola?
5/ Where was preventive care recommended for Ebola, Malaria, Typhus and Cholera?
6/ Where are viruses for Ebola, Hepatitis B, Cholera, Typhus, Malaria considered active?
7/ Which preventive measures are recommended for traveling to Africa?
8/ Where do people have to stay in quarantine after being in contact with a patient that tested positive for Ebola, Malaria, Typhus or Cholera?
9/ Which organs are first attacked in the case of Ebola, Malaria, Typhus or Cholera?
10/ Which hospitals treat patients with Ebola, Malaria, Typhus or Cholera?

To build your search engine you will consider the following factors of relevance:

X1=Diagnosis (disease)
X2= Symptoms/Signs
X3= Tests

X4= Treatment
X5 = Prevention
X6= Body organ (e.g. skin)
X7= Location (country, hospital, etc)
X8 = Infectious agent
X9 = Spatial Signal (e.g. around, near, behind)

Examples for the recognition of the relevance factors in acquired paragraphs are provided.

Tasks:

TASK 1: 2-3 students shall work on the relevant text collection acquisition to (1) formulate queries (manually) , (2) select paragraphs from those retrieved by a search engine (manually) (3) produce judgments and record the judgments (positive or negative) as well the explanations in the provided format, which considered a relevance factor X_k(name of factor)= Text expression followed by ";".
A minimum of 10 queries that enable the acquisition of relevant paragraphs, which contain a minimum of 3 relevant factors and are judged either positive or negative need to be formulated independently by each student performing this task for each of the 10 information needs before November 15th, and a minimum of 10 additional queries as well as their judgments documents need to be provided to the rest of the team before November 22nd.
The students performing this task will also produce judgments of the results of the search engine in the period 1-5 December. They will also write the description of their work in the task for the final report and provide their data for the deliverable of the project.

TASK 2: 2 students: One student shall develop the incremental paragraph index by deciding the representation of the relevance factors in the index and by acquiring automatically additional paragraphs. The second student will develop software that will read the paragraphs, the judgments and the relevance factors provided by the students performing Task 1 as well as additional software that identifies automatically factors in the new paragraphs by matching them against factors recognized by the students performing Task 1.
The students performing this task will also generate additional queries to be tested by students from Task 1 who will produce judgments of the results of the search engine in the period 1-5 December. The index format needs to be decided together with the student performing Task 3. The students performing Task 2 They will also write the description of their work in the task for the final report and provide their software for the deliverable of the project.

TASK 3: 2 students. Both students shall design the spatial relevance model and build the Temporal IR systems that uses the paragraph index , the relevance factors and will be able to rank high the paragraphs that were judged positive and low those that were judged negative. They will collaborate with the students performing Task 2. They will also formulate additional set of 25 queries and explain the performance of

their relevance system based on the judgments performed by students from Task 1. The system needs to be completed before December 1st.
The students performing this task will also generate an interface for the demo of the project, such that any query can be entered and the results of the system can be presented in a browser. They will also write the description of their work in the task for the final report and provide their software for the deliverable of the project.

Project Presentation:
A powerpoint presentation shall be provided to the entire class as well as a demo. Additionally, the report shall be delivered along with the CD with the report, data, software and README file describing how the software can be used. The work and contribution of each student shall be described in the report.