# Word Clouds to Visually Present the Most Informative Words in Subject Categories
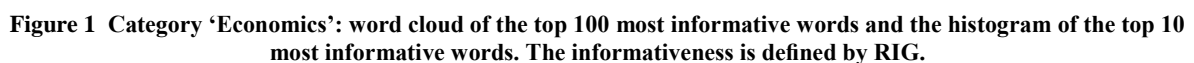
April 2020 by Neslihan Suzen, PhD student at the University of Leicester (ns433@leicester.ac.uk / suzenneslihan@hotmail.com )

Supervised by Prof Alexander Gorban and Dr Evgeny Mirkes

This publication presents word clouds of the most informative words in Web of Science (WoS) categories [1,2]. The clouds are created with words of the Leicester Scientific Dictionary-Core LScDC [3,4]. We consider the list of words with their Relative Information Gain (RIGs) in the corresponding category. For all categories, words are sorted by their RIGs in descending order and the top 100 words are shown in the word clouds [5]. The bigger size the word in word clouds, the more informative it is for the category. This study is a part of the research on the quantification of the meaning of research texts.

Word clouds for the top 100 most informative words and histograms of RIGs for the top 10 most informative words for each of 252 categories can be found in the archive published along with this description. The most informative 100 words with their RIGs for each of categories are presented in tables published.

As an example, the following figure shows the word cloud of the top 100 most informative words and the histogram of the top 10 most informative words for the category 'Economics' . The informativeness is defined by RIG of words. The most informative word in this category is 'market', followed by 'econom' and 'price'.



**Figure 1  Category 'Economics': word cloud of the top 100 most informative words and the histogram of the top 10 most informative words. The informativeness is defined by RIG.**

In tables published, the top informative 100 words for each of 252 categories are listed in descending order of RIG. The structure of the tables is shown with the top 15 words for the category 'Economics' in Table 1.

**Table 1 The list of the top 100 words in the category Economics with RIGs**

| No. | Word | RIG |
|---|---|---|
| 1 | market | $9.1 \times 10^{-2}$ |
| 2 | econom | $6.4 \times 10^{-2}$ |
| 3 | price | $6 \times 10^{-2}$ |
| 4 | polici | $5.8 \times 10^{-2}$ |
| 5 | economi | $5.1 \times 10^{-2}$ |
| 6 | countri | $5.1 \times 10^{-2}$ |
| 7 | financi | $4.8 \times 10^{-2}$ |
| 8 | firm | $4.4 \times 10^{-2}$ |
| 9 | empir | $3.4 \times 10^{-2}$ |
| 10 | capit | $3.2 \times 10^{-2}$ |
| 11 | invest | $3.2 \times 10^{-2}$ |
| 12 | trade | $3.1 \times 10^{-2}$ |
| 13 | sector | $2.7 \times 10^{-2}$ |
| 14 | find | $2.6 \times 10^{-2}$ |
| 15 | incom | $2.6 \times 10^{-2}$ |

## Published archive contains following files:

1. **Word_Clouds.pdf**: A file that contains all word clouds of the top 100 most informative words and the histogram of the top 10 most informative words for each of 252 WoS categories.
2. **Lists_of_Words .pdf:** Lists of the top informative 100 words for each of 252 WoS categories.

## References

[1]     Web of Science. (15 July). Available: https://apps.webofknowledge.com/
[2]     WoS Subject Categories. Available:
        https://images.webofknowledge.com/WOKRS56B5/help/WOS/hp_subject_category_terms_t
        asca.html
[3]     Suzen, Neslihan (2019): LScDC (Leicester Scientific Dictionary-Core). figshare. Dataset.
        https://doi.org/10.25392/leicester.data.9896579.v3
[4]     Suzen, N., Mirkes, E. M., & Gorban, A. N. (2019). LScDC-new large scientific
        dictionary. *arXiv preprint arXiv:1912.06858*.
[5]     Suzen, Neslihan (2020): LScDC Word-Category RIG Matrix. figshare. Dataset.
        https://doi.org/10.25392/leicester.data.12133431.v1