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, Harnishfeger, and Stowe (1988) 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, Vinson, Lewis, and Garrett (2004; see also Andrews, Vigliocco, and Vinson, 2009) 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 and Lynott, 2012; Lynott and Connell, 2009 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

1.2 Variables chosen

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.