Company Type Analysis

By Angela Gonzalez-Curci for WSDOT as part of University of Washington/Open Data Literacy Internship

# Goals:

Find the types of organizations which are requesting the most records.

Assess the feasibility of automating the analysis of company type through the use of a decision tree classifier.

# Analysis Strategy:

Take a sample of full dataset and code entries in the Company Name field. First, split sample into two subsets: Public and Private, and base subsequent subset on patterns found in keywords and syntax in company names. Record keywords and syntax structures used to differentiate different class types.

After classification, frequency of classes and subclasses were taken with Excel Countif function. The percentage of each class and subclass was then calculated by dividing class count by the total sample count.

# Data Definitions:

## Company:

Any organization incorporated or not, in which multiple individuals work towards a common goal.

## Public

Organizations not structured to make a profit. Includes individual citizens who are not incorporated and therefore look more like public groups than private groups.

### Citizen

Individuals not explicitly associated with a company.

### Government

Ruling, regulatory, and administrative public organizations at all levels of jurisdiction.

Expectation: Public Utilities look more like Government organizations than private companies even though they are natural monopolies motivated by profit.

### Academic

Educational institutions at any degree level and of any type. For example, local elementary schools, Commercial Driving schools, and universities are all considered academic.

### Nonprofit

All other nonprofit organizations not governmental or academic in purpose

## Private

Organizations motivated by profit.

### Business

All profit organizations not in law or communications.

### Law

For-profit organizations in the legal field. Non-profit legal organizations, such as, Clinics, Prosecutor Offices, Attorney General Offices are not classified as law.

### Media

For-profit and non-profit organizations with the communications field, such as blogs, websites, TV and radio stations, and newspapers. Public broadcasting are considered media despite being non-profit organizations, since all other Media types are for profit.

## Unknown

Organizations of whose classification could not be discerned.

# Method

Data analysis was done in an Excel spreadsheet titled Company Type Analysis. A sample of 3993 entries from the PDR dataset.

After classification, frequency of classes and subclasses were taken with Excel Countif function. The percentage of each class and subclass was then calculated by dividing class count by the total sample count.

Since the sample is such a large portion of the whole population of the dataset, the margin of error needed to be adjusted with the Finite Population Corrector:

Finite Population Corrector =

N = Population Size

n = Sample Size

With this adjustment the Margin of Error was calculated for each class and subclass:

Margin of Error = FPC \* Z-Value \*

FPC = Finite Population Corrector

Z-Value for 99% confidence level = 2.58

P = Percentage of Sample

n= Sample Size

# Analysis:

## 

# Results:

## Frequency Counts:

## 

|  |  |  |  |
| --- | --- | --- | --- |
| Class Type | Count | % of Sample | Margin of Error |
| Business (B) | 1698 | 42.5% | 0.028% |
| Citizen ("C") | 919 | 23.0% | 0.023% |
| Government(G) | 844 | 21.1% | 0.023% |
| Law(L) | 273 | 6.8% | 0.014% |
| Media (M) | 158 | 4.0% | 0.011% |
| Academic (A) | 41 | 1.0% | 0.006% |
| Nonprofit(N) | 41 | 1.0% | 0.006% |
| Unknown | 19 | 0.5% | 0.001% |
| Total | **3993** | **100.0%** |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Class Type | Count | % of Sample | Margin of Error |
| Public | 1847 | 46.26% | 0.03% |
| Private | 2132 | 53.39% | 0.03% |
| Unknown | 14 | 0.351% | 0.003% |
| Total | **3993** | **100.00%** |  |

## Proposed Decision Tree:

### Proposed Classes with Suggested Keywords and Syntax

|  |  |  |
| --- | --- | --- |
| Class Type | Keywords | Syntax |
| Media | FM, AM, Press, News, Reporter, Media, Newspaper, Herald, Journal, Bulletin, Observer, Tribune, Chronicle, Messenger, Sun, Television, TV, Radio, Advocate | Entry starts with “The” |
| Firm | PA: P.A, P.A.  PLLC: P.L.L.C, PLLC., PLL.C  Company: Corp, Corp., C.O, CO.  PC: P.C, P.C.  Inc: Inc.  PS: P.S, P.S.  SpA  Ltd: LTD  Group, Associates, Firm | Entry Ends with Keywords  Entry contains a last name |
| Non-Profit | Church, HOA, Taskforce, Foundation, Club, Volunteer, Local, Federation, Union, Society, Fund, Alliance, Homeowners |  |
| Public | Department, Dept. | Contains name of City, County, State or Country in entry |
| Citizen |  | No value in Company Type Field, but a name in the First Name Field |

### Proposed Firm Subclasses with Suggested Keywords and Syntax

|  |  |  |
| --- | --- | --- |
| Subclass | Keywords | Syntax |
| Law Firm | Law Group, Law Office, Law, Legal, Attorney, Attorneys, Counselor | Three or more last names separated by comas, with last name separated by an ampersand. |
| Business | Timber, Forestry, Freight, Broadband, Lumber, Solar, Reconstruction, Geotechnical, Geomatics, Survey, Surveying, Materials, Plan, Planners, Manufacture, Insurance, Design, Commercial, Construction, Solutions, Architects, Mapping, Consultants, Industries, Distribution, Cable, Transparent, Enterprises, Products, Real Estate, Logging, Asphalt, Concrete, Contactor, Gravel, Trucking, Systems, Strategies, Hauling, Crane, Rigging, Truck, Analysis, Logistics, Management, Polymers, Claim |  |

### Proposed Public Subclasses with Suggested Keywords and Syntax

|  |  |  |
| --- | --- | --- |
| Subclass | Keywords | Syntax |
| Academic | Institute, University, College, School, Technology, Academy, Faculty, Laboratory, Student |  |
| Government | Public Works, PUD, Council, Wastewater, Police, Assessor, Sheriff, Commerce, Penitentiary, Correction, Appraiser, Earth and Space, Health, Housing, Licensing, Natural Resources, Interior, Economic Development, Permitting, Floodplain, Traffic Engineering, Utilities, GIS, Transit, Power, Agriculture, Wildlife |  |