Instructions

This is a guide on how to replicate this exercise for data-sets for other districts and courts.

Data Source

The data is publicly on from the e-courts district website.

1. Navigate to Case Status
2. Select the State, District, Court Complex and Court Establishment. The first three are relevant for all districts. Establishment is applicable only in certain court complexes
3. On the horizontal toolbar, the default selection is "Party Name". Click on "Act" to filter by Act
4. Within the Act Type dropdown, select "Negotiable Instruments Act"
5. In the Under Section, type 138
6. Choose Pending / Disposed. The activity will have to be repeated for both.
7. Enter captcha, and click go

It should display a table. Wait while the data loads. It can take a while depending on the district.

Extracting the Data

1. Once the data has been loaded, use the relevant browser option (typically a right-click menu) to "Inspect".
2. This will open up the HTML data. Look for
3. Copy the entire HTML block. You should be able to right-click and use one of "Copy HTML" or "Copy Outer HTML" or "Copy Element".
4. If the case status selected in Step 6 of the Data Source steps above is "Pending", save it into the 'district'-pending-cases-all.txt, if it is "Disposed", save it into the 'district'-disposed-cases-all.txt
5. Paste all HTML blocks one below the other in the text file

It may be useful to use a tool like Notepad++ or SublimeText or any other text editor that makes it easy to read HTML

Converting the data into csv

Run court-cases-cleanup.ipynb

Remember to replace the name of the district in the code so it picks up the right txt file.

Requirements

You need to have the following libraries installed

- Pandas
- BeautifulSoup
- lxml

Interim Output

The output will be 2 files. In the case of Mumbai, these files are 'pending-cases-mumbai.csv' and 'disposed-cases-mumbai.csv'

Running the Litigant Classifier

Run litigant-classifier.ipynb

Requirements

You will need the following libraries

- Pandas
- NLTK

Code Structure

Run the jupyter blocks one at a time

- The first code block identifies all the words in the names of all the litigants
- The second code block exports the words in the names that are also english words in the wordnet corpus
- The third block is the classification logic based on keywords into 3 types - Financial Firms, Non Financial Firms and Individuals

Classification Logic

Litigants are classified through a process of looking for keywords in the name:

1. Financial Firms typically have the terms bank, finance, invest, loan, and related keywords and variations.
2. Non Financial Firms have terms like ltd, pvt, corporation.
3. Non Financial Firms may contain common nouns from the English Language.
4. Litigants with the term proprietor in the name were categorised as individuals.
5. Those that did not fit these criteria were categorised as individuals.

Manual Step

The classification logic in Step 3 above depends on identifying common nouns. The corpus has a mix of proper and common nouns so a manual cleaning process is required.

- Before executing the third block, you will need to make a copy of the file englishterms.csv and rename it englishterms-clean.csv
- Within the column Ignore, add "Yes" against any proper nouns or other terms you don't want to use to classify a litigant as a Non Financial Firm (lawyer, representative, alias etc.)
- For our exercise, we manually categorized words that constitute 95% of the volume. The columns "Count" and "Coverage" in the englishterms-clean.csv are used to track when to stop. Whether to do 80%, 95% or 99% is left to the user, based on how much one wants to reduce the error rate

Output

Once you run Block 1, Block 2, do the Manual Step, run Block 3.

The output will be a csv e.g. 'mumbai-all-cases-classified.csv'.

The classifier obfuscates the names of the litigants. Of course, while you run the data on your system, feel free to remove the drop_columns code if you want to retain the original names.

Scope for Improvement

The methodology above suffers from some known issues and a small rate of misclassification. Some names like Banku and Chitra containing the terms Bank and Chit could be classified incorrectly. We do not account for firms that have common nouns in their name from languages other than English. In many cases, an individual proprietorship may have the term company or finance in their name. The methodology does not take into account spelling errors.

We hope to continue to improve this methodology. Please write to us with suggestions, or contribute to improving this project.