Post Training Survey Data Analysis

Ken Vu

March 25, 2022

Introduction

Here's an example of analysis of survey data for a training held at the San Jose Conservation Corps. We'll be using it to generate a word cloud to study the keywords used in the survey responses to the question, "What skills did you hope to learn/improve on through this training session?"

Code for Survey Analysis

First, let's load all of the relevant libraries.

```
library(pacman)
```

```
## Warning: package 'pacman' was built under R version 4.1.2
```

```
p_load(dplyr, tidytext, tidyverse, stringi, wordcloud, igraph, ggraph)
```

Let's load the survey data.

```
df <- read.csv("training_response_sample.csv")
head(df)
```

```
##      id
## 1 hI7d5
## 2 AkW3n
## 3 h3Q9s
## 4 BKFrF
## 5 lpzep
## 6 vroUV
##                                                                                skills_gain
## 1                                                        Learn about environmental justice
## 2
## 3                                                                    Awareness of pollution
## 4         My knowledge about can i help my community with contamination and how to prevent it
## 5        I learned about different neighborhoods and how economically it impacts some areas.
## 6 More awareness of enviro injustices in San Jose area, and the tools to combat this at work
```

Looking at Individual Tokens

Let's see what the attendees think of the training in terms of commonly used words related to the skills they hope to gain and/or work on.

Data Preparation and Processing

Break down the words into individual tokens

```r
# Tokenize the words
data_words<- df %>% unnest_tokens(word, skills_gain) %>%
  anti_join(stop_words) %>% count(word, sort = T)
```

```
## Joining, by = "word"
```

```r
# Removing words related to the title of the training or the city of
# "San Jose" (a bigram and name that makes no sense when separated into tokens)
data_words_small <- data_words[-c(1,2, 25,31),]
data_words_small
```

```
##             word n
## 3      awareness 2
## 4      knowledge 2
## 5       learning 2
## 6      pollution 2
## 7        ability 1
## 8          check 1
## 9          clean 1
## 10        combat 1
## 11     community 1
## 12 contamination 1
## 13  economically 1
## 14   electronics 1
## 15        enviro 1
## 16   environment 1
## 17     expanding 1
## 18    experience 1
## 19    government 1
## 20     hazardous 1
## 21       impacts 1
## 22    injustices 1
## 23        issues 1
## 24           job 1
## 26         learn 1
## 27       learned 1
## 28         legal 1
## 29 neighborhoods 1
## 30       prevent 1
## 32         stuff 1
## 33  surroundings 1
## 34         teach 1
## 35         tools 1
## 36         waste 1
```

Visualization

Make a word cloud out of it

```r
data_words_small %>% with(wordcloud(word, n, max.words = 40, scale = c(2.5, 0.25)))
```

Bigrams

Let's do the same analysis as before, except through bigrams. ### Data Preparation and Processing

```
data_bigrams <- na.omit(df %>% unnest_tokens(bigram, skills_gain,
                                             token = "ngrams",
                                             n = 2))
head(data_bigrams)
```

```
##      id               bigram
## 1 hI7d5           learn about
## 2 hI7d5   about environmental
## 3 hI7d5 environmental justice
## 5 h3Q9s          awareness of
## 6 h3Q9s          of pollution
## 7 BKFrF          my knowledge
```

Get bi-grams into separate words so we can remove stop words

```
bigrams_separated <- data_bigrams %>% separate(bigram, c("word1", "word2"),
                                               sep = " ")
bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

# new bigram counts:
bigram_counts <- bigrams_filtered %>%
```

```
  count(word1, word2, sort = TRUE)

head(bigram_counts)
```

```
##           word1      word2 n
## 1 environmental    justice 3
## 2        enviro injustices 1
## 3     expanding  knowledge 1
## 4    government      stuff 1
## 5     hazardous      waste 1
## 6           job experience 1
```

Looks like most of the words are titles and names linked to those titles. Let's bring both words in the bi-grams together.

```
bigrams_united <- bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ")
head(bigrams_united)
```

```
##      id                bigram
## 1 hI7d5 environmental justice
## 2 vroUV     enviro injustices
## 3 vroUV              san jose
## 4 1fMAt        job experience
## 5 1fMAt   expanding knowledge
## 6 1kY4J       hazardous waste
```

Visualization

Create an igraph object from tidy data

```
library(igraph)

# original counts
head(bigram_counts)
```

```
##           word1      word2 n
## 1 environmental    justice 3
## 2        enviro injustices 1
## 3     expanding  knowledge 1
## 4    government      stuff 1
## 5     hazardous      waste 1
## 6           job experience 1
```

```
# only look at relatively common combinations
bigram_graph <- bigram_counts %>% graph_from_data_frame()
bigram_graph
```

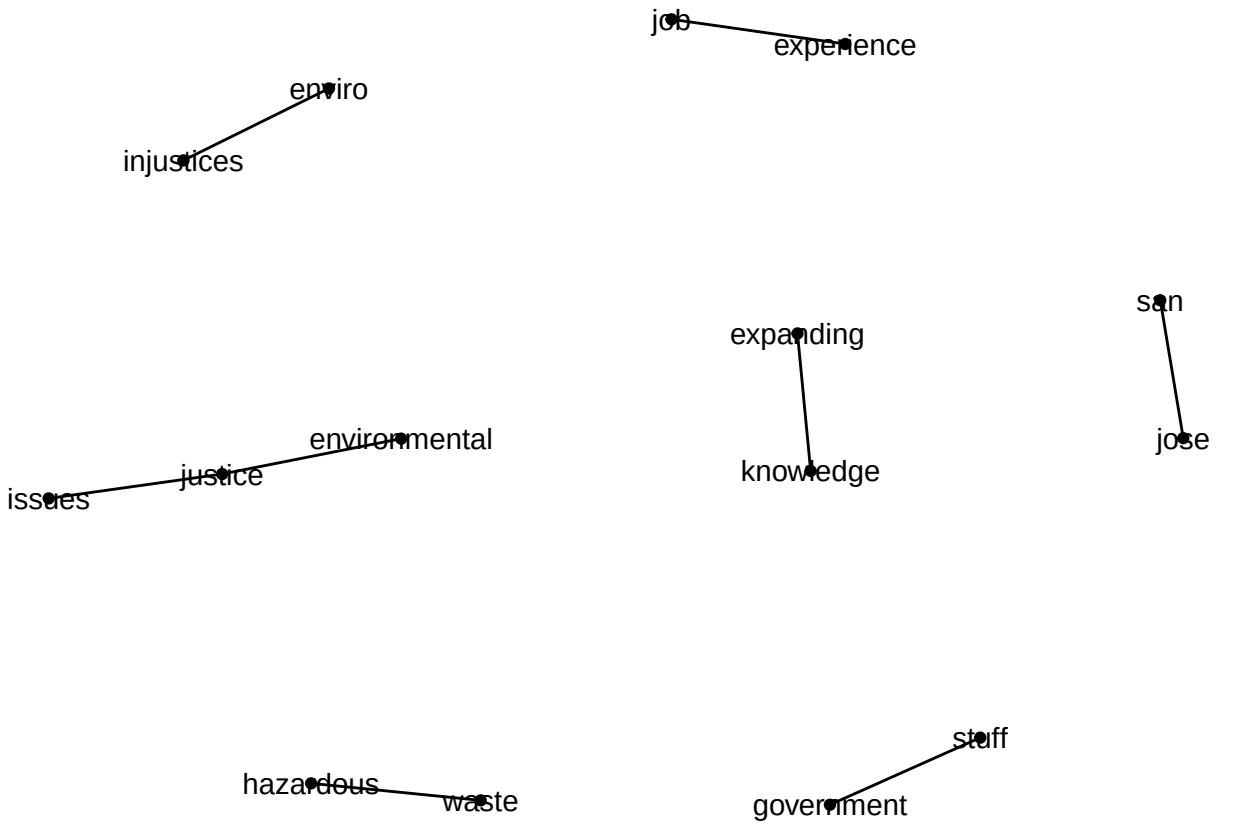
```
## IGRAPH e296986 DN-- 15 8 --
## + attr: name (v/c), n (e/n)
## + edges from e296986 (vertex names):
## [1] environmental->justice    enviro       ->injustices
## [3] expanding    ->knowledge  government   ->stuff
## [5] hazardous    ->waste      job          ->experience
## [7] justice      ->issues     san          ->jose
```

We can convert an igraph object into a ggraph with the "ggraph" fcn. Then, add layers to it - nodes, edges,

text.

```
set.seed(2017)

ggraph(bigram_graph, layout = "fr") + geom_edge_link() + geom_node_point() +
  geom_node_text(aes(label = name), vjust = 0.5, hjust = 0.5) + theme_void()
```



NOTE: The analysis isn't as useful here as there aren't enough responses to fully get a diverse portfolio of bigrams. However, despite the small population of the survey respondents (at the time this survey was deployed), we can see some common word associations within the responses of the training attendees.

Here, we can see that when it comes to the environment, there's a strong drive among attendees to address social injustices and issues related to the environment. We also see a desire to gain knowledge and work experience on how to tackle these issues with a brief mention of government policy (?) and San Jose itself (the city at which the attendees are based at).