Lab 2 - Linguistics Data, Stat 215A, Fall 2019

Your name

September 23, 2019

Please use this structure for your report, but you do not have to slavishly follow this template. All bullet points are merely suggestions and potential points to discuss in your writeup. Your report should be no more than 12 pages, including figures. Do not include *any* code or code output in your report. Indicate your informal collaborators on the assignment, if you had any.

1 Introduction

- Describe the problem of interest and put your analysis in the domain context. Read the introduction of the two Nerbonne and Kertzschmar papers for some help here.
- What do you aim to learn from this data?
- Outline what you will be doing in the rest of the report/analysis

2 The Data

- What is the data that you will be looking at?
- Provide a brief overview of the data
- How is this data relevant to the problem of interest? In other words, make the link between the data and the domain problem

2.1 Data Cleaning

- This dataset isn't as bad as the redwood data, but there are still some issues. You should discuss them here and describe your strategies for dealing with them.
- Remember to record your preprocessing steps and to be transparent!

2.2 Exploratory Data Analysis

- This is where you compare pairs of questions with discussion and plots.

3 Dimension reduction methods

- This is where you discuss and show plots about the results of whatever dimension reduction techniques you tried - PCA, variants of PCA, t-SNE, NMF, random projections, etc.
- What do you learn from your dimension reduction outputs
- Discuss centering and scaling decisions

4 Clustering

- This is where you discuss and show plots about the results of whatever clustering methods your tried - k-means, hierarchical clustering, NMF, etc.

5 Stability of findings to perturbation

- What happens to your clusters when you perturb the data set?
- What happens when you re-run the algorithm with different starting points?

6 Conclusion

- What are the main takeaways from your exploration/clustering/stability analysis?