Disease Text Classification

XIE Tian, WANG Chang, WANG Sidi

12/19/2019

1. Overview

Every day, work-related injury records are generated. **In order to alleviate the human effort expended with coding such records**, the Centers for Disease Control and Prevention (CDC) National Institute for Occupational Safety and Health (NIOSH), in close partnership with the Laboratory for Innovation Science at Harvard (LISH), is interested in **improving their NLP/ML model to automatically read injury records and classify them** according to the Occupational Injury and Illness Classification System (OIICS). Our project is inspired by this initiative.

This project represents a **text classification** problem that is expected to be solved using efficient **big dataset** handling techniques and various **classificiation algorithms**. Through exploration, we hope to achieve better accuracy and higher efficiency in injury records classification.

2. Data Inspection

A random sample of 153,956 records with the outcome event column included. The data have 4 column (text, age, sex and a response label): We have 48 unique OIICS response label in total.

```
Train <- read.csv("CDC_Text_ClassificationChallenge_TrainData.csv")
head(Train, 3)
```

```
##                                                                                                                text
## 1                    57YOM WITH CONTUSION TO FACE AFTER STRIKING IT WITH A POST POUNDER WHILE SETTING A FENCE POST
## 2                                                  A 45YOM FELL ON ARM WHILE WORKING HAD SLIPPED ON WATER FX WRIST
## 3 58YOM WITH CERVICAL STRAIN  BACK PAIN S P RESTRAINED TAXI DRIVER IN LOW SPEED REAR END MVC NO LOC NO AB DEPLOYED
##   sex age event
## 1   1  57    62
## 2   1  45    42
## 3   1  58    26
```

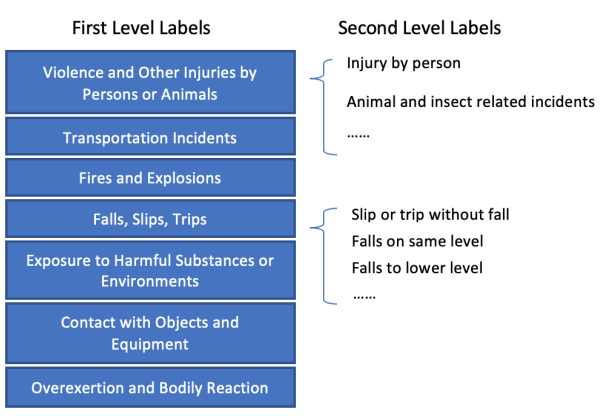
```
dim(Train)[1]
```

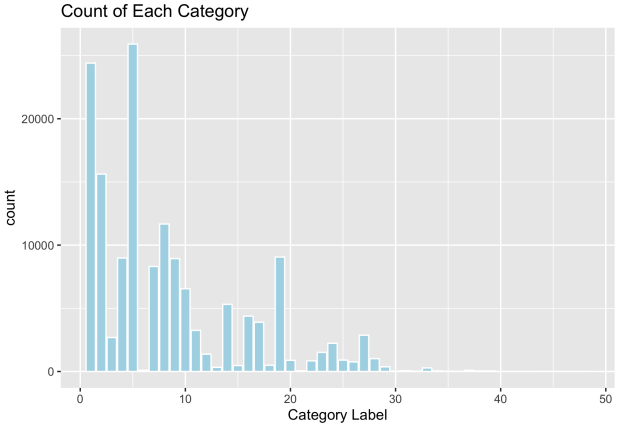
```
## [1] 153956
```

```
summary(Train$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   27.00   37.00   38.29   48.00   96.00
```

The response variable is hierarchical (for example, if one label is 32, then it belongs to the third class "Fires and Explosions", and the second type). Figure on the left is showing the 7 first level labels and a few second level labels (48 in total). Although there are 48 possible outcomes, the frequencies is extremely unbalanced, as shown in the figure on the right. This needs to be taken into account while training different models.





Data Processing and Prelimiary Analysis [1] Tian

Approach [1] Tian, Chang

Result [1] Tian, Chang, Sidi

Conclusion and Future [1/3] Sidi