splot

Data in general can be represented at many levels, from datum to statistic. Each level can be fun to look at.





lusi research.php for study material and data 'NSd guides resources.php for software and epts.ttu.



LIWC visualizer

E LIWC Visualizer displaying z-scored frequencies

Each LIWC category is made up of individual words, so it can be useful to see which words are contributing to the category value in a given text. This sort of thing can be done by referring to the dictionary, or by using the color-code feature in LIWC, but this visualizer offers a convenient way to see multiple categories (though it may not always line up with LIWC).

Each LIWC category is made up of individual words, so it can be useful to see which words are contributing to the category value in a given text. This sort of thing can be done by referring to the dictionary, or by using the color code feature in LIWC, but this visualizer offers a convenient way to see multiple categories though it may not always line up with LIWC. adverb, conj. cogproc, differ, funct

ID	WC	achieve	adj	adverb	affect	affiliatior
103	69	-0.64	-0.94	-0.4	-0.53	-0.63
•						۲
Menu		u	Pro	Clear		

www.depts.ttu.edu/liwc

lingmatch

LIWC output is a special case of a document-term matrix (DTM), with categories as terms.



Language Style Matching (LSM) uses inverse Canberra distance to measure stylistic similarity within this space:

$$1 - \frac{|7 - 5|}{7 + 5} = .83$$

This matrix, and the distance between its rows can be represented in many ways.

Cosine similarity between uncategorized rows is quite a bit lower:

$$\frac{\sum D1D2}{\sqrt{\sum D1^2 \sum D2^2}} = .59$$

Weighting by square root dampens the impact of term counts, which increases similarity:

	a	an	the		article
D1	1.41	1	2	\rightarrow	4.41
D2	1	1.73	1		3.73

miserman.github.io/lingmatch

Data Types:

lingmatch can be used with raw text, a DTM, or LIWC output.

Pre-Processing:

Inputs can be weighted by combinations of term and/or document frequencies, then dimensionally reduced by dictionary-based categorization (like LIWC) or semantic mapping (latent semantic analysis).

Comparisons:

Comparisons can be made between each row of a matrix (one to one), or each row can be compared with some standard (such as a group mean or prompt text; one to many).

Rows can also be compared between speakers and groups.

Metrics:

lingmatch offers a variety of similarity/distance metrics, including inverse Canberra distance for the standard form of LSM, and cosine similarity for a standard form of latent semantic similarity.