

text manipulations in stata

adam okulicz-kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Wednesday 7th April, 2021 14:23

outline

intuition

text analysis

one-on-one

- a reminder: let's meet one on one per your project !
- at least twice!

outline

intuition

text analysis

text as data

- text are just rich data
- we can quantify anything, e.g.:
 - feelings (survey data)
 - faces (image recognition)
 - text

setup

- we will begin with simple string functions in stata
 - like excel functions
- and then talk about regular expressions
- and we we will do some examples
- we will continue with text as data in Python

string functions

- string functions are not that complicated
 - help string functions
- and they are very useful
 - you can apply them in most data sets
- dofile

regular expressions

- did anybody heard of regular expressions ?
- http://en.wikipedia.org/wiki/Regular_expression
- they are used to match characters
- e.g. “*” matches any character
- they give you **very powerful** toolbox
- they are great for general, automated code to replace humans in pattern recognition
- they are similar in Python, but better
- we will do more under Python
- dofile

outline

intuition

text analysis

links

- https://www.stata.com/meeting/spain15/abstracts/materials/spain15_escobar.pdf
- <https://www.stata-journal.com/sjpdf.html?articlenum=dm0077>
- https://www.tcd.ie/Political_Science/wordscores/stata_manual/manual.html
- https://www.tcd.ie/Political_Science/wordscores/stata_manual/wordfreq.html
- <https://www.stata-journal.com/sjpdf.html?articlenum=dm0077>
- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2759033
- <http://casus.usal.es/blog/modesto-escobar/files/2013/01/Escobar2015.pdf>

simple word count

- these programs from the previous page
- may be overkill for simple search for number of occurrences
- simply use `https://www.google.com/search?ei=7x7LW8DlHI-xggfx6bjQCg&ins=&q=stata+dind+number+of+word+occurrences`
- and see dofile