Using Natural Language Processing to Analyze the DSM-5

Syeda Zainab Aqdas Rizvi ‘18, Zhu Shen ‘19

SDS 400: NLP Special Studies

Professor Jordan Crouser

Introduction to NLP in Python

The corpus for our project is the text from the Diagnostic and Statistical Manual of Mental Disorders, 5th edition: DSM-5. Specifically, we will be looking at bias in the paragraphs titled ‘Culture-Related Diagnostic Issues’ that correspond to diagnoses in the DSM henceforth referred to as documents. This project uses NLTK, TextBlob and scikit-learn libraries in Python. Before analyzing each document, we removed stop words, tokenized and stemmed it followed by transforming the corpus in vector-space using term frequency-inverse document frequency. We then calculated the cosine distance between each document to serve as a measure of similarity.

Cluster Analysis of Culture-Related Diagnostic Issues

We ran the K-means algorithm on the culture-related diagnostic text in each of the diagnoses in the DSM. K-means operates with a predetermined number of cluster and our estimate for the optimum number of clusters is 8 which we achieve through silhouette analysis.

We sorted the clusters to identify the top words that are nearest to the cluster centroids and they give a sense of the topics in that cluster.

The clusters show some clear patterns such as that almost 60% of the variance in clusters can be explained by the chapter the diagnosis corresponds to. We also explored the use of passive voice at the sentence level

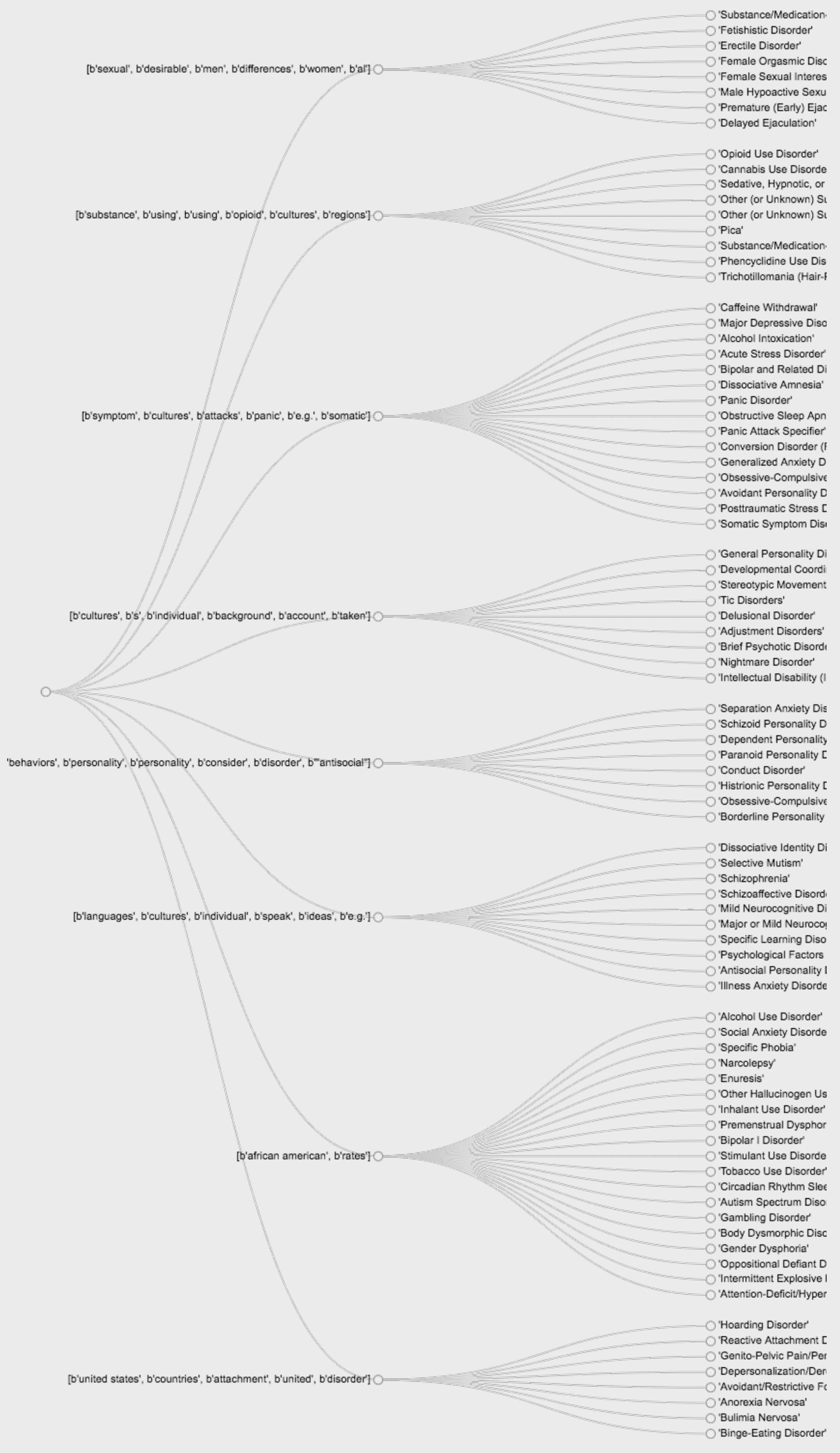
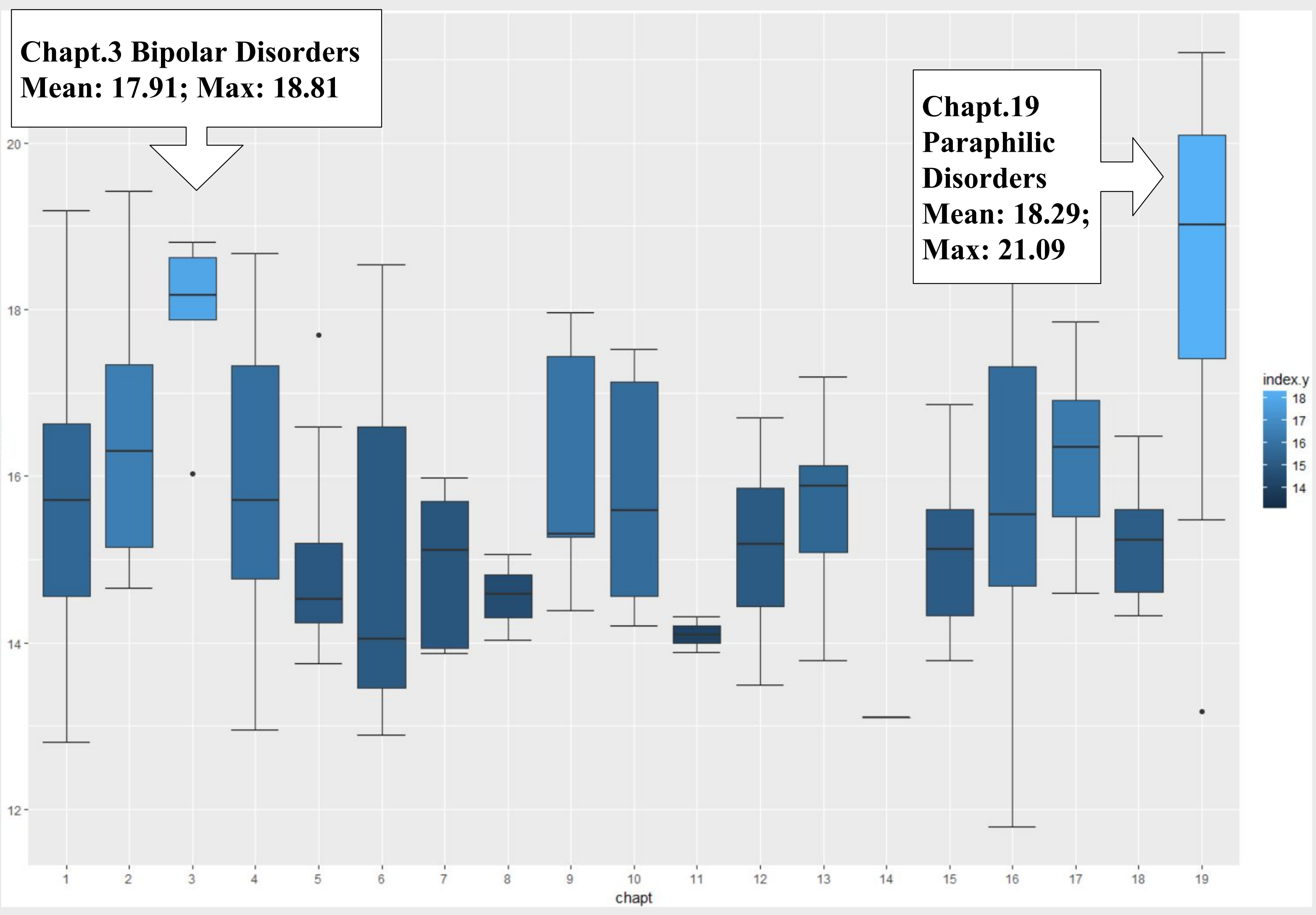


Figure 1: Dendrogram showing the distribution of the 87 diagnoses in 8 clusters.

Smog Index and Readability Test

Another source of bias comes from the understanding gap between social workers, who actually use DSM, and psychiatrists: the professionals who write this manual. Specifically, we will test on the difficulty level of language used in diagnostic and associated features for each diagnosis, and determine whether the level relates to the likelihood of misdiagnosis. Non-character notations, reference and disorders' names were removed before computing the difficulty level for each sentence in the diagnosis. Unusually difficult sentences for each diagnosis were evaluated and removed if they are irrelevant of making diagnosis. Then, we calculated the Smog Index, a measure of text readability for each of the 158 diagnosis.


Chapt.3 Bipolar Disorders
Mean: 17.91; Max: 18.81

Chapt.19 Paraphilic Disorders
Mean: 18.29; Max: 21.09

Figure 2: Boxplot showing the Smog Index of 19 chapters in DSM-5.

Chapter 19 (Paraphilic Disorders) and Chapter 3 (Bipolar and Related Disorder) have significantly higher Smog index than other chapters. Such result somehow explains the high misdiagnosis possibility of Bipolar Disorder. However, some other commonly misdiagnosed disorders, such as Borderline Personality Disorder and Schizophrenia, have Smog indexes which are below the average index of all diagnoses in DSM.

Such discrepancy suggests two possible reasons for misdiagnosis: (1) diagnostic feature is too difficult to understand; (2) diagnostic feature is relatively simple but fails to contain enough information to distinguish.

References

1. “A Comparative Study of TF*IDF, LSI and Multi-Words for Text Classification.” *Expert Systems with Applications*, Pergamon, 6 Sept. 2010.
2. “A Readability Assessment of Online Parkinson’s Disease Information.” *J. R Coll Physicians Edinb.*, 2010.

Acknowledgements

This project was completed in partial fulfillment of the requirements of SDS 400: Special Studies. This course was taught by R. Jordan Crouser during Fall 2017.