TITLE: Automated identification of RDoC construct from PubMed abstracts

TEAM: Suraj Subramanian (Pitt), William Schwarzmann (Pitt)

INTRODUCTION:

The Research Domain Criteria (RDoC) is a comprehensive framework developed by the National Institute of Mental Health (NIMH) for describing mental illness in multiple dimensions. This approach is a novel mechanism for mental health classification compared to one-dimensional methods (Sanislow et. al). One of the 2019 RDoC Task challenges is aimed at information retrieval and extraction. The competition provides a dataset of PubMed article abstracts annotated with RDoC constructs. The goal is to submit a ranked list of relevant articles to each RDoC construct. A given Average Precision evaluation metric is used to qualify each construct independently, and then averaged across constructs to compute a Mean Average Precision. Our goal is to develop an approach to this task that retrieves indexed abstracts for a given RDoC “search,” providing accurate results by the challenge criteria.

METHODS:

The RDoC 2019 Task challenge provides a corpus of 250 PubMed abstracts annotated with RDoC constructs. This is a very small dataset to train a model, so we obtained 350K abstracts from Medline. We chose this set based on the relevance of their MeSH headings to mental health topics.

Text preprocessing of the corpus is performed with Python code and the Natural Language Toolkit Python library (NLTK). To establish a baseline for performance, we minimally preprocess the documents. Stop words such as “a,” “why,” and “but” are then removed from the PubMed abstracts through the NLTK to reduce the data down to only relevant words. Composite words separated by hyphens and slashes are split into their constituents. We develop a simple abbreviation expander that extrapolates the full-form phrases of abbreviations from preceding words. All punctuations are discarded. Each document is represented as a single line.

Beyond these, we want to try the following preprocessing steps:

1. Using a POS tagger to keep only nouns, noun phrases, and adjectives. The algorithm powering Word2Vec is a Convolutional Neural Network, most commonly used in spatial (image) analysis. Unlike Recurrent NNs, CNNs cannot determine links between consequent words (Karpathy et. al). We believe most of the information that Word2Vec can capture about a term, is provided by the entities (nouns) in the term’s context.
2. Lemmatization might serve to reduce noise but we are unsure of the quantum of benefit. So far we have avoided it since it could have unpredictable results on medical jargon.

We use the gensim Word2Vec package to train a hyperspace on the 350K abstracts. Again, we use this corpus to establish a baseline. We suspect the hyperspace thus obtained would be sorely specific to the research it has trained on, and risks learning and overfitting esoteric representations of concepts. To provide a more general representation of concepts, we intend to augment out hyperspace with definitions from medical ontologies (like OntoBee).

RESULTS:

Firstly, to check if this approach is promising, we test it on the 250 annotated abstracts provided on the Task page. Definitions for each RDOC construct are obtained from the NIMH webpage (<https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/definitions-of-the-rdoc-domains-and-constructs.shtml>). A vector representation for the constructs is obtained by summing the vectors for each token in the preprocessed definition. Similarly, document vectors are obtained by summing up the token-vectors in each of the 250 abstracts.

We compare the relevance to 8 constructs (queries) used in the annotation.

The AP and MAP for 4 models are obtained.

**Unweighted** uses no weighting scheme to sum token-vectors

**Tfidf** uses tf-idf to weight token vectors before summing

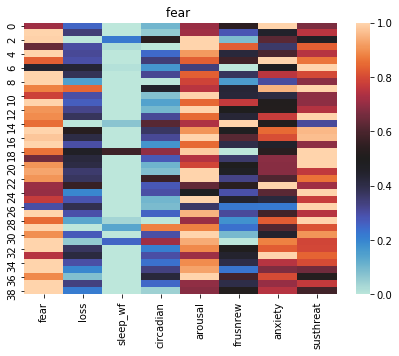
**Tfidf-posfilt** strips out verbs, prepositions, adverbs and modals.

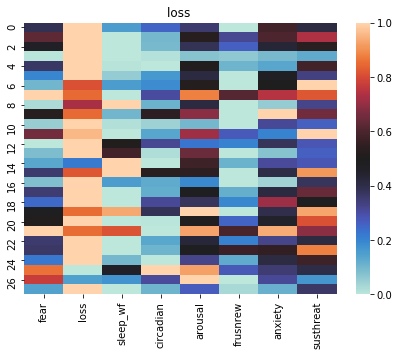
**Sg\_w10\_tfidf\_pos** increases the window size from 5 (used in all the previous models) to 10

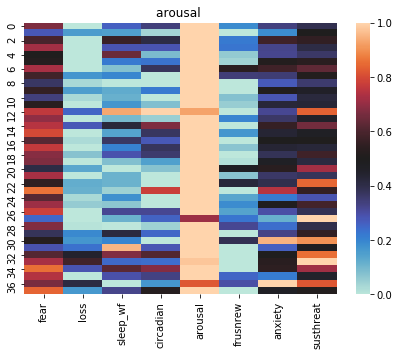
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **unweighted** | **tfidf** | **tfidf\_posfilt** | **sg\_w10\_tfidf\_pos** |
| **fear** | 0.331225 | 0.394218 | 0.408762 | 0.383806 |
| **loss** | 0.267024 | 0.742487 | 0.780494 | 0.840315 |
| **arousal** | 0.313597 | 0.579911 | 0.537592 | 0.506523 |
| **circadian** | 0.439948 | 0.278955 | 0.284175 | 0.406311 |
| **frusnrew** | 0.061585 | 0.071679 | 0.07701 | 0.079673 |
| **anxiety** | 0.356331 | 0.316097 | 0.341975 | 0.434728 |
| **sleep\_wf** | 0.36036 | 0.520229 | 0.563937 | 0.415623 |
| **susthreat** | 0.232491 | 0.213766 | 0.226111 | 0.234838 |
| **MAP** | **0.29532** | **0.389668** | **0.402507** | **0.412727** |

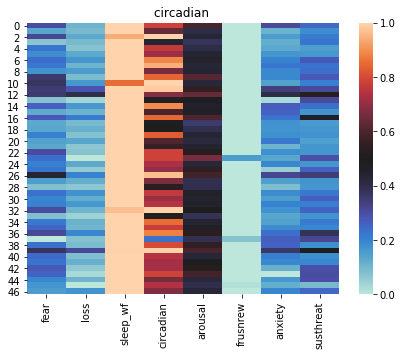
These numbers are much lesser than what the competition winners have obtained (the winner has an MAP of 0.86). However, those results were computed on a separate test set containing many more documents that we do not have access to.

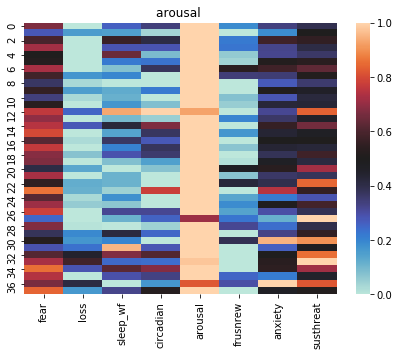
To get a better sense of what the model is doing, we provide the true annotation and similarity heatmaps of the best model. Given a document, we obtain the similarities with each RDoC. These similarities are independent of each other, so we scale these values between 0 to 1 (using MinMax scaling) to obtain a ranked similarity. We then generate heatmaps based on this scaled relevance to see which RDoC constructs are most often confounded with each other.

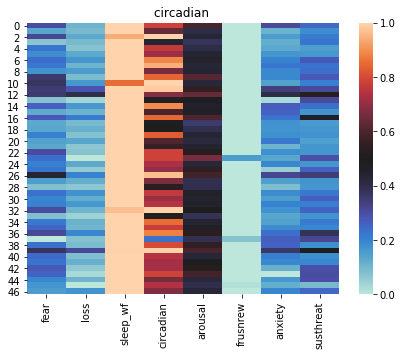


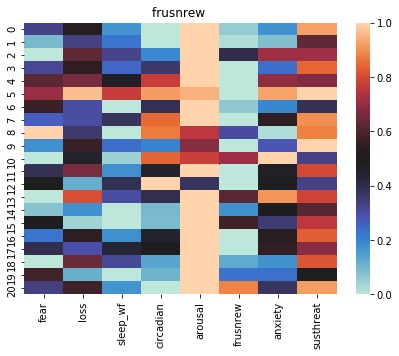


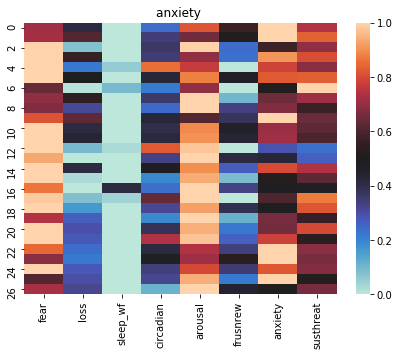


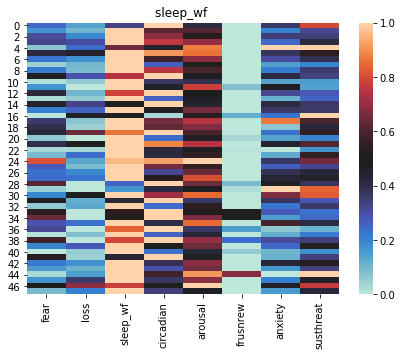


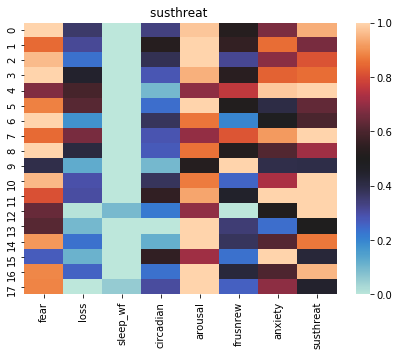












LIMITATIONS & NEXT STEPS:

While the system we have right now is not a very robust IR system, this general approach seems to show some promise.

A colloquialism in machine learning is “Garbage In, Garbage Out”. While our input data is far from garbage, for our model to achieve desired performance, it is pertinent to craft this input to allow our model to focus on what is important. Medical literature is different from free language corpora in phrasing and jargon. This can trip up statistical models that are hitherto used to free text in more general contexts. For example, the Stanford POS tagger sometimes tags ‘tryptophan’ as an adjective instead of a noun.

Yet another challenge is the ambiguity of certain phrases; the intended meaning is identified from the context, and even then requires a certain degree of subject matter expertise. Such aspects of the corpora restrict a naive and simplistic application of ML algorithms, and require some craftiness.

We are using the RDoC constructs as our queries. The NIMH has built the RDoC as a multidimensional representation of certain mental health issues. However in our current system, we have only represented them based on the textual definitions provided. This might have inadequate specificity, leading to conflation and retrieval of ‘non-relevant’ documents. For example, the definition of ‘frustrated nonreward’ contains terms that figure highly in ‘sustained threat’ and ‘arousal’. Work needs to be done in clamping down on very specific and exclusive definitions to obtain higher precision scores for such definitions.

We have identified the next steps that we believe will improve performance:

1. Only keep nouns and noun-phrases, to allow the model to focus on the co-occurring entities.
2. Tweaking the window size of Word2Vec. Currently we are using 5 (the default).
   1. A larger window allows for longer turns of phrases that appear to be common in medical literature abstracts. Mikolov et al. use the entire sentence for context.
   2. A smaller window restrict contexts to immediate words, allowing for tighter representations, good for analogies.
3. Augment hyperspace-training corpus with definitions from medical ontologies. Right now, the model is learning patterns of contexts only from research abstracts. On average, each abstract is a few sentences, and possibly does not have enough statistically significant patterns to build tight representations of context. Learning a wikipedia-like reference before training on the abstracts would allow stronger linkages among concepts.
4. Increase the number of training iterations (epochs). Since the corpora consists of short abstracts, it would help the model to go over it a higher number of times.
5. Use Doc2Vec to represent documents and definitions
6. Lemmatization, depluralization to further eliminate noise

WORKS CITED:

Sanislow CA, Ferrante M, Pacheco J, Rudorfer MV, Morris SE. Advancing Translational Research Using NIMH Research Domain Criteria and Computational Methods. Neuron. 2019 Mar 6;101(5):779-782. doi: 10.1016/j.neuron.2019.02.024. PubMed PMID: 30844398.

Karpathy, Andrej, Joulin A, Fei-Fei L. “Deep Fragment Embeddings for Bidirectional Image Sentence Mapping.” *ArXiv*, 22 June 2014, arXiv:1406.5679v1 [cs.CV].

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Schütze, H., Manning, C. D., & Raghavan, P. (2008, June). Introduction to information retrieval. In Proceedings of the international communication of association for computing machinery conference (p. 260).