Linking EEOC Case Investigation Records by Employer Name using Text Analytics

Ada Harris

Government Advances in Statistical Programming (GASP) Presentation

September 23, 2019

Overview

- Background
- Introduction
- Integrated Mission System (IMS)
- Problem
- Approach
- Data and Methods
- Results
- Summary
- Next Steps

Background

EEOC Investigates complaints of discrimination in the workplace

- Race
- Color
- Religion
- Sex
- National Origin
- Age
- Disability
- Retaliation

Introduction

Why is this important to the EEOC?

- Expand the use of administrative data to understand trend across the country
- Efficiently leverage data into insight to support the mission of the agency

Integrated Mission System (IMS)

- EEOC's internal information management system and data repository for case investigations – "charge data"
- Database repository for intake, investigation, settlement and closure of charges of employment discrimination
- Five main tables
  - Charge
  - Allegation
  - Respondent Name (Employer)
  - Action
  - Charging Party Name (Employee)

Problem

There is no key or ID number to link different charges against the same employer or corporation. Respondent name must be used to link charges.

A Respondent Name must be recorded for a formal charge. Variations of the employer's name, business name changes over time, and errors in manual entry make it difficult to link cases.

The purpose is to use text analysis to identify unique respondent names and determine whether they should link to other respondents.

Challenge Example

- EEOCRespondent#1 vs. EEOC-Respondent #1
- EEOC Respondent Store #1784
- Business Name d/b/a EEOC Respondent
- Corporate Name vs. Franchise Name

“Private Company” Dental Insurance

“Private Company” Tools

“Private Company” Faucet

Approach

- Develop search query for Respondent Name to model the degree of similarity between the Respondent Name and all Respondents within IMS.

- How close are two pieces of text in lexical similarity and semantic similarity?

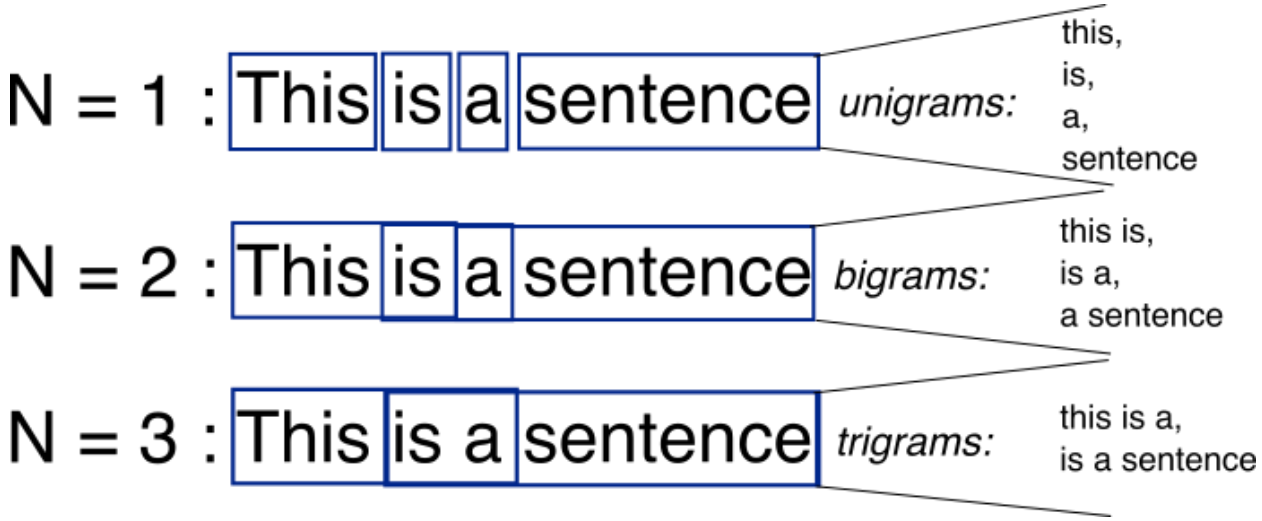
Represent Respondents as vectors of features and compare all Respondents to a single respondent of interest by measuring the distance between the features.

Data and Methods

- Data from IMS from 2008-2018 from the Respondent Table were used to for text analytics (1.4 million observations)
- Data cleansing
  - Remove punctuation
  - All lowercase
  - Trim White Space
  - Remove stop words (such as “the”, “a”, “an” and “in”)
- Deterministic deduplication by unique string of a Respondent Name(681,625 observations)

Data and Methods

- N- Grams – sequence of N contiguous items (in this case character)
  - Example : EEOC Respondent 3-Gram
  - 'EEO', 'EOC', 'OC ', 'C R', ' RE', 'RES', 'ESP', 'SPO', 'PON', 'OND', 'NDE', 'DEN', 'ENT'



Source: http://recognize-speech.com/language-model/n-gram-model/comparison

Data and Methods

Term Frequency – Inverse Document Frequency (TF-IDF) – measures how important a word is to document in a corpus

Term Frequency – how frequently a word occurs in a document

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

IDF – measures the importance of the term

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

Example

Consider a document containing 100 words where the Respondent Name appears 5 times. The TF for the Respondent Name is (5/100) = 0.05. Assume we have 10 million documents and the Respondent Name appears in 1,000 of these. Then, the IDF is log(10,000,000/1000) = 4. The TF_IDF is the product of these quantitates 0.05* 4 = 0.20

Data and Methods

- Cosine Similarity – to calculate the similarity between TF-IDF values
  - Measure of orientation and not magnitude
  - Normalized dot product

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

RESULTS EXAMPLE

Query Term	Document Term	Variations	Similarity
“Private Company” Tool	“Private Company # 580” Tool	Name entry variation	0.9948
	“Privte Comany” Tool	Spelling Variation	0.9948
	“Company Private” Tool	Name entry variation	0.9937
	“Private&Company” Tool	Name entry variation	0.9937
“Private Company” Faucet	“The Private Company” Faucet	Name entry variation	0.9854
	“Private Company INC” Faucet	Name entry variation	0.9742
	“d/b/a Private Company” Faucet	Name entry variation	0.9548
	“Private and Company” Faucet	Name entry variation	0.9398

RESULTS

EEOC Respondent #1

Manual String Search: EEOC Respondent #1 classified 7,783 of 681,625 observations

Entity Linkage Approach: EEOC Respondent #1 appears 16,088 of 681,625 observations

EEOC Respondent #2

Manual String Search: EEOC Respondent #2 classified 4,856 of 681,625 observations

Entity Linkage Approach: EEOC Respondent #2 classified 10,859 of 681,625 observations

Summary

- Multiple entries corresponding to the same entity is a problem often occurring in databases and can lead to loss of information.

- Growing data sources create interesting challenges and opportunities for linking data

- Record linkage helps to find information about Respondents and extract actionable data from the IMS database to support leveraging internal data to understand Respondent charge frequency

Next Steps

- Decrease time to results
  - Parallel processing
- Integrate business name databases as golden standard for correct name
- Analyze cut-off scores for classification
- Manually classify output to for future automation