Course Project: ICD code and Mortality

August 10, 2019

1 Background

Every year, centers for disease control and prevention (CDC) provides detailed statistics of deaths and their underlying causes in the United States. The CDC mortality data is used by various industries like medical, health and insurance to provide better services. It provides the basis of numerous researches and is widely cited in public papers.

Because of your outstanding knowledge, you are hired as a data scientist by a prestigeous insurance company. The VP of your department wants to launch a life insurance product but would like to do an analysis on the cause of death in the US population first. The following questions needs to be answered:

2 Problems

- What are the major causes of death in the US
- For different causes of death, how does the death distribution look like against age (For example, histogram of 5 year age band)?
- For each age band (5-year band), what are the top 3 causes of death? Do they differ?
- Are the causes of death chaning over time? Are there any significant increasing or decreasing trends in some causes?
- You have medical transcripts of 5000 patients "**medicaltranscriptions.csv**". You want to determine if any patients have medical conditions that are associated with ICD codes of major causes of death.
  - Design a similarity measure metric
  - Calculate the similarity measure between ICD code descritption and medical transcirpts
  - Assign ICD code to a medical transcripts only if the similarity score is above certain threshold.

3 Data

You are provided with the CDC mortality data between 2005 and 2015, under the mortality directory i.e. "**data/mortality/2005_data.csv**", "**data/mortality/2006_data.csv**" ... "**data/mortality/2015_data.csv**". Every row in the files represents a death file in CDC. Each file contains 77 columns and the fields are selectively described as below:

- current_data_year: the year of the statistics
- detail_age: death age
- icd_code_10th_revision: ICD10 code, a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. A detailed descriptions of the code can be found in "**data/ICD10/allvalid2011 (detailed titles headings).txt**"

4 Project Requirements

Please use the data provided and do the analysis that are needed to answer the questions listed above. Put together a report and present it to the VP of your department. Specifically,

- Form a team of 4-5 people and work together to accomplish the tasks.
- Prepare a report that describe the project's background, your methods, results (including graphs) and conclusions.
- Put together a powerpoint and presents to the results.
- Submit the report together with the codes.