Populate Keywords Database

Anthony Arroyo, Josh Forster

10/22/2022

Datasets

We pulled a variety of job descriptions regarding data science jobs. Shown below are some the tweakable parameters to create the database.

```
job = jsonlite::read_json(here::here("data/job_description_data.json"))
jobIdVector = sapply(job, function(x) { x$link }) %>%
  magrittr::set_names(paste0("linkedin_", 1:length(.)), .)

data.frame(job_id = jobIdVector, job_url = names(jobIdVector)) %>%
  write.csv(., here::here("data/keyword-posting-crosswalk.csv"), row.names = FALSE)

writeFiles = rep(here::here("data/keywords_linkedin/"), length(jobIdVector))

dontWrite = TRUE
if (dontWrite) {
  writeFiles = NA
}

head(jobIdVector, 2)
```

```
## https://www.linkedin.com/jobs/view/data-scientist-%E2%80%93-operations-research-at-bosch-usa-3302078
##
##                                 https://www.linkedin.com/jobs/view/data-scientist-at-big-cloud-330939
##
```

Captures from the English Dictionary

To determine our stop words, we used Webster's dictionary to select nouns, verbs, and adjectives.

```
captureGroups = c("n.", "a.", "v.")
dictionary = here::here("data/dictionary.json") %>%
  readLines(.) %>%
    jsonlite::fromJSON(.)

head(dictionary[-1:-705, -3])
```

```
##      pos        word                       synonyms
## 706   v.      ACCORD                           <NA>
## 707   a.  ACCORDABLE                           <NA>
## 708   n.  ACCORDANCE Harmony; unison; coincidence.
## 709   n.  ACCORDANCY                           <NA>
## 710   a.   ACCORDANT                           <NA>
## 711 adv. ACCORDANTLY                           <NA>
```

Finished Aggregates

The data created shows the keyphrase, number of occurences across all inputs, and number of words. We can use this to determine the most common keyphrases in Data Science job descriptions.

```
values = GenerateKeywords(job, jobIdVector, writeFiles, dictionary, captureGroups, GrabLinkedin)
aggregates = SumFreq(values)
singularAggregates = aggregates %>% .[.$numWords == 1, ]
writeReports = FALSE

if (writeReports) {
  write.csv(aggregates, here::here("data/outputs/aggregateLinkedinPhrases.csv"), row.names = F)
  write.csv(singularAggregates, here::here("data/outputs/aggregateLinkedinKeywords.csv"), row.names = F)
}

head(aggregates[-1:-10, ], 20)
```

```
##             keyword sumFreq numWords
## 11             WORK      50        1
## 12     DATA SCIENCE      44        2
## 13 MACHINE LEARNING      42        2
## 14          MACHINE      41        1
## 15         ANALYSIS      39        1
## 16              ARE      39        1
## 17             WILL      37        1
## 18        ANALYTICS      35        1
## 19             TEAM      33        1
## 20           MODELS      31        1
## 21      ENGINEERING      30        1
## 22               BE      29        1
## 23         RESEARCH      28        1
## 24         MODELING      27        1
## 25              NEW      27        1
## 26           STRONG      27        1
## 27          ABILITY      26        1
## 28           OF THE      26        2
## 29      STATISTICAL      26        1
## 30       ABILITY TO      25        2
```