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 and 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 Mean Concreteness

The minimum mean concreteness value is 1.07 and the variable ranges from 1.07 to 5.0. Median and mean values are 3.66 and 3.53 respectively. While the first quartile ranges to 2.67, the third quartile peaks at 4.46.

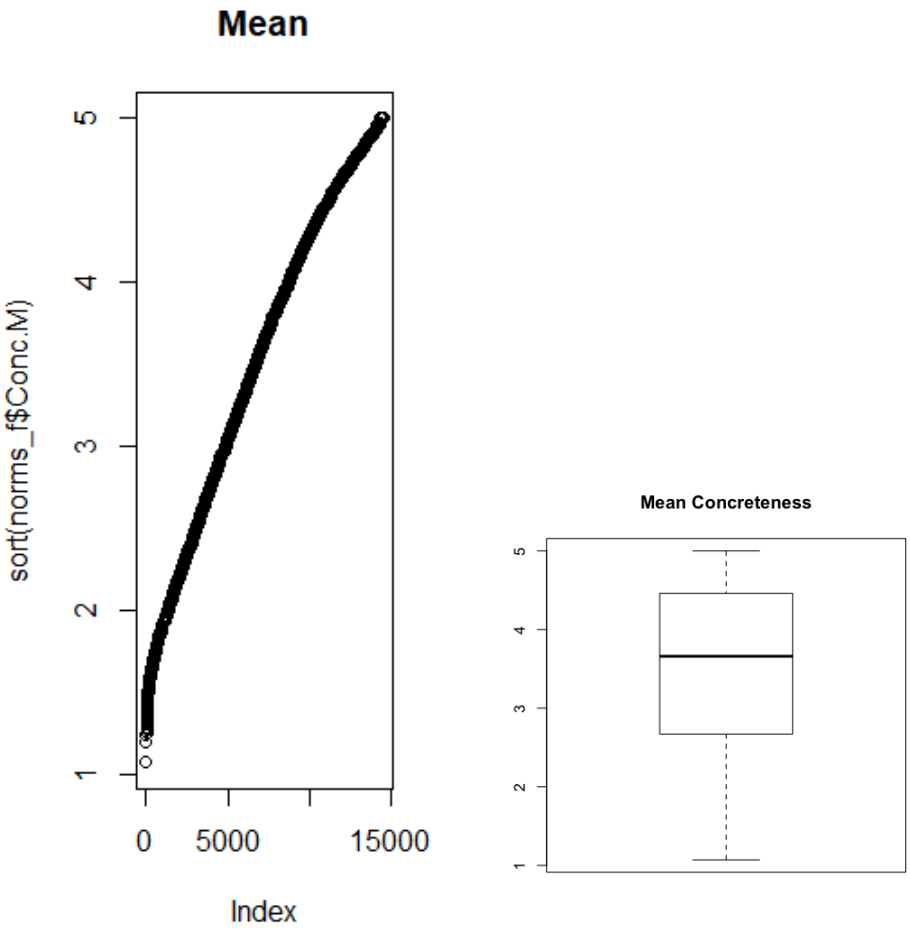

Figure 1: Mean concreteness

3.2 Standard Deviation of Concreteness

The minimum value is 0 and the variable ranges from 0 to 1.890. Median and mean values are 1.2 and 1.109 respectively. While the first quartile ranges to 0.920, the third quartile peaks at 1.370.

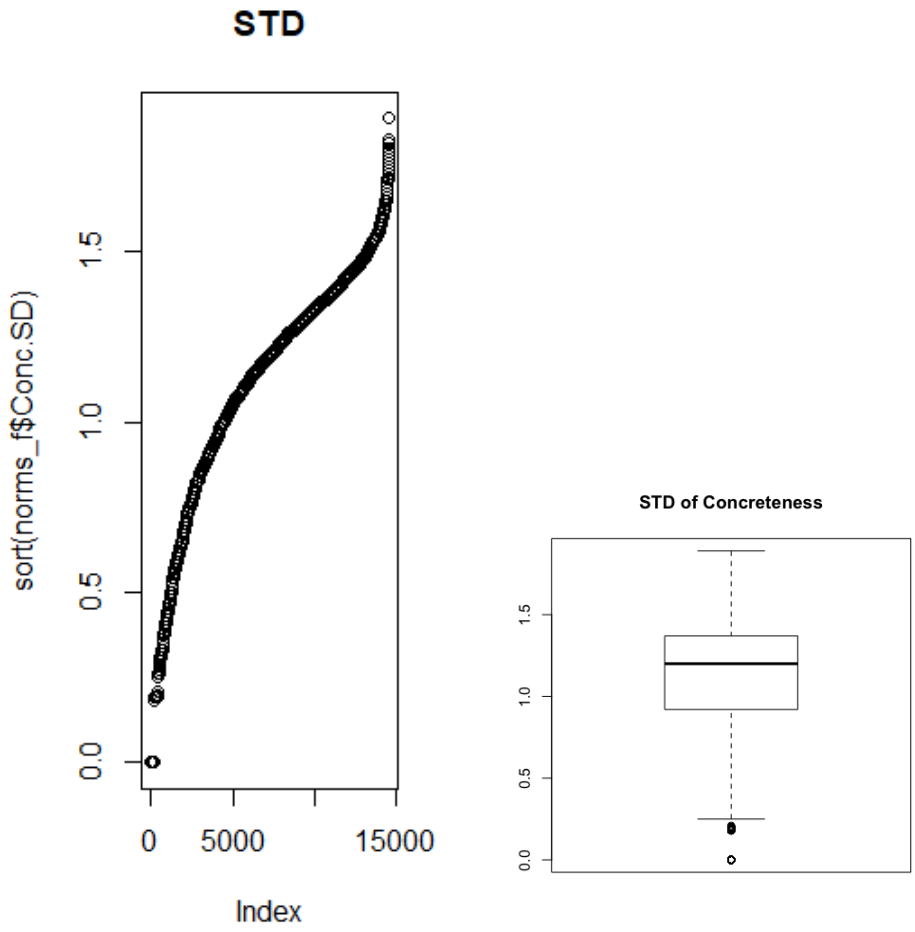

Figure 2: Standard deviation of concreteness

3.3 Percentage Known

The minimum value is 0.8462 and the variable ranges from 0.8462 to 1. Median and mean values are 1 and 0.9720 respectively. While the first quartile ranges to 0.9630, the third quartile peaks at 1.

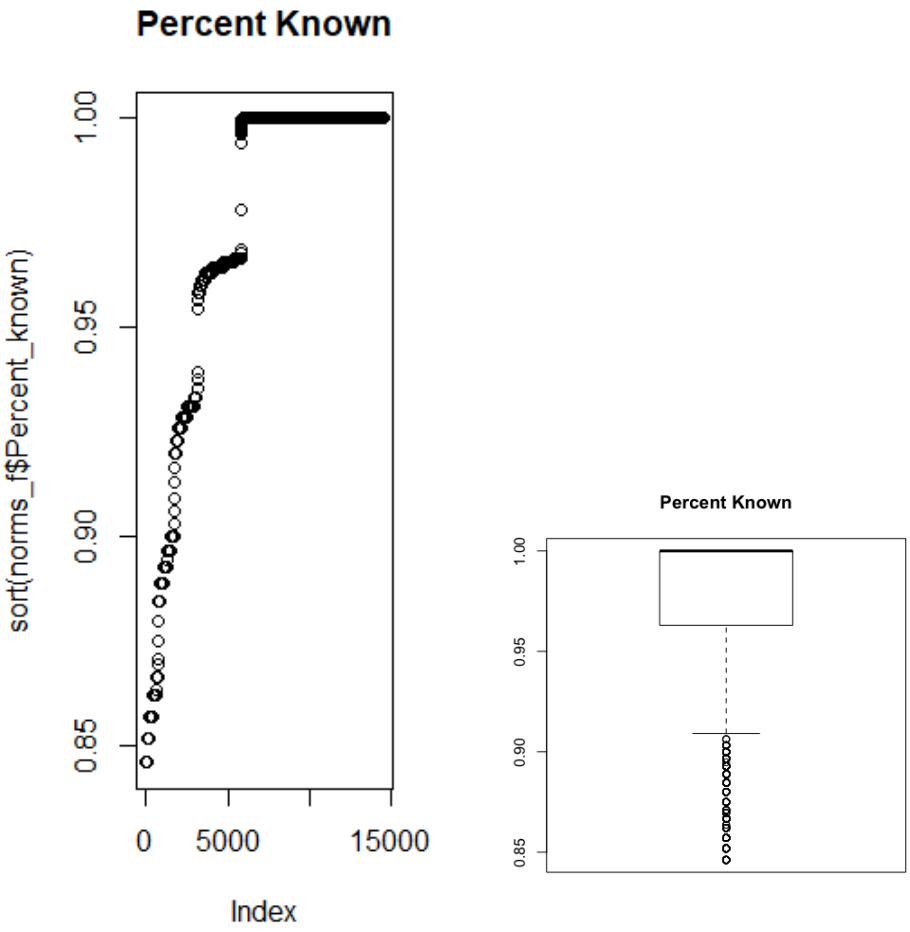

Figure 3: Percentage Known

4 Hypothesis Testing Methods

4.1 The ratings of the chosen variables are related to each other across the target words

The Windows dataset contains multiple instances of the same dataset as it also contains the context word related to each of them. To minimize them, the means of context frequency and lmi for the words taken and put into a vector. Then norms for the target words were loaded to run correlation on based on the norm variables. For correlation the lsr[8] package was used where Pearson Correlation Coefficient[9] is the default metric for finding correlation.

The correlation results were as follows:

target	mean	std	percent known
frequency	-0.029	0.040	0.010
lmi	0.033	-0.008	0.020

4.2 The target ratings in the norms are related to target corpus frequency

For this hypothesis, norms were first loaded for the target words and then their correlation was measured. The result of the correlation was the following:

target	mean	std	percent known
freq	-0.008	0.016	-0.013

4.3 The target ratings in the norms are related to the semantic diversity of their distributional nearest neighbours

- Target words and their norms were loaded into vectors
- Frequency and lmi of the target words were normalized using $(x - min(x))/(max(x) - min(x))$
- $k = 13$
- Categorical data was converted to numeric data using R method as.numeric
- Feature and target (Word) vectors were created
- Cosine score was computed for each of the nearest neighbors of the target words using

$$cos(u, v) = \frac{uv}{\|u\| \, \|v\|}$$

  and then their mean was computed
- The mean was correlated with Norms variables for measuring semantic diversity.

The results were as follows:

target	mean	std	percent known
avg cos	-0.029	0.026	-0.007

5 Insights

5.1 Hypothesis 1

- Frequency is positively correlated with standard deviation and percent known variables. Which indicate that the words with higher frequency are more prominent and more known to persons who were asked to rate the words

- For lmi, positive correlation is seen with the mean and percent known variables.
- Given the hypothesis, it can be said that only percent known variable is positively correlated with both frequency and lmi of the target words. Other variables are not related with all the target word variables.

5.2 Hypothesis 2

- Only standard deviation of norms have a positive correlation with target word frequencies. Form this result it can be said the the frequency of the target words are more spread out than the ones in the norms.

5.3 Hypothesis 3

- Here too, only standard deviation is positively correlated to average cosine score of the target words. Here too can be said that the Norms dataset can not fully capture the semantic features of the target words, as presented by Brysbaert et al. [1] that the norm datasets, despite having a large number of entries may fail to capture all the semantic features.

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.

[8] D. Navarro, *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5)*. University of Adelaide, Adelaide, Australia, 2015. R package version 0.5.

[9] "Pearson correlation coefficient, wikiwand," 2020.