WSDOT DISCLOSURE REQUESTS

Week Four: Cleaning Disclosure Dataset

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

# Purpose

The purpose of this document is to document and describe the work I have done in cleaning and the dataset, show some of basic statistical results gained through this data cleaning, add new questions based on this results of this cleaning, and document what the next cleaning and analysis steps will be.

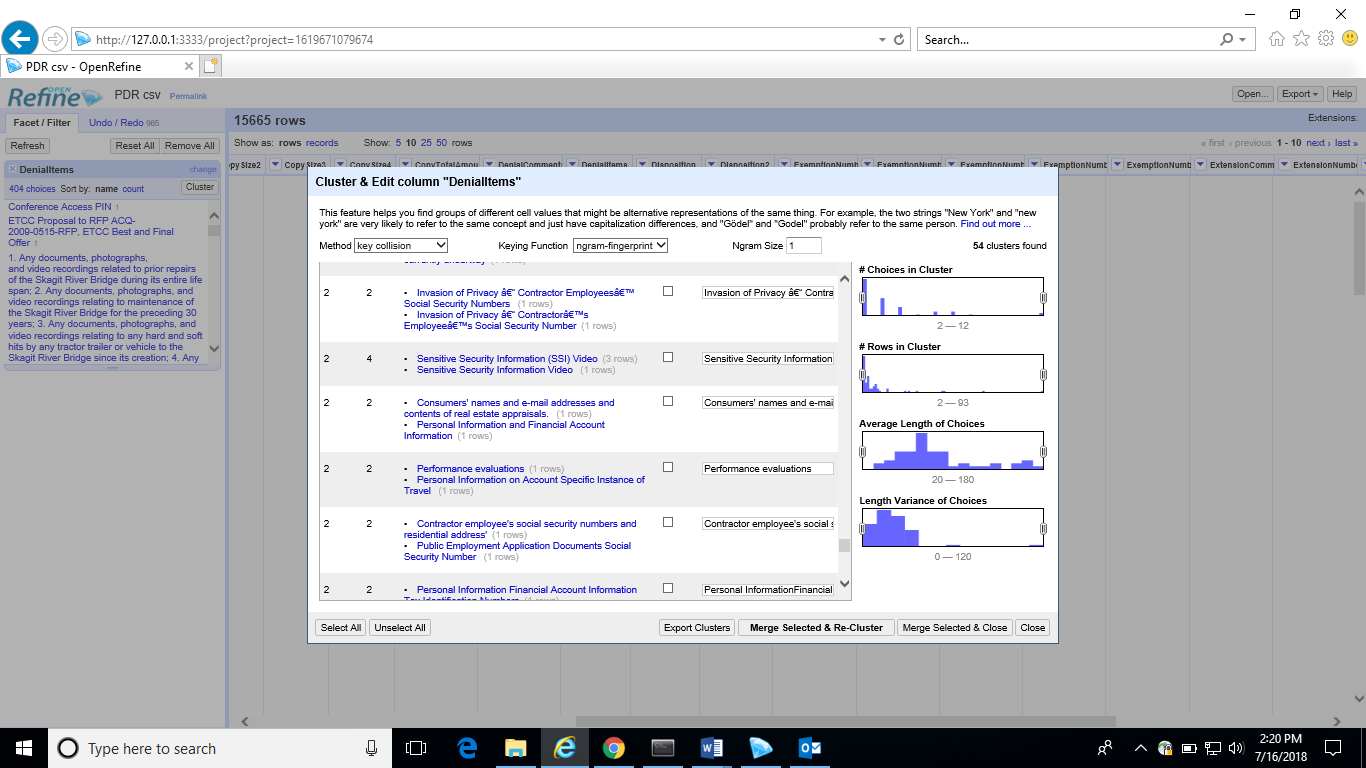
# Analytics Workflow

Progress as of Week Four

# Data Cleaning

## Method

This week I moved from data cleaning in python with Jupyter Notebook to record my steps, see attached Jupyter Notebook page, to using Open Refine. Open Refine makes it easier to clean data values: replacing incorrectly or unstandardized terms with the correct values, and find missing values based on patterns in other columns, by creating “clusters” or sets of values that have a high probability of meaning the same thing based on different algorithmic processes.



Example of clusters in the “DenialItems” column found using the Key-collision method and an N-gram size of one. An N-gram is set of words highly probable to be found concurrently. An N-gram of one is looking for the probability one word will be found concurrently with another word.

## Rules for Clustering in Open Refine

Sometimes it was very obvious what the correct value of an entry was. For example, white space, misspellings, incorrect capitalization. However, a majority of the changes made were not clear errors, but rather an issue of preferred terms. Terms deemed too narrow, such as acronyms for Government Agencies or too narrow, such as nine digit zip codes, were standardized to a preferred term on a set of rules I created based on patterns I was seeing in the column and my understanding of the types of analysis I want to run on the data.

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Rule for Clustering | Example | Exception |
| Assigned | Full Name | Patti Thomas = Patricia Thomas |  |
| Company | The requesting organization doing business as another organization | Lancaster Engineering dba the City of Tacoma = Lancaster Engineering. |  |
| Company | Formal Business Name | PACE = PACE Engineering, Inc. | Media organizations with multiple affiliates within the State of Washington  CBS News = KIRO News or KIMA News |
| Company | Joint Request are treated as one unique organization | Azure|Green Consultants, LLC |  |
| Company | National, state, county and  local agencies are preferred terms. Sub-agencies are removed. Municipalities are accepted if no agency is defined | City of Sequim – Accepted term  City of Sequim Public Works – preferred term  City of Sequim Public Works Sanitation Dept = City of Sequim Public Works | Oregon Department of Transportation- Motor Vehicle Transportation Division and WSP Collision Reports are frequent requestors. Important information would be ignored if merged |
| Company | University Name is the preferred term, department is a narrow term | UW Dept. of Civil Engineering = University of Washington |  |
| ExemptionNumber (1-5) | Blue Book Citation as Standard for Statutes and Court Rules | Court Rules :  WA Superior Court CR 26(4)(b)  Statute:  RCW 42.56.270(2) | Superior Court case rulings are not available online and I can’t easily find the proper citation. Court Name, Case  Number creates a unique and easily findable value. |
| FullNameFML | Company Names and Email are not considered Names and are considered Null Values |  |  |
| FullNameFML | First M. Last is preferred, if known | Angela Gonzalez = Angela G. Gonzalez |  |
| FullNameFML | One request filed by two or more people is considered one unique request |  |  |
| FullNameFML | First Name only are considered Null Values unless uniquely identified through a specific company |  |  |
| FullNameFML | Professor Titles are not a part of a name | Dr. Angela Gonzalez = Angela Gonzalez |  |
| State | If not United State, full country name is acceptable | CANADA  ITALY  AUSTRIA |  |
| Zip | Five digit US and Six digit with space Canadian zip code | 99403  V4G 1A5 |  |

Note: this is not uniform thought-out the dataset, as clustering is probabilistic and did not pick up all the terms which are not perfectly unique. For example, dept. and department were not picked up as equivalent terms. During Week 5 I will be using basic text-mining techniques to create a list of similarly used terms to use in my analysis.

# Frequency Results

## Top Ten Assigned Coordinators

|  |  |  |  |
| --- | --- | --- | --- |
| Coordinator | Requests | Closed Requests | Group |
| Julie Brown | 2968 | 2962 | Transit Data Office |
| Eric Van Hooser | 1223 | 1213 | [North West Region](javascript:%7b%7d) |
| Bryan Ray | 1098 | 1096 | [Engineering](javascript:%7b%7d) |
| Robert Lee | 1075 | 1035 | [Headquarters](javascript:%7b%7d) |
| Rob Fossett | 987 | 981 | Engineering |
| Patricia Thomas | 868 | 862 | [South Central Region](javascript:%7b%7d) |
| Tanya Joblonski | 727 | 713 | Transit Data Office |
| April Bickar-Pierce | 672 | 636 | [Headquarters](javascript:%7b%7d) |
| Trevor Alexander | 623 | 603 | [Headquarters](javascript:%7b%7d) |
| Celeste Dimichina | 579 | 579 | [South West Region](javascript:%7b%7d) |

## 

## Top Ten Requesting Companies

|  |  |  |
| --- | --- | --- |
| Company | Number of Requests | Type of Requestor |
| Oregon Department of Transportation | 621 | State Government |
| Transpo Group, Inc. | 275 | Transportation and Engineering Firm |
| WSP Collision Records | 273 | State Government |
| Gibson Traffic Consultants, Inc. | 267 | Transportation and Engineering Firm |
| Gray & Osborne Consulting Engineers, Inc. | 151 | Engineering Firm |
| H. Lee & Associates, PLLC | 150 | Engineering Firm |
| [Transportation Engineering NorthWest](javascript:%7b%7d) | 124 | Engineering Firm |
| KIRO News | 113 | Media |
| The Seattle Times | 99 | Media |
| Kittelson & Associates, Inc. | 95 | Transportation and Engineering Firm |

## Top Ten Individual Requestors

|  |  |  |
| --- | --- | --- |
| Full Name | Number of Requests | Company |
| Denise S. Rothermund | 193 | ODT- Motor Carrier Transportation |
| Patricia Gibbs | 180 | WSP Collision Records |
| Brian K. Ulmer | 132 | ODT- Motor Carrier Transportation |
| Grant Stonex | 100 | H. Lee Associates, PLLC |
| Jim Morelli | 99 | WSP Collision Records |
| Julian Tarver | 81 | Clallam Bay Correction Center |
| Matthew Palmer | 75 | Gibson Traffic Consultants, Inc. |
| Graham Johnson | 68 | KIRO-News |
| Sarrunas Farler | 63 | Gibson Traffic Consultants, Inc. |
| Spencer Haynic | 63 | Transportation Engineering NorthWest |

# Conclusions

## New Questions

From looking at the results of these frequencies that there are companies and specific requestors within