text manipulations in stata

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Thursday 7th April, 2022 14:22

outline

intuition

text analysis

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 regular expressions
- we will continue with text as data in Python

string functions

- string functions are easy

  help string functions

- and they are very useful
  - need them in most data sets; sometimes they're really necessary
- 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
- **powerful** toolbox: replace humans in pattern recognition
- similar in Python, but better
- we will do more under Python
- dofile

outline

intuition

text analysis

can do little in stata, but py way better

- 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