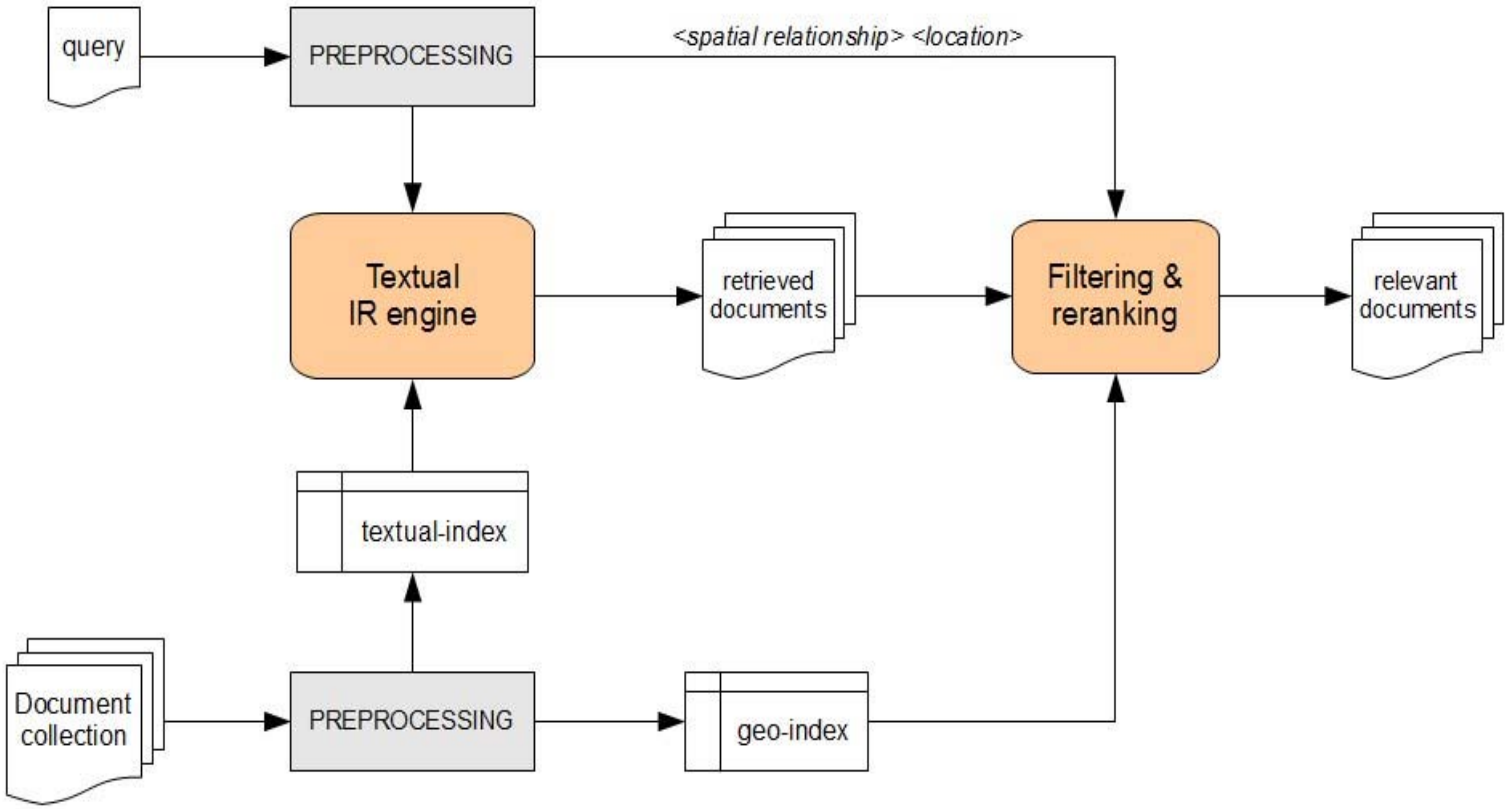
# SINAI-UJAEN system overview



# 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: ***inL2*** (implemented by default in Terrier)
  - Automatic query expansion: ***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

$$sim_{geo}(q,d) = \frac{\sum_{i \in geoEnts(d)} match(i,GS,SR) \cdot freq(i,d)}{|geoEnts(d)|}$$

  - ***geoEnts(d)***: geographical entities found in document *d*
  - ***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
  - ***freq(i,d)***: means frequency of the geographical entity *i* in the document *d*

# Filtering and reranking

- Reranking:
  - if $sim_{geo}(q,d) = 0$ then the document *d* keeps its original RSV score (it is not reranked)
  - If $sim_{geo}(q,d) > 0$ then it is calculated a new RSV score (RSV')

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