TEXT MINING

MIS 612 - 675

GROUP 4

TABLE OF CONTENTS

01

INSTALLATION OF R PACKAGES

- ❑ Load packages
- ❑ Import data

02

DATA TRANSFORMATION

- ❑ Create Corpus
- ❑ Data Pre-processing
- ❑ Tokenizing by BI-gram

03

TEXT ANALYSIS

- ❑ Document Text Matrix
- ❑ Term frequency statistic and term correlation

04

TEXT VISUALIZATION

- ❑ Bar plot
- ❑ Word Cloud

DATA TRANSFORMATION

1 STEP 1: Create a corpus to put the dataset

2 STEP 2: Tokenization and cleaning of the uninformative dataset using tm_map()

3 STEP 3: Create a Document text matrix using bigrams to tokenize 2-grams terms and order by most frequency terms.

```
student verification           work ethic           team player        student great        quick learner
                  85                   50                    48                   46                   38
 communication skill
                  36
```

4 STEP 4: Explore the most frequent terms and their correlations

```
$`work ethic`
strong work
        0.46

$`team player`
  great team    good team player alway
        0.46         0.37         0.31

$`communication skill`
 great communication verbal communication       written verbal
                0.36                 0.33                 0.33
```

- **Work ethic** is 46% associated to the term “strong work”.
- **Team player** is also associated to 3 different terms 46%, 37%, and 31% accordingly.
- **Communication skill** is associated with the term, great communication, verbal communication, and written verbal

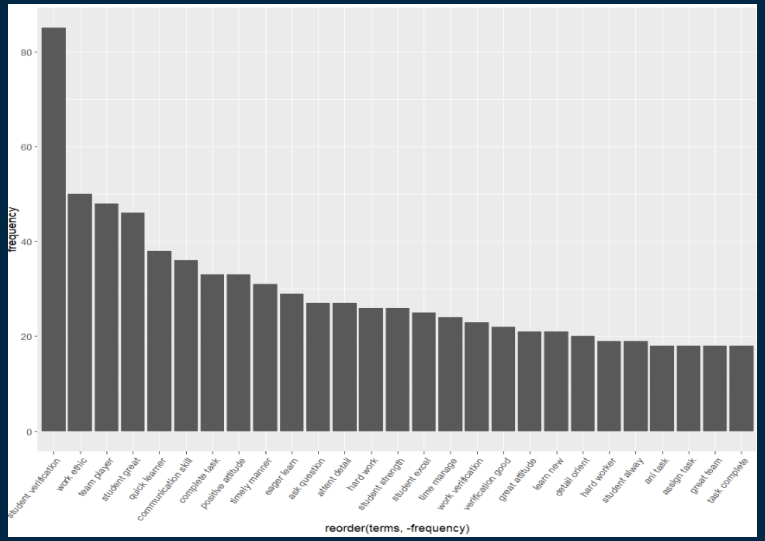
TEXT ANALYSIS

Term	Frequency	pct_corp	culm_sum
student verification	85	0.008900524	0.008900524
work ethic	50	0.005235602	0.014136126
team player	48	0.005026178	0.019162304
student great	46	0.004816754	0.023979058
quick learner	38	0.003979058	0.027958115
communication skill	36	0.003769634	0.031727749
complete task	33	0.003455497	0.035183246
positive attitude	33	0.003455497	0.038638743
timely manner	31	0.003246073	0.041884817
eager learn	29	0.003036649	0.044921466
ask question	27	0.002827225	0.047748691
attent detail	27	0.002827225	0.050575916
hard work	26	0.002722513	0.053298429
student strength	26	0.002722513	0.056020942
student excel	25	0.002617801	0.058638743
time manage	24	0.002513089	0.061151832
work verification	23	0.002408377	0.063560209
verification good	22	0.002303665	0.065863874
great attitude	21	0.002198953	0.068062827
learn new	21	0.002198953	0.070261780

According to bi-grams text analysis, the last 4 most frequent terms make up almost 7% of the total terms.

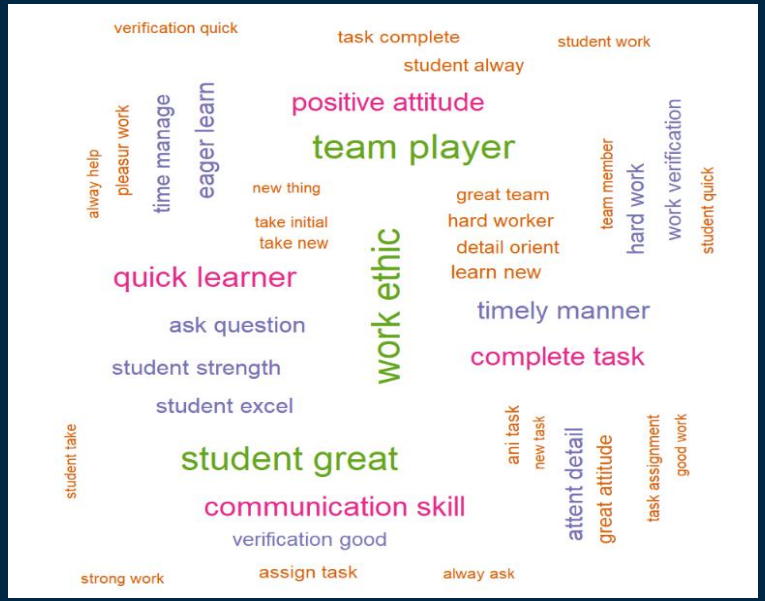
1. Team manage
2. Work verification
3. Great attitude
4. Learn new

TEXT VISUALIZATION



Bar plot for frequency terms

Word cloud for frequency terms



REFERENCES

Anandarajan, M., Hill, C., & Nolan, T. (2019). Practical Text Analytics. In Advances in Analytics and Data Science. Springer International Publishing. https://doi.org/10.1007/978-3-319-95663-3

Girdher, H. (2021, July 30). TDM (Term Document Matrix) and DTM (Document Term Matrix). Analytics Vidhya. https://medium.com/analytics-vidhya/tdm-term-document-matrix-and-dtm-document-term-matrix-8b07c58957e2

Hutson, G. (2021, March 17). Text Mining – Term Frequency analysis and Word Cloud creation using the tm package | R-bloggers. https://www.r-bloggers.com/2021/03/text-mining-term-frequency-analysis-and-word-cloud-creation-using-the-tm-package/

Kirenz, J. (2019, September 16). Text Mining in R. Jan Kirenz. https://www.kirenz.com/post/2019-09-16-r-text-mining/

Lonkar, S. (2019, January 28). Text Mining in Data Science. Text Mining in Data Science— a Tutorial of Text Mining in R Using TM Package. https://medium.com/text-mining-in-data-science-a-tutorial-of-text/text-mining-in-data-science-51299e4e594