AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Check for updates

The Stata Journal (2018)
18, Number 2, pp. 379–386

Content analysis: Frequency distribution of words

Mehmet F. Dicle
Loyola University New Orleans
New Orleans, LA
mfdicle@gmail.com

Betul Dicle
Research and Teaching Associates
Mandeville, LA
bkdicle@gmail.com

Abstract. Many academic fields use content analysis. At the core of most common content analysis lies frequency distribution of individual words. Websites and documents are mined for usage and frequency of certain words. In this article, we introduce a community-contributed command, `wordfreq`, to process content (online and local) and to prepare a frequency distribution of individual words. Additionally, another community-contributed command, `wordcloud`, is introduced to draw a simple word cloud graph for visual analysis of the frequent usage of specific words.

Keywords: dm0094, wordfreq, wordcloud, word counting, frequency distribution, content analysis, word cloud

1 Introduction

One of the most cited studies in content analysis in political science, Laver, Benoit, and Garry (2003), compares the efficiency of traditional methods with their method of word frequencies. On one side, there is the method of hand collecting, which requires much time and effort and is therefore costly. On the other side, there is machine automation of the content, which can be quite reliable and replicable. However, sophisticated phrase-recognition algorithms can be expensive and need frequent adjustments. Most importantly, phrase algorithms may not be as available in every language as they are for English. In fact, Laver, Benoit, and Garry (2003, 323) refer to their word-frequency systems as the "language-blind word scoring technique". Hopkins and King (2010) provide a detailed summary of historical use of content analysis in political science and propose a new nonparametric method. More recently and again in political science, Grimmer and Stewart (2013) emphasize the importance of content analysis and provide a detailed evaluation of some of the most popular models.

Within the context of psychology, Chung and Pennebaker (2013) summarize how computer automated systems can be used in lab and clinical studies. They emphasize the importance of individual words: "That is, much of the variance in language to identify psychopathologies, honesty, status, gender, or age, was heavily dependent on the use of little words such as articles, prepositions, pronouns, etc., more than on content words (for example, nouns, regular verbs, some adjectives and adverbs)" (Chung and Pennebaker 2013, 2). Authors refer to a word-frequency software (Linguistic Inquiry and Word Count) developed by Pennebaker, Francis, and Booth (2001) that is used to predict health status improvements based on the use of words.

© 2018 StataCorp LLC
dm0094

Similar content analysis studies have appeared in other fields. Here are a few important works in their respective fields: Downe-Wamboldt (1992) evaluates the issue for healthcare, Roberts (1989) for linguistics, Kassarjian (1977) and Kolbe and Burnett (1991) (a review of 128 studies) for consumer research, and Scott (1955) (one of the oldest studies in content analysis literature) for public opinion.

`wordfreq` is a simple code that would assist researchers in their specific content analysis research projects. It provides a word list as inclusive as possible without much modifications to avoid bias. Finally, `wordcloud` provides a sample word cloud graph that uses Stata's own scatter graphs. While the word cloud chart is simple, the code that generates the chart is provided to the user for possible modification, betterment, and adaptation to individual needs.

2 Word-frequency distribution: wordfreq

2.1 Title

`wordfreq` downloads a webpage or a local file and prepares frequency distribution of all different words.

2.2 Syntax

`wordfreq using` *filename* [`, min_length(`*integer*`) nonumbers nogrammar nowww`
`nocommon clear append`]

filename is the filename to process. It can be an Internet address to download, in which case it must start with http or https. It can also be a local file with any extension. The ASCII source of the file will be processed.

2.3 Description

`wordfreq` processes a webpage or a local file and prepares frequency distribution of all different words contained in the processed file. Once the content is processed as a single string, all noncharacters are replaced with space characters. An ASCII character list includes all characters between A–Z, a–z, and 0–9. Characters also include non-English letters. The entire string, stripped from noncharacters, is then split by the space character. In terms of the online content, many websites include news as part of a JavaScript code (for example, cnn.com, finance.yahoo.com, etc.). Thus, the content string is not limited to text between meaningful HTML tags (for example, table, td, tr, etc.) and includes text between code-related tags as well (for example, script). Text within tags, however, is eliminated (that is, "td width=80%" within "<td width=80%>" is eliminated). Because the text between code-related tags is not eliminated, the word list includes nonwords that are included within these sections (for example, var, int, foreach, forval, etc.). These may result in long variable names that web developers use in their coding. Four different lists of exclusion are made available to users for convenience. All words that contain numbers, that are related to grammar, that are related to http or html, and that are most commonly used in everyday English can be dropped using these word lists.

2.4 Options

`min_length(`*integer*`)` specifies the minimum number of characters required in a word to keep it in the frequency distribution. The default is `min_length(0)` (that is, keep all words).

`nonumbers` specifies to drop the words that contain numbers. The default is to keep them.

`nogrammar` specifies to drop words that are part of common grammar (for example, is or are). The default is to keep them. The full list is available at http://researchforprofit.com/data_public/wordfreq/wordfreq_grammar.txt.

`nowww` specifies to drop words that are related to http or html (for example, html, http, or chrome). The default is to keep them. The full list is available at http://researchforprofit.com/data_public/wordfreq/wordfreq_www.txt.

`nocommon` specifies to drop most common and ordinary words (for example, over, after, or about). The default is to keep them. The full list is available at http://researchforprofit.com/data_public/wordfreq/wordfreq_common.txt.

`clear` clears the data in the memory.

`append` specifies to append the new word-frequency distribution to an existing word-frequency distribution.

2.5 Installation and updates

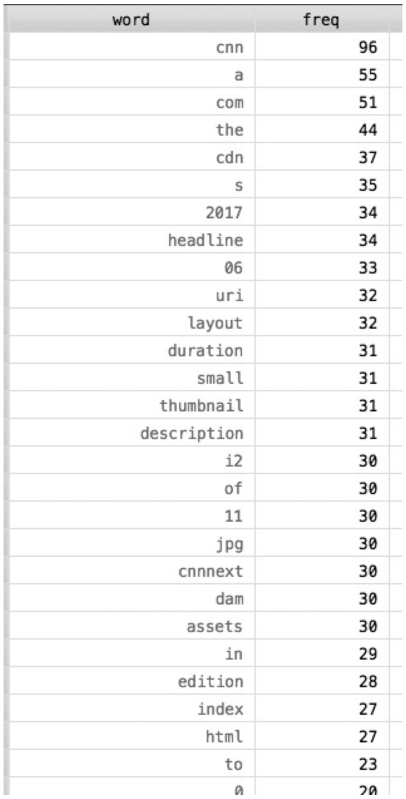
```
. net install "http://researchbtn.com/stata/110/wordfreq.pkg"
```

2.6 Usage

▷ Example: Simple word-frequency table

Figure 1 shows a simple word-frequency table downloaded from http://www.cnn.com on June 19, 2017. No common words, numbers, or grammar-related words are dropped. The minimum word length is not specified. Therefore, there are single characters as words.

```
. wordfreq using http://www.cnn.com
```



word	freq
cnn	96
a	55
com	51
the	44
cdn	37
s	35
2017	34
headline	34
06	33
uri	32
layout	32
duration	31
small	31
thumbnail	31
description	31
i2	30
of	30
11	30
jpg	30
cnnnext	30
dam	30
assets	30
in	29
edition	28
index	27
html	27
to	23
0	20

Figure 1. Word frequency for http://www.cnn.com on June 19, 2017

◁

3 Word cloud graph: wordcloud

3.1 Title

`wordcloud` draws a word cloud graph based on unique words and their frequencies.

3.2 Syntax

`wordcloud` *stringvar numericvar* [`,` `min_length(`*integer*`)` `nonumbers` `nogrammar`
`nowww` `nocommon` `style(1|2)` `showcommand` *twoway_options*]

stringvar is the variable name for the string variable that is to be used for the unique words. *numericvar* is the variable name for the numeric variable that is to be used for the frequency of the unique words.

3.3 Description

wordcloud draws a word cloud graph based on unique words included in a string variable and their associated frequencies. The command is a series of twoway scatter graphs with different mlabsize() values used for each. The size used for mlabsize() is based on the frequency distribution of the unique words. There are two styles provided with the command that differ mainly in mlabsize(). Users can specify the showcommand option to see the entire twoway graph command.

3.4 Options

min_length(*integer*) specifies the minimum number of characters required in a word to keep it in the frequency distribution. The default is min_length(0) (that is, keep all words).

nonumbers specifies to drop the words that contain numbers. The default is to keep them.

nogrammar specifies to drop words that are part of common grammar (for example, is and are). The default is to keep them. The full list is available at http://researchforprofit.com/data_public/wordfreq/wordfreq_grammar.txt.

nowww specifies to drop words that are related to http or html (for example, html, http, or chrome). The default is to keep them. The full list is available at http://researchforprofit.com/data_public/wordfreq/wordfreq_www.txt.

nocommon specifies to drop most common and ordinary words (for example, over, after, or about). The default is to keep them. The full list is available at http://researchforprofit.com/data_public/wordfreq/wordfreq_common.txt.

style(1|2) is the specific style of the graph to be drawn. Users can change mlabsize() in each graph to determine the readability of the graphs.

showcommand lists the command that is used to draw the graph produced by wordcloud.

twoway_options are any of the options documented in [G-3] ***twoway_options***. These additional options are simply added to the end of the command.

3.5 Installation and updates

```
. net install "http://researchbtn.com/stata/110/wordcloud.pkg"
```

3.6 Usage

▷ Example: Word cloud (style(1)) with exclusions

Figure 2 shows a word cloud (`style(1)`) for the word-frequency table downloaded from http://www.cnn.com on June 19, 2017, excluding word lists for numbers, grammar, http or html, and common English words. The minimum word length is set to three.

```
. wordfreq using https://www.cnn.com, clear
. wordcloud word freq, min_length(3) nonumbers nogrammar nowww nocommon style(1)
  (output omitted)
```

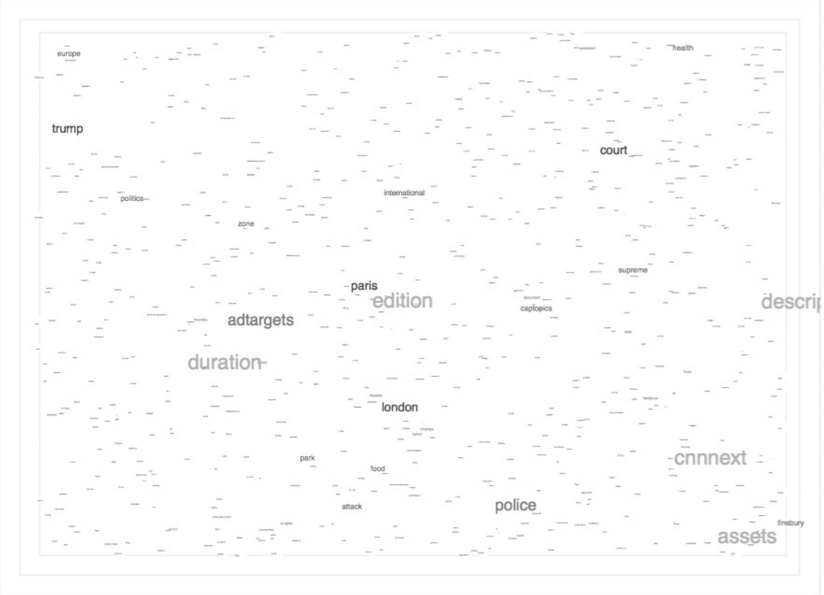


Figure 2. Word cloud (`style(1)`) for the word-frequency distribution for http://www.cnn.com on June 19, 2017

◁

▷ Example: Word cloud (style(2)) with exclusions

Figure 3 shows a word cloud (`style(2)`) for the word-frequency table downloaded from http://www.cnn.com on June 19, 2017, excluding word lists for numbers, grammar, http or html, and common English words. The minimum word length is set to three.

```
. wordfreq using https://www.cnn.com, clear
. wordcloud word freq, min_length(3) nonumbers nogrammar nowww nocommon style(2)
  (output omitted)
```

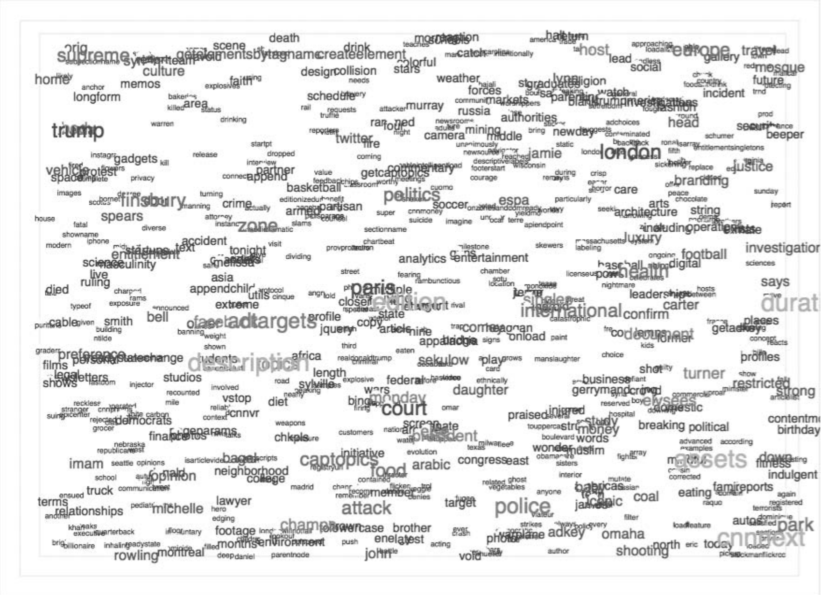


Figure 3. Word cloud (`style(2)`) for the word-frequency distribution for http://www.cnn.com on June 19, 2017

◁

4 Conclusion

Content analysis receives significant attention in literature for many academic fields. While phrase-based analysis is common, human-based evaluations can be biased and may be costly. Automated phrase-analysis systems are commercially available and provide replicable results. Word frequencies, however, are suggested as competing methods to resource-consuming phrase-based models (Laver, Benoit, and Garry 2003). The literature also emphasizes use of individual words (Chung and Pennebaker 2013).

We provided details for two community-contributed commands. `wordfreq` processes content (online and local) and provides a word-frequency distribution. `wordcloud` draws a word cloud graph based on unique words and their frequencies. These two commands are provided as the first step in content analysis to be modified to fit individual researcher needs.

5 References

Chung, C. K., and J. W. Pennebaker. 2013. Counting little words in big data: The psychology of individuals, communities, culture, and history. In *Social Cognition and Communication*, ed. J. P. Forgas, J. László, and O. Vincze, 25–42. New York: Psychology Press.

Downe-Wamboldt, B. 1992. Content analysis: Method, applications, and issues. *Health Care for Women International* 13: 313–321.

Grimmer, J., and B. M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21: 267–297.

Hopkins, D. J., and G. King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54: 229–247.

Kassarjian, H. H. 1977. Content analysis in consumer research. *Journal of Consumer Research* 4: 8–18.

Kolbe, R. H., and M. S. Burnett. 1991. Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research* 18: 243–250.

Laver, M., K. Benoit, and J. Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97: 311–331.

Pennebaker, J. W., M. E. Francis, and R. J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC2001.* Mahwah, NJ: Lawrence Erlbaum.

Roberts, C. W. 1989. Other than counting words: A linguistic approach to content analysis. *Social Forces* 68: 147–177.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19: 321–325.

About the authors

Mehmet F. Dicle is an associate professor of finance at Loyola University New Orleans with a PhD in financial economics from the University of New Orleans, New Orleans, LA.

Betul Dicle is a research associate at Research and Teaching Associates with a PhD in political science from Louisiana State University, Baton Rouge, LA.