Large Scale Data Exploration with R

Ashraf, Shawon

26 June, 2020

1 Norms

1.1 Dataset

For this data exploration project, I have used the Norms dataset presented by Brysbaert et al. [1] which contains the concreteness ratings for 40 thousand generally known English word lemmas. According to the authors, concreteness is the measurement of the concept a word demotes to an entity. The concept of concreteness of words came Paivio's dual-coding theory [2] which states that concrete words are easier to recall and activate in memory compared to non concrete words. Also, Schwanenflugel et al. [3] presented that concrete words are easier to recall because of the supporting memory context imposed by the words on entities to the degree abstract words can not. Vigliocco et al. [4] and Andrews et al. [5] presented a semantic theory which states that the learning process of words are more based on direct experience of the learners.

The authors based their presentation of the Norms based on Connell et al. [6] which states that despite words being learnt on direct experiences, the existing concreteness ratings were too much focused on visual perception whereas Lynott et al. found that the concreteness ratings were correlated not only on visual perception but also on touch and smell. The overcome the limitations of the existing datasets, the authors came up with the dataset in use which was collected by asking English speakers to rate the concreteness of the words based on their knowledge on them.

The Norms dataset consist of 8 columns :

1. Word
2. Whether the word is a Bigram or not
3. Mean concreteness rating
4. Standard deviation of the concreteness ratings
5. Number of persons not knowing the word
6. Total number or persons rating words
7. Percentage of persons knowing the word
8. SUBTLEX-US frequency count of the word [7]

1.2 Variables chosen

- Mean concreteness rating
- Standard deviation of the concreteness ratings
- Percentage of persons knowing the word

2 Preprocessing

The provided datasets were processed using unix tools such as awk, cat, etc.

2.1 Windows

- POS : NN
- Frequency count : Greater than 20000
- Excluded characters : ":::"
- The processed output was converted to a csv file for importing into R
- Reduced from 49776264 items to 11046 after processing

2.2 Frequencies

- POS : NN
- Frequency count : Greater than 5000
- The processed output was converted to a csv file for importing into R
- Reduced from 326775 items to 59947 after processing

2.3 Norms

The Norms dataset was already provided in .xlsx format and did not need preprocessing via Unix tools. Norms with words having the POS tag Noun were extracted using R function 'subset' during analysis.

3 Descriptive Statistics and plots of Norms

3.1 Descriptive Statistics

GG

References

[1] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known english word lemmas," *Behavior Research Methods*, vol. 46, no. 3, pp. 904–911, 2013.

[2] M. Sadoski and A. Paivio, *Imagery and text*. Routledge, 2013.

[3] P. J. Schwanenflugel, K. K. Harnishfeger, and R. W. Stowe *Journal of Memory and Language*, vol. 27, no. 5, pp. 499–520, 1988.

[4] G. Vigliocco, D. P. Vinson, W. Lewis, and M. F. Garrett, "Representing the meanings of object and action words: The featural and unitary semantic space hypothesis," *Cognitive Psychology*, vol. 48, no. 4, pp. 422–488, 2004.

[5] M. Andrews, G. Vigliocco, and D. Vinson, "Integrating experiential and distributional data to learn semantic representations.," *Psychological Review*, vol. 116, no. 3, pp. 463–498, 2009.

[6] L. Connell and D. Lynott, "Strength of perceptual experience predicts word processing performance better than concreteness or imageability," *Cognition*, vol. 125, no. 3, pp. 452–465, 2012.

[7] M. Brysbaert and B. New, "Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english," *Behavior Research Methods*, vol. 41, no. 4, pp. 977–990, 2009.