Understanding Presidential Speeches and Executive Orders with Natural Language Processing

Computational Content Analysis

Lily Grier
lilygrier@uchicago.edu
The University of Chicago

Linh Dinh
ldinh@uchicago.edu
The University of Chicago

Roberto Barroso-Luque
barrosoluquer@uchicago.edu
The University of Chicago

Abstract

BLABLABLA

Preprint. Under review.

Table 1: Number of principal component vectors and ratio of variance explained.

Dataset	PCA Vectors Used	Variance Explained
Heart Disease(continuous only)	1	.25
Heart Disease	2	.41
Divorce	1	.31
NBA Rookies	2	.55
Mushrooms	1	.34

1 Introduction

BLABLABLA

2 Methodology

2.1 Datasets

BLABLABLA

2.2 Data Pre-processing

BLABLABLA

2.3 Other methods

BLABLABLA

2.4 Even more methods

BLABLABLA

3 Results

3.1 Some subsection

BLABLABLA

3.2 Another subsection

BLABLABLA

3.3 Conclusion

BLABLABLA

4 References

[1] Chipman H. & Gu H. (2006): Interpretable dimension reduction. *Journal of Applied Statistics*

[2] Ding J. , Condon A. & P.Shah S. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications*

[3]Hosseini B. & Hammer B (2019): Interpretable Discriminative Dimensionality Reduction and Feature Selection on the Manifold. *BiorXiv*

Table 2: Original features with highest correlation to chosen principal component vectors.

Dataset	Original feature (label)	Correlation to PC vector(R2)
Heart Disease(continuous only)	Thalach (max heart rate)	.89
Heart Disease	Age	.69
Divorce	Atr7	.45
NBA Rookies	Offensive Rebounds	.80
Mushrooms	Stalk Shape-enlarging and shape-tapering	.58

Table 3: Original features with highest correlation to chosen "interpretable direction" vectors

Dataset	Original feature (label)	Correlation to ID vectors(R2)
Heart Disease(continuous only)	Thalach (max heart rate)	.79
Heart Disease	Oldpeak	.58
NBA Rookies	Offensive Rebounds	.68
Mushrooms	Odor-musty, ring-type-none, ring-number-none	.88

Table 4: Prediction accuracy using PCs, IDs, and original feature space

Dataset	Full features space (Least Squares)	PCA Least Squares	ID Least Squares
Heart Disease(continuous only)	.72	.70	.40
Heart Disease*	.59	.77	.70
Divorce	1.00	.50	NA
NBA Rookies	.67	.56	.41
Mushrooms	.48	.47	.47

Note: For all classification tasks the feature matrix was broken into 80% training set and 20% testing set. Accuracy shown in the table is based on the training set. *These errors are not what we expected. See findings section.