Linguistic Analysis of the bioRxiv Preprint Landscape

This manuscript (permalink) was automatically generated from greenelab/annorxiver_manuscript@28cc311 on May 24, 2020.

Authors

- **David N. Nicholson**
  0000-0003-0002-5761 · danich1
  Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania · Funded by GBMF4552; T32 HG00046

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Introduction

1. What is a preprint

2. Why are preprints important?

3. Mention how preprints are being integrated into scientist’s everyday workflow

4. Talk about biorxiv and discuss how it is one of the repositories that maintains preprints along with citation of others such as arxiv/medrxiv etc.

5. Discuss works that analyze biorxiv from an audience perspective (quantifying tweets etc.)

6. Mention the gap which consists of analysing the language of biorxiv preprints (first to do this)

7. ^ Why is this important? What will this allow for future research projects?

8. Provide list of contributions within this manuscript

Methods

Datasets

bioRxiv

1. Describe how bioRxiv was obtained
2. Describe metadata statistics on bioRxiv (number of preprints, number of preprints with multiple versions)

PubMed Central

1. Describe how PubMed central was obtained
2. Describe metadata statistics on PubMed central (number of articles, how many articles were processed

Comparing Corpora

1. Spacy to process text via - Lemmatization, removal of stop words
2. Describe counting frequencies of each lemma
3. Describe using chi-square test
4. Describe how to calculate the likelihood and log odds ratio

Visualizing the Preprint Landscape

Generate Document Representation

1. Describe how word2vec works
2. Talk about training word2vec on entire biorxiv repository
3. Discuss how to generate a document representation using word2vec model

Dimensionality Reduction of Document Embeddings

1. Explain how tSNE works (paragraph one)
2. Explain how PCA works (paragraph two)
3. Discuss how words were mapped onto PC components via cosine similarity
4. ^ Explain cosine similarity

Recommending Journals/ bioRxiv Audience Analysis

1. This title will update as analysis is completed
2. This section will describe how the above process is conducted

Results

Comparing bioRxiv to PubMed Central

Global View

1. Create a treemap visualization of top X terms that are different between bioRxiv and PubMed Central (based on odds ratio)

Published Preprint Differences

1. Create a treemap visualization of top X terms that are different between Preprint and Published documents (based on odds ratio)

The bioRxiv Preprint Landscape

1. Provide the tSNE figure of the bioRxiv
2. Discuss the results of the tSNE figure and highlight that there are category clusters within the figure

Topic Analysis of Principal Components

1. Provide an example of the word cloud for principal components
2. Show plot of the principal components and the scatterplot
3. Mention that the word clouds can be found at xyz

Journal Recommendations/Audience Associations

1. Title will change once analysis is finished
2. Provide key figure for this section and take-home message

Discussion

Conclusion

Acknowledgements

References