TAD Week 2 Assignment

Tommy Klein

Part 1

1. Supervised learning is machine learning with labelled data. With supervised learning, the machine learning algorithm receives training data, which is labelled. It then uses these labels to learn about the underlying correlations in the data, and can then apply the labels to un-labelled data. For example, if you had the text of books, and the genre of each book, you could use a supervised learning method to learn how to classify books into pre-defined genres. You could then apply this model to other books to label them with a genre.
2. Supervised learning must have labelled data. The algorithm needs to have some data where the outcome of interest (e.g. sentiment, topic, etc.) has already been provided. It then uses this data to learn how to apply those outcomes to data that it hasn't yet seen.
3. A training data set is a labelled data set that is fed to a supervised learning algorithm. The algorithm uses the training data to learn how to the data is correlated with the outcome variable of interest. After using the training data, the algorithm can then apply the outcome variable to data that it hasn't yet seen.
4. Unsupervised learning is machine learning without labelled data. Unsupervised learning algorithms classify data into different groups. Since the data is not labelled, the groups that are created are not pre-defined, and thus might not be easily interpretable.
5. Unsupervised learning differs from supervised learning primarily in how the algorithm is trained. Supervised learning algorithms are trained on labelled data, and learn how to apply pre-defined labels to data. Unsupervised learning models are trained on unlabeled data, and learn how to group data in to undefined groups.
6. Dictionary methods map words to different labels, such as sentiment. For example the General Inquiry Database maps ~3,600 words to either a positive or negative sentiment, or the Valence Aware Dictionary for Sentiment Reasoning, which maps ~7,500 terms to different emotions and emotional strenghts. After words have been mapped, they can be used to classify the documents into different labels, such as positive or negative.

Part 2

Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidytext)
```

Read in CSV

```
leg_data <- read_csv('19_20_Legislation_Title_Clean.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `Legislation Number` = col_character(),
##   URL = col_character(),
##   Congress = col_double(),
##   `Bill Text Type` = col_character(),
##   `Bill Text Type Code` = col_character(),
##   `Latest Title` = col_character(),
##   `HTML Url` = col_logical(),
##   `PDF Url` = col_character()
## )
```

```
leg_data %>%
  head()
```

```
## # A tibble: 6 x 8
##   `Legislation Numb~ URL            Congress `Bill Text Type` `Bill Text Type C~
##   <chr>              <chr>             <dbl> <chr>            <chr>
## 1 H.R. 8900          https://www.c~      116 Public Law       PL
## 2 H.R. 8472          https://www.c~      116 Public Law       PL
## 3 H.R. 8337          https://www.c~      116 Public Law       PL
## 4 H.R. 8276          https://www.c~      116 Public Law       PL
## 5 H.R. 8247          https://www.c~      116 Public Law       PL
## 6 H.R. 7440          https://www.c~      116 Public Law       PL
## # ... with 3 more variables: Latest Title <chr>, HTML Url <lgl>, PDF Url <chr>
```

Unnesting Text and Counting

```
leg_data %>%
  unnest_tokens(word, `Latest Title`) %>%
  nrow()
```

```
## [1] 3548
```

```
leg_data %>%
  unnest_tokens(word, `Latest Title`) %>%
  select(word) %>%
  distinct() %>%
  nrow()
```

```
## [1] 1118
```

There are 3548 tokens in the data, and of those, 1118 are unique.

Tokens in the 5th Document

```r
leg_data[5, ] %>%
  unnest_tokens(word, `Latest Title`) %>%
  nrow()
```

```
## [1] 5
```

There are five tokens in the fifth document.