Quantitative Text Analysis of readers’ contributions on Japanese daily newspapers

Yasuto NAKANO

1. School of Sociology, Kwansei Gakuin University, JAPAN
‘Contact author: yasuto@soc-nakano.net

Keywords: text mining, correspondence analysis, readers contributions

0.0.1 Purpose

The purpose of this paper is to mine the big data of readers’ contributions in Japanese daily newspapers. Methodologies used here are text mining and correspondence analysis.

0.0.2 Data

There are several daily newspapers distributed throughout Japanese nation wide. Each of these papers is circulated for round 10 millions. The biggest two papers cover almost 40% of Japanese population. In such newspapers, there is a page for readers’ contributions, placing about 10 articles contributed by readers’. Occasionally there is a special topic offered by the editor, but in general, the topic of contributions depends on contributors’ concerns. We have set up cumulative data of this page for 20 years. One contribution includes not only text of the artice but also its contributor’s name, age, occupation and address. We extract these contributor’s attributes from the article text, and make them related to it. Analyzing cumulative contributions data tells us opinions of certain groups of Japanese. It tells us transitions and stability of opinions. Furthermore, this data tells us relations between topics and contributors’ attributes.

To analyze Japanese text data, we have to conduct morphological analysis on the data. We use RMeCab package, which utilize a morphological engine 'MeCab' from R. RMeCab provides us a Document Term Matrix(DTM). In case of ’ASAHI’, there are 50 thousands contributions. We morpholize about 70 thousands different words. Therefore our data of ’ASAHI’ consists of a matrix of 50 thousands article rows and 70 thousands words columns. A matrix of occupation and words, 3 thousands occupations rows and 70 thousands words columns, is also available. These matrix can be visualized with the package ca and igraph.

0.0.3 Analysis

For example, we choose topic of ’war’. Amongst all articles, 8% of ’ASAHI’ articles includes ’war’. Frequencies of war-related articles are changed by month. August is a month of ’war’. ’Teachers’ tend to mention on ’war’ more frequently than other occupations. The DTM tells us what kind of words are closely related to ’war’. It reveals how contributers have descrived on ’war’. Words of ’Misery’ and ’peace’ have kept close relation for 20 years. ’Responsibility’ had close relation to ’war’ 20 years ago, but it is rarely appeared in war article nowadays. Recently words of ’father’ and ’mother’ are conspicuous. Story of ’war’ has been changed from contributers’ own story to their parents’ story. ’WWII’ has been a main topic of war articles. But a word ’terror’ appears a new topic in 00’s.

This analysis is a example of textmining on our dataset with R.