Extract relevant text/sentences from Wikipedia using Knowledge Base

The Knowledge Base contains facts in tuple format of the form

*Subject Predicate Object*

We use these facts to extract similar predicates from Wikipedia

Using Lucene search Object in the subjects Wikipedia page. Pick up the sentence(s) containing the object

Input: Wikipedia Text corpus + Knowledge Base

Output: Sentence(s), tuple pair

|  |  |
| --- | --- |
| Input | Output |
| * Wikipedia Text corpus * Knowledge Base | Knowledge Base Tuple, Sentence(s) |

Uploading data to Neo4j as graph for the purpose of querying

After we collect the predicate-phase pair with their category information. Upload the data in Neo4j Graph Database using the Neo4j batch importer tool

The categories form the nodes

The bag of predicates along with the actual predicate form the edges

|  |  |
| --- | --- |
| Input | Output |
| * Categories Hierarchy graph as Nodes list * Bag of Predicates as Edges | * Graph in Neo4j with Categories as nodes and Bag of Predicates on Edges |

Generate categories of subject… and everything else

1. Using the *subject* provided by the user generate all of its categories.
2. Pick those categories which have the input *predicate* in their edges
3. Now traverse the graph to reach the destination *object-category*
4. For all the matching edges in path take the most weighted phrase and search it in the *subject’s* Wikipedia Page/Knowledge Base for all the matching predicates. If the subject and predicate in the sentence match with what we have then the output is the object entity in that sentence.

|  |  |
| --- | --- |
| Input | Output |
| * User Query as : *subject predicate object-category* | * The corresponding *object* |