Applying Text Mining in Industry and Occupation Narratives Conversion

Hantao Wang and Thomas Salter

Tennessee Occupational Health Surveillance Program

BACKGROUND

Industry and Occupation (I/O) variables are used to identify risky work groups in Occupational Health (OH) surveillance and research. Most data sources that OH epidemiologists deal with store I/O variables in text narratives. NIOCCS is a machine learning based software developed by NIOSH to convert narratives to existing I/O classification system. In practice, NIOCCS does not perform as expected due to following reasons:

- Data collectors lacking professional education
- Same phrases having multiple meanings
- Lacking training data

In this presentation, a new Natural Language Processing based method is purposed to solve those difficulties in I/O narratives conversion with performance comparison to NIOCCS.

METHODS

Word2Vector (W2V) model is commonly used in text mining. W2V is an embedding model where every word is represented in a high dimension word vector (WV). Directions of each WV can measure the similarity of meanings. By adding or subtract WVs, the calculated WV can represent aggregate meanings of a group of words, which is usually a sentences or paragraph. Given that, W2V model is able to tell that King – Man + Woman = Queen. In this study, a list of keywords was first mined from the manuals of NAICS and SOC, which was used as estimated meaning of each I/O. The top possible I/O were then ranked by weighting the term-frequency of a given entry and comparing its similarity to list of key words from first step. GoogleNews-vectors was adopted as pre-trained W2V model in this project. In addition, since the vectors can group phrases with similar meanings, such process does not necessarily require a training set.

RESULTS

On a sample of 277 industry and occupation pairs extracted from death file, the proposed method can convert all 277 pairs (100%) with 83.39% accuracy (231/277). The NIOCCS was able to convert 145 records (42.6%) with 91.2% accuracy (118/145). The overall correctness of this method is 83.39% while 42.6% for NIOCCS given its low conversion rate. The two methods have 116 records with agreement. The purposed method is significantly faster than NIOCCS. On another sample run of more than 15,000 records, the purposed method finished in half hour on a local 16GB-RAM PC, while such size is too large for NIOCCS to handle.

Figure 1

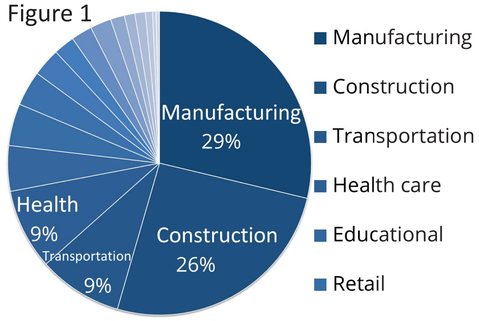
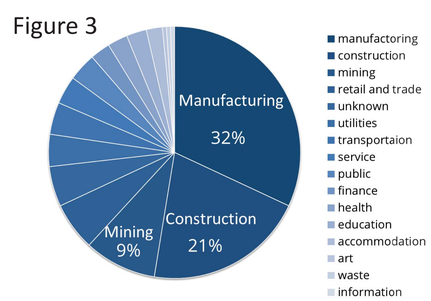


Figure 2

Site Type	%
PROSTATE GLAND	24.6%
LUNG & BRONCHUS	19.8%
BLOOD, BONE MARROW, & HEMATOPIETIC SYS	5.92%
URINARY BLADDER	5.21%
LARGE INTESTINE	5.15%

Figure 3



CONCLUSION

The result on this sample testing shows that this new method is more efficient in converting industry and occupation narratives. The results may be improved further with a labeled training set. Tennessee Cancer Registry and death statistical file both have I/O variables in either classified coding system or narratives. Using the purposed method, we were able to increase understanding of targeted group more precisely. **Figure 1** is the industry distribution of mesothelioma cases reported to TN Cancer Registry since 1983. **Figure 2** is the most common cancer sites of Firefighters in TN. **Figure 3** is the distribution of pneumoconiosis mortality by industry in from 2012 to 2015.

Scan the button for a link to the demo website. If you are interested in getting more details regarding this poster, please contact Hantao.Wang@tn.gov.