PDF Extraction (Part 1: Text)

Ryan Atkinson

2022-04-08

A brief introduction to pdf extraction, tokenization, data type conversion, and cleaning methods

Partitioning the data from the pdf into complete sentences as a tibble

Below are the first ten complete sentences pulled from the pdf. The pdf is from the Bureau of Labor Statistics, Monthly Labor Review, for June 2021.

sentences
1 Larry Akinyooye akinyooye.larry@bls.gov Larry Akinyooye is an economist in the Office of Employment and Unemployment Statistics, U.S.
Bureau of Labor Statistics.
Eric Nezamis nezamis.eric@bls.gov Eric Nezamis is an economist in the Office of Employment and Unemployment Statistics, U.S.
Bureau of Labor Statistics.
As the COVID-19 pandemic affects the nation, hires and turnover reach record highs in 2020 Data from the Job Openings and Labor Turnover Survey (JOLTS) highlight the effects of the coronavirus disease 2019 (COVID-19) pandemic and the results of efforts to mitigate its spread in 2020.
With the challenges of the pandemic, many of the JOLTS data elements experienced shocks early in the year before returning to previous trends.
In fact, many of the data elements experienced series highs.
For example, the hires level reached a series high of 8.3 million in May 2020, bouncing back from a depressed level of 3.9 million in April 2020.
The total separations level, also referred as turnover, reached a series high of 16.3 million in March 2020, boosted largely by a spike in layoffs and discharges.
The Job Openings and Labor Turnover Survey (JOLTS) data show that job openings, hires, and total separations experienced large movements early in 2020 in the wake of an economic recession because of the coronavirus disease 2019 (COVID-19) pandemic.1 After the initial economic downturn, many of the JOLTS data series started to return to prepandemic levels.

Partitioning the data from the pdf into individual tokens

The data below have been "tokenized." In brief, a token is the string of characters delineating a word. After tokenizing, "stop words" – words like articles, conjunctions, and related filler words – are removed. Lastly, instances of a word within a document and the proportion of that word within the whole documented are printed.

Table 2: Words, Counts, and Proportions

word	instance_of_word	prop
separations	122	1.700348
level	99	1.379791
region	98	1.365854
million	97	1.351916
percent	96	1.337979
december	90	1.254355
layoffs	90	1.254355
total	90	1.254355
discharges	89	1.240418
services	83	1.156794

Wordcloud

A wordcloud of the most common words within the documented is printed.