Indexing with Named Entities Recognition

Information Retrieval Mini-Project Report

Paarth Iyer (MCS202218)
Sagnik Dutta (MCS202112)
Shubhankar Varshney (MCS202204)

What is NER?

A named entity is a physical or abstract object that can be denoted by a proper name, such as John Doe, Paris, 15th of December, UNESCO etc. With NER, our goal is to tag these entities from our documents based on some predefined classes.
For our implementation, we are using the Stanford NER, with a 3-class model. (PERSON, ORGANIZATION, LOCATION)
For example :

Vincent Willem van Gogh was born on 30 March 1853 in Groot-Zundert, in the predominantly Catholic province of North Brabant in the Netherlands. He was the oldest surviving child of Theodorus van Gogh, a minister of the Dutch Reformed Church, and his wife, Anna Cornelia Carbentus.

(We are using the python 'sner' package)
In our code, before running anything, we start a server for the Stanford NER with an option "split-hyphenated = false". This just helps avoid split names with hyphens in them. Now, for the text we want to tag, we send the text to the server. It returns the text with each word labelled as one of the four: 'O' (meaning no tag), 'PERSON', 'LOCATION', 'ORGANIZATION'. We remove the words labelled 'O', attach consecutive words with same labels to join the names (since each word is tagged separately) and remove repeats. We could keep repeats and they would help with slightly higher ranking of relevant results depending on type of implementation, but it becomes a tradeoff with memory usage.

Indexing with SOLR

For our example, we initialize our solr core with two fields, namely "headlines" and "text", both with the field type 'text_en' to perform some general stopword removal and stemming. Each doc is tagged using NER and then all the documents are indexed with the mentioned fields along with the fields formed by the tags, using default solr settings.

Querying

We have implemented a basic search which boosts the named entity in its tag.
We first tag the entities in the query. Due to some limitations of the SNER model, we need to use proper capitalization in the query so that SNER can recognize the entities. This can be mitigated by using a caseless model (seems to perform worse) or using a True case annotator (would require us to use coreNLP along with).
We then use this info to form a solr query with the named entities being boosted by a factor of 2. We then send this query to solr and return the results.

Some improvements can be made

- Using a caseless model or a True case annotator while querying (as discussed above)
- Using a class hierarchy, so for example, a 'brand' class could be referenced as a 'name' or and 'organization', so search the name in all of those fields for more relevant results.
- In the current implementation, the query searching is done with individual words instead of the names as queries. Searching them as queries would increase matching
- Saving frequencies of an entity in a document to boost it more in the results

What else can we do?

- We could also make a field which stores the 'actions'/verbs in a document. So, if we are searching for a person/place also with some action, we get more relevant results. This effect can be made even stronger if we consider the correlation with the entities which are related to the action.
- We can ask who/when/where questions and return results based on the frequencies in the relevant fields.
- We could calculate correlation between entities and use that information to better the queries.