Entity Search App

Materials Needed

In addition to this document, you should have received the following two files:

1. `entities.json`
2. `companies.csv`

The CSV is a list of company names and websites. The JSON file contains a subset of the entities in our system (i.e. the "nouns" of our Intelligence Graph). In the JSON file, the keys are the entity IDs, which are unique identifiers we use internally, and the values are the entity attributes such as name and type of entity.

Acceptance Criteria

Your task is to build a web app that takes a CSV as input and returns the best matched entity from the JSON for each row or indicates that there is no reasonable match. Because some company names can be ambiguous, the mapped entities should be presented in a way that a user can easily review and verify whether the quality is adequate. The tool also needs to be able to handle the case where users have made multiple submissions to be tracked separately.

The standard user workflow is the following:

1. Submit a CSV to be mapped to entities.
2. Review mapped results.
3. Approve or Reject submission.

Users must be able to:

1. Upload CSVs.
2. In the UI, view which CSV rows match which entities in the mock ontology.
3. Download an updated CSV indicating which rows matched which entities in the mock ontology.
4. In the UI, view all CSVs uploaded/submitted.
5. In the UI, view a summary of results for each uploaded CSV (e.g. total rows, duplicates, matches, etc.).
6. In the UI, update an uploaded CSV's status from "Pending Review" to "Under Review", "Approved," or "Rejected".

In short, this is a tool for assessing results of a bulk fuzzy search. The user uploads search terms via CSV; the JSON data in "rf-company-data.zip" consists of the potential search matches; and the UI is means of reviewing results of bulk searches performed.

Tips

Given enough time to prepare results, a brute force approach of hardcoding the matching entities for each provided company name is possible. However, in the real world this tool would need to be able to handle arbitrary names and find the best match. This is a challenging problem, and we do not expect you to get all matches. We are much more interested in your ability to build an intuitive tool that finds matches for some cases and makes it easy for users to manually verify than we are in the accuracy of your matches.

Our data model can be complicated, so it’s very reasonable that you might run into questions. That’s expected, and is part of the process, so please get in touch when you run into things you’d like to ask about or discuss!