SINAI-UJAEN system overview
Document collection and query preprocessing

• **Query** preprocessing
  – **Goal**: to detect and recognize the *geographical scope* and the *spatial relationship* (if any)
  – **Tools**:
    • TreeTagger (POS tagger)
    • Lexical syntactic rules (e.g. *preposition*+*proper noun*)
    • Geo-NER (based on GeoNames and Wikipedia)
    • Stopwords and Snowball stemmer

• **Document collection** preprocessing
  – **Goal**: to identify the geographical entities in each document
  – Generation of two main indexes:
    • **Geographical** (using Geo-NER)
    • **Textual** (stopwords, Snowball stemmer), including proper nouns
Indexing and retrieval

• We have used Terrier as a search engine
  – Weighting scheme: \textit{inL2} (implemented by default in Terrier)
  – Automatic query expansion: \textit{Bo1} (implemented by default in Terrier)
  – 1,000 documents retrieved per query
Filtering and reranking

- **Goal**: to modify the RSV score obtained for each document according to the *geographical similarity* with the query

- Geographical similarity query-document

\[
\text{sim}_{\text{geo}}(q,d) = \frac{\sum_{i \in \text{geoEnts}(d)} \text{match}(i, GS, SR) \cdot \text{freq}(i, d)}{|\text{geoEnts}(d)|}
\]

- \( \text{geoEnts}(d) \): geographical entities found in document \( d \)
- \( \text{match}(i, GS, SR) \): returns 1 if the geographical entity \( i \) of the document \( d \) satisfies the geographical scope of the query (\( GS \)) and its spatial relationship (\( SR \)); 0 otherwise
- \( \text{freq}(i, d) \): means frequency of the geographical entity \( i \) in the document \( d \)
Filtering and reranking

• **Reranking**:  
  – if $\text{sim}_{geo}(q,d) = 0$ then the document $d$ keeps its original RSV score (it is not reranked)  
  – If $\text{sim}_{geo}(q,d) > 0$ then it is calculated a new RSV score (RSV’)

$$RSV'_d = RSV_d + \log(RSV_d) + \text{sim}_{geo}(q,d)$$