Lab-12

your name here

4/20/23

Preface

The goal of this assignment is to help you gain more familiarity with processing text data. As always, please come to office hours and reach out to your teaching staff if you have any questions.

Data

We will work with data on data scientist job postings in the U.S. scraped from popular job boards by JobSpikr.

```
job_posts <- read_csv("data_scientist_united_states_job_postings.csv") |>
  select(-cursor, -contains("contact"), -uniq_id, -html_job_description) |>
  relocate(crawl_timestamp, url, .after = last_col())
job_posts |>
  head(5)
```

```
# A tibble: 5 x 17
  job_t~1 categ~2 compa~3 city  state country infer~4 infer~5 infer~6 post_date
  <chr>   <chr>   <chr>   <chr> <chr> <chr>   <chr>   <chr>   <chr>   <date>
1 Enterp~ Accoun~ Farmer~ Wood~ CA    Usa     Woodla~ Califo~ Usa     2019-02-06
2 Data S~ <NA>    Luxoft~ Midd~ NJ    Usa     Middle~ New je~ Usa     2019-02-05
3 Data S~ <NA>    Cincin~ New ~ NY    Usa     New yo~ New yo~ Usa     2019-02-05
4 Data S~ Accoun~ BlackR~ New ~ NY 1~ Usa     New yo~ New yo~ Usa     2019-02-06
5 Senior~ biotech CyberC~ Char~ NC    Usa     Charlo~ North ~ Usa     2019-02-05
# ... with 7 more variables: job_description <chr>, job_type <chr>,
#   salary_offered <chr>, job_board <chr>, geo <chr>, crawl_timestamp <chr>,
#   url <chr>, and abbreviated variable names 1: job_title, 2: category,
#   3: company_name, 4: inferred_city, 5: inferred_state, 6: inferred_country
```

1. Let's start by looking at the job title. We see from the first few entries that most job titles include "data scientist." Tokenize `job_title` to bigrams (i.e., n-grams with n=2), and use a bar chart to show the top ten bigrams that appear in `job_title`. What are they? Do they make sense to you?

2. From question 1 we see that some of the job titles include words indicating the job level, such as "senior", "sr", "lead", "principal", etc. Use `str_detect()` to classify jobs into three different levels: "junior", "senior", and "principal", based on the description of the job title. Then use a bar chart to show the corresponding number of postings for each level.

3. Let's look at the category of the job. Tokenize `category` into individual words and use a bar chart to show the top 10 words.

4. Try using a word cloud to visualize the `category` text. Use your tokenized text from question 3 to make a word cloud plot using `wordcloud()` function. Include `scale = c(2, .5)` (or similar) as an argument to ensure all the words render properly in the pdf. Does the plot seem easy to digest?

5. Where are these jobs located? Use a bar chart to show the number of job postings of the top 10 cities.

6. What software skills are most commonly required for these jobs? To find out, create logical variables to indicate whether each `job_description` contains skill requirements, such as excel, python, R, tableau, java, sql, matlab, etc. Then calculate the share of postings that require each of these skills, and show them in a bar plot. Do your results make sense? If not, can you improve them?

7. Do something else interesting with the data. For example, you may explore a bit more about the job description, and see whether you might be able to find useful information such as minimun working experiences, salary ranges, etc. Another option would be to explore education requirements for the jobs. Get creative and have fun!

Note: unlike in lab-11, this question is NOT optional.