Disability Trends NLP Project Progress

Noah Duggan Erickson Raghav Vivek Brady Deyak

04 June 2024

1 Noah's Progress

1.1 Accomplishments

- **Data Acquisition**: Current document count: 3300
- **Preprocessing**: Built basic script to convert rtf to standardized JSON. Experimented with `nltk` WordNet lemmatizer and Porter stemmer.
- **Word Embeddings**: Ducttaped `gensim` FastText model following an unreflected SciPy depreciation.
- **Data Vis & Analysis**: Modified TSNE embedding visualization code from A4, began looking into data more deeply.

1.2 Tasklist

- **Data Acquisition**: Continue to scrape and convert documents. Potentially increase corpora to 1000 docs/term/decade (currently 500) if Nexis ratelimits allow.
- **Preprocessing**: Continue to refine preprocessing pipeline, specifically lemmatization (and adjacent).
- **Data Vis & Analysis**: Continue refining TSNE embedding visualization, finish WordCloud visuals, continue deeper analysis, build connections between points (network-like?).
- **Deliverables**: Enhance plan for report, refine data vis for presentation.

1.3 Challenges

- **Preprocessing**: Lemmatization 2020 SoTA[1] neural seq2seq (pg. 3) insufficient. Considering falling back to regex. 2018 BRNN SoTA[2] untested.
- **Being a Student**: Learning in-situ. *So many manpages...*

[1] arXiv:2003.07082v2 [cs.CL] 23 Apr 2020

[2] arXiv:1808.03703v2 [cs.CL] 27 Aug 2018

2 Raghav's Progress

2.1 Accomplishments

2.2 Tasklist

2.3 Challenges

3 Brady's Progress

3.1 Accomplishments

3.2 Tasklist

3.3 Challenges