Classification of death records using topic modeling and supervised classifiers

Maya Bhat-Gregerson
August 2020

Overview

- Cause of death information is required on all death records and is used for public health research and disease prevention efforts.
- Often, cause of death statements are poorly worded making it difficult to select a single, initiating disease leading to death.
- This project examines the potential for topic modeling and supervised machine learning classifiers to use all available information on cause of death statements to assign these problematic records to specific category of death.

The typical flow of mortality data

Death record creation

- In WA, cause of death completed by medical provider, medical examiner, or coroner.
- Ideally, a causal sequence of conditions leading directly to death

Death record processing

- Medical terms in cause of death statement are assigned ICD-10 codes
- One condition in the sequence is selected as underlying cause of death

Epidemiological analysis

- Analyses of ICD-10 codes provide information on leading causes of death, subpopulations affected disproportionately, trends etc.

The problem

- Annually, there are over 57,000 deaths registered in WA
- Medical providers don’t receive much training in writing good cause of death statements.
- Each year, between 6 and 7 percent of death records have poorly worded cause of death statements resulting in ICD-10 codes that are not useful for analysis.

Potential solutions

- Provide training to cause of death certifiers
  - Available, but medical providers don’t have motivation or time
- Natural language processing (topic modeling, machine learning classifiers using cause of death statements)
  - This could be either medical terms (words) or ICD-10 codes

Data cleaning and preparation

- Death records for 2016-2019 for persons who died in WA
- Check for missing and out of range values
- Standardize text fields:
    - Remove punctuation, white spaces
    - Tokenize text for natural language processing (NLP)
- Label records with garbage code categories

Exploratory analysis

- Greater proportions of records with garbage codes for underlying cause among:
  - Older decedents (80 yrs +) vs. 0-19 year old decedents
  - American Indian/Native Alaskan vs. white non-Hispanic or Asian non-Hispanic
- Physician assistants most likely to submit death records with poorly written cause of death

Exploratory analysis (cont'd)

- Between 6.4% and 7% of death records annually had garbage codes (roughly 7,100)
- Most common garbage codes were for underlying causes that fell into one of three groups:
    - Septicemia
    - Unspecified heart disease
    - Unspecified cancer

Topic Modeling - Latent Dirichlet Allocation

- Unsupervised learning algorithm used with corpus of ICD-10 code tokens
- Specified 6 topics
- Models with abbreviated unigram ICD-10 codes (letter + 2 digits) showed the most distinct topic areas (in terms of code frequency)
- Higher coherence values did not always mean that the topics modeled made sense in term of the groups of most common codes

Latent Dirichlet Allocation (cont'd)

- Adding more records to corpus may improve clustering of ICD-10 codes into distinct topics
- Future steps:
  - Mallet LDA - better algorithm for short text
  - Add records with valid ICD-10 underlying cause codes to provide more information to algorithm

Supervised learning: classifiers

- Multinomial Naive Bayes and linear Support Vector Machine
- Classes included 10 leading causes of death in WA
- Training and testing data consisted of records with **valid** underlying cause codes because we know which cause of death to use as labels.
- Can’t use garbage coded records for training- true underlying cause labels not available without medical expertise.

Supervised learning: classifiers (cont'd)

- Both MNB and linear SVC yielded mediocre accuracy when using multiple classes in algorithm
- One vs. all binary classifiers for each cause of death category had much greater accuracy but not such great recall and F1 scores for the positive (with disease) class.
- Imbalanced class problem present that needs to be addressed in the future

Next steps

- LDA model
  - Apply Mallet LDA algorithm - more suited to short text topic modeling
  - Add death records with valid codes to corpus
- Linear SVM
  - Recreate training and testing data with balanced classes and rerun binary classifiers for each disease

Next steps

- Linear SVM
  - Recreate training and testing data with balanced classes and rerun binary classifiers for each disease
  - Create narrower classes for some causes of death (cancer, heart disease)