Research Topic

I will work next week to refine our research topic, our hypotheses, and our research methodology while incorporating advice from the lecturer.

Torrenting Reddit

The pushshift project publishes web-scrapes of Reddit. I was able to torrent several relevant subreddits (takin data spanning from 2005 to 2023). I downloaded r/AmItheAsshole, r/AskDocs, and r/medical_advice, which took several hours. The files are in .zst format, which is a compressed format, but with the zstgrep tool it is possible to search the compressed files without decompressing them on disc. r/AmItheAsshole is nearly 17Gb compressed.

Summary: The Development and Psychometric Properties of LIWC2015

Author: James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn
Year: 2015

The LIWC2015 program processes texts and produces linguistic and textual measures, as well as proxy measures of various psychological and developmental attributes.

The proxy measures are essentially relative word counts. For instance the 'negative emotion' proxy is calculated by counting occurences of words associated with negative emotion.

The 'negative emotion' proxy is subdivided into anxiety, anger, and sadness. Words which contribute to any of these also contribute to the 'negative emotion' score of the text. Interestingly 'positive emotion' is not subdivided, meaning there is a richer analysis of negative emotion. I also saw in the paper "Recognition of Affect, Judgement, and Appreciation in Text" that negative affect was given a richer subcaterogization that positive affect.

The authors describe the process by which the LIWC2015 dictionaries and word-categorisations were developed. It involved a significant manual process with collaboration of several judges to create an initial dictionary. The dictionary was then analysed and ammended to fix ommissions or internal inconsistencies. The set of categories was chosen based on attributes commonly study in social, health, and personality psychology. The authors give an explanation as to why the internal consistency of linguistic-based psychometrics can be lower or have wider margins than a questionnaire (it comes down to the fact that people rarely repeat themselves in spontaneous natural language, whereas questionnaires make heavy use of repetition, or at least rhyming). Perhaps of interest to our group's analysis is the claim that Spearman-Brown prediction formula is generally are more accurate approximation of a word-category's internal consistency than is Cronbach's alpha.

Data Collection

Using the pushshift archives I have managed to download (by torrenting) the entirety of submissions and comments for several subreddits. The time to download the files was several days. The sizes of the raw datasets in compressed format are as follows.

```
$ du -h *
16G     AmItheAsshole_comments.zst
691M    AmItheAsshole_submissions.zst
502M    AskDocs_comments.zst
491M    AskDocs_submissions.zst
110M    medical_advice_comments.zst
108M    medical_advice_submissions.zst
```

Next week I will a write a program to generate summary statistics and distributions of the datasets, e.g. number of submissions, number of comments on a submission, number of submission authors, number of comment authors, submission sentence/word/character counts, comment sentence/word/character count.

Time permitting I would also like to generate summary statistics and distributions of the occurence of acronymns, NTA, YTA, NAH, ESH.

Finally, it may be beneficial to distribute these raw datasets to my group's members on hard media such as USB flash drives or SD cards.