Status Internet Argument Data Set

Goal

Create a data set based on Internet Argument Corpus data which includes

- original post
- response
- topic
- mechanical turk annotations

link to source: https://nlds.soe.ucsc.edu/iac2

Current status

Data set with

- original post
- response
- topic
- 1 out of 3 mechanical turk data incorporated
  - yet to include other 2 mechanical turk data sources

See files:

- script: script/extractCombineTables.R
- output: ./data/output/quoteResponseMTurk.RData

Observations

Summary count:

topic	count	count_wth_mturk	mean_sarcasm_yes
	315557	108	0.18
communism vs capitalism	1222	56	0.07
abortion	54043	2167	0.14
evolution	76863	3970	0.14
death penalty	2955	86	0.10
School Uniforms	481	0	
climate change	2793	94	0.15
gay marriage	28657	918	0.15
minimum wage: pro or con	368	0	
gun control	46628	1909	0.16
marijuana legalization	922	80	0.12
obamacare	754	9	0.16
women in the military	302	0	
existence of God	7337	543	0.15
legalized prostitution	168	0	
immigration	215	0	
vegetarianism	244	0	
gays in the military	134	0	
socialized medicine	15	2	0.17
TOTAL	539658	9942	0.15

- Only a small sample of post-responses have been labeled by Mechanical Turk
- 10K vs 540K Based on 1 out of 3 Mechanical turk data

Proposed next steps

- Include other 2 mechanical turk data sources
- Create response variable that only includes the portion of the response that responds to the quoted text
    - had problems creating this variable due to encoding