PATSTAT / Fleming Algorithm TODO File

Mark Huberty

October 18, 2012

This file was started after the 17 October 2012 call with Georg and Amma. It outlines the steps needed to implement the Fleming Smallheiser / Torvik (FST from here on out) disambiguation algorithm on the PATSTAT database.

1 Prepare

1.1 DONE *2012-10-17 Wed* ID PATSTAT fields for use in the FST multi-dimensional similarity profile

Resolved: we will match on the following fields:

- Name
- Company legal identifier
- Country
- Top N IPC codes
- Top N Co-Authors

We will not match on address; the address fields are likely too complicated.

1.2 TODO Check whether the Yahoo GeoDict service can handle the number of addresses we would need to formalize

UPDATE: Yahoo's Placemaker service probably limits requests to 50,000 every 24 hours. That's about one every 1.8 seconds. I count about 4.6

million unique addresses (of a total of 12 million non-blank addresses) in the PATSTAT `person` tables.

This means that disambiguation would take about 92 days.

There's an open-source academic solution provided by the Unlock project at the University of Edinburgh. I've emailed their primary contact to see if she has advice on how to handle this.

Finally, there's a fully open solution called Geodict that uses it's own MySQL database of 2 million placenames to extract data. But I suspect it's good at cities/countries, and less good at addresses.

The final option would be to use theData Science Toolkit's geocoder service, which I could run locally on my own server. Then we can just query away with no limits.

1.3 TODO Determine the right N for the IPC code field

Amma will check the summary statistics (min/mean/median/max) on how many IPC codes the average patent has

1.4 TODO Determine the right N for the co-authors field

Amma will check the summary statistics (min/mean/median/max) on how many co-authors or co-assignees the average patent has

1.5 TODO Generate text files for use in the disambiguation process

We need to build files with one row per author-patent instance for use in the disambiguation process. This will require querying the data of interest from PATSTAT, formatting it, and writing it out.

We will do the following:

1. Query by country
2. Write out by country (format: tab-separated file; filenames to have country code in them)
3. Clean the names/addresses in each file using the `psClean` logic
4. Contatenate the files into a single text file
5. Load them to the sqlite database for further use

Before doing this, we will run tests on only countries beginning with the letter "A".

2 Configure

2.1 Adapt the primary config files

Both `sp.desc` and `data.desc` need to be updated per the documentation for the format of our data and the details of our multidimensional similarity profile.

This should wait until we are 100% clear about the structure of the data file itself, and the specific comparisons between records that we will make.

2.2 TODO Generate "training" and "testing" datasets

This should use exact matches of relatively rare names; and mismatches of relatively rare names. We probably want to use names whose counts are in some range, like the 5-10th percentile, rather than the rarest; otherwise we could get weird noise from bad data formatting.

Exact matches should be defined as an exact name match plus an exact country match.

2.3 TODO Define blocks

The first disambiguation round should use blocks based on country and exact match of one word in the first name

3 Test

4 Run

5 Evaluate