USING NLP AND UNSUPERVISED LEARNING TO CLUSTER TWEETS FROM MEMBERS OF THE 116TH CONGRESS

Jennifer Brown
General Assembly
May 2021

TABLE OF CONTENTS

INTRODUCTION

- Problem Statement
  - The United States of America has two major political parties: Republican and Democrat with a few Independents here and there.
  - But, while Congress Persons typically affiliate with one of these two political parties, can they be grouped differently based on the language used in throughout their body of tweets?
- Audience
- Importance

	116th Congress		
	Democrat	**Independent**	**Republican**
Senators	47	2	50
Representatives	196	0	236
Total:	**243**	**2**	**286**

THE DATA: COLLECTION & CLEANING

- Data Sources:
  - Tweets of Congress
    - A GitHub repository collecting daily tweets for all members of Congress and affiliates.
  - Twitter Handles
  - Wikipedia
- Cleaning
  - Keep relevant Twitter accounts
  - Remove duplicates
  - Keep relevant columns

THE DATA: PREPROCESSING

- Remove:
  - Hyperlinks, emojis, symbols, characters
- Remove stopwords and contractions
- Keep important parts of speech
- Lemmatization

MODELING

- Unsupervised Machine Learning Cluster Algorithms
  - K-Means
  - DBSCAN
- Text Vectorization
  - spaCy word embeddings
  - CVEC or Count Vectorizer
  - TF-IDF or term frequency–inverse document frequency

MODELING

	Model 1	Model 2	Model 3
	1) SpaCy word vectors 2) TSNE	1) CVEC 2) Scaled 3) PCA 4) TSNE	1) TF-IDF 2) Scaled 3) TSNE
Silhouette Score	0.529	0.547	0.492
Inertia Value	49197.659	50542.696	43583.468
k (# of clusters)	2	2	2

MODELING

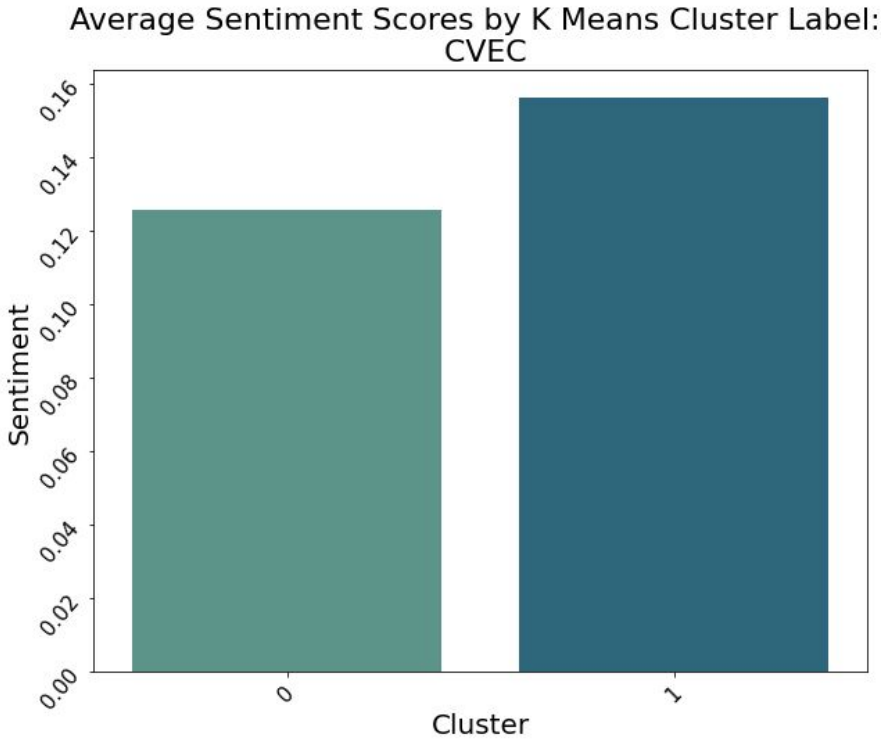
	Model 1	Model 2	Model 3
	1) SpaCy word vectors 2) TSNE	1) CVEC 2) Scaled 3) PCA 4) TSNE	1) TF-IDF 2) Scaled 3) TSNE
Silhouette Score	0.529	0.547	0.492
Inertia Value	49197.659	50542.696	43583.468
k (# of clusters)	2	2	2

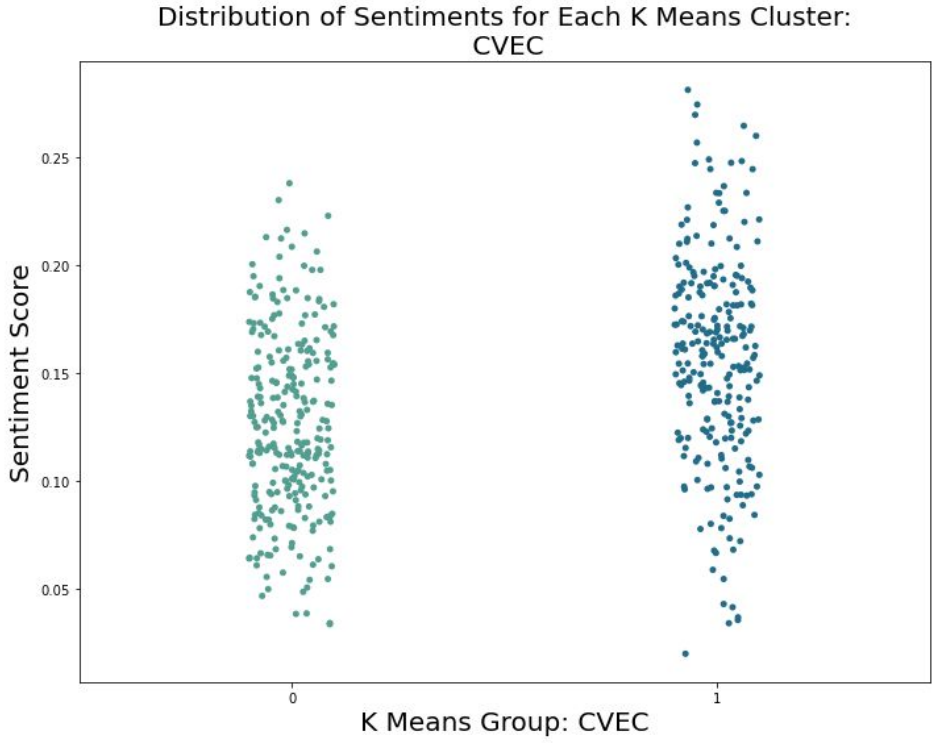
Cluster 0: 278 documents
Cluster 1: 243 documents

RESULTS

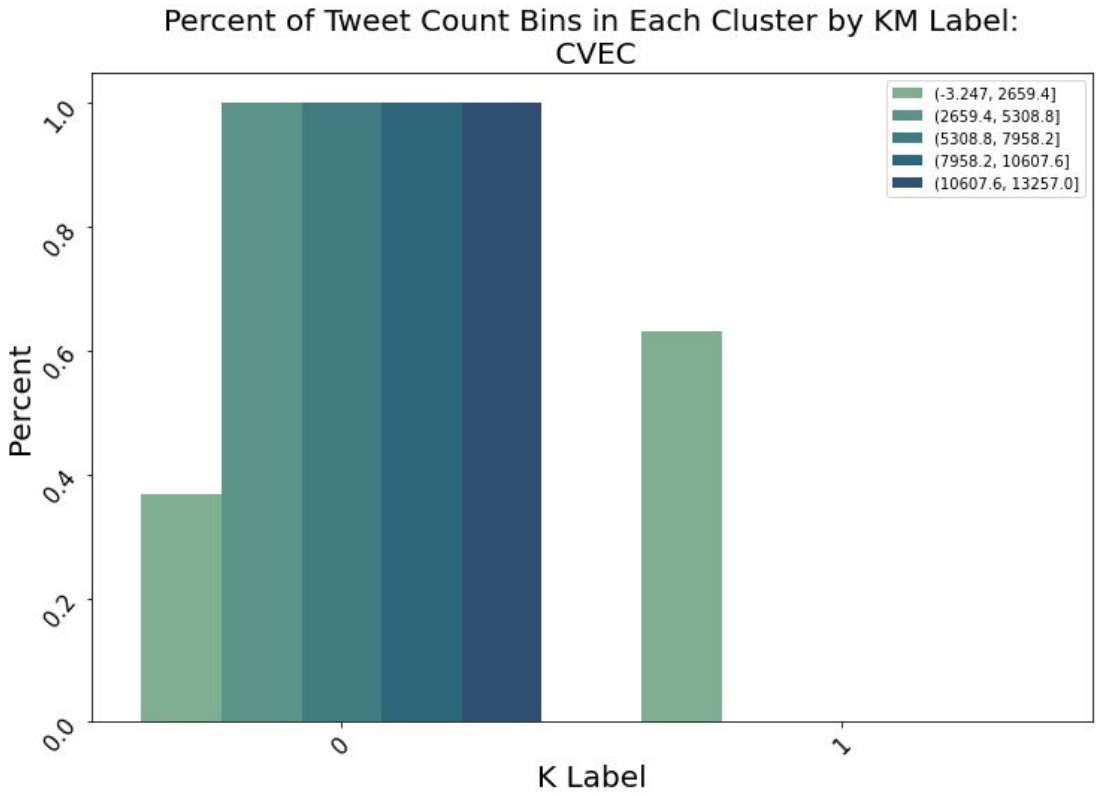
Most Frequent Words					
Complete Corpus		**Cluster 0**		**Cluster 1**	
today	136868	today	105011	today	31857
work	128641	work	101290	work	27351
need	115104	need	95138	thank	23990
trump	100434	trump	88466	help	21055
help	100053	help	78998	need	19966
people	92826	people	76404	great	18678
act	92378	act	75588	year	17610
president	91200	president	74882	house	17381
thank	89566	health	71351	time	17161
community	88413	community	71267	community	17146

RESULTS

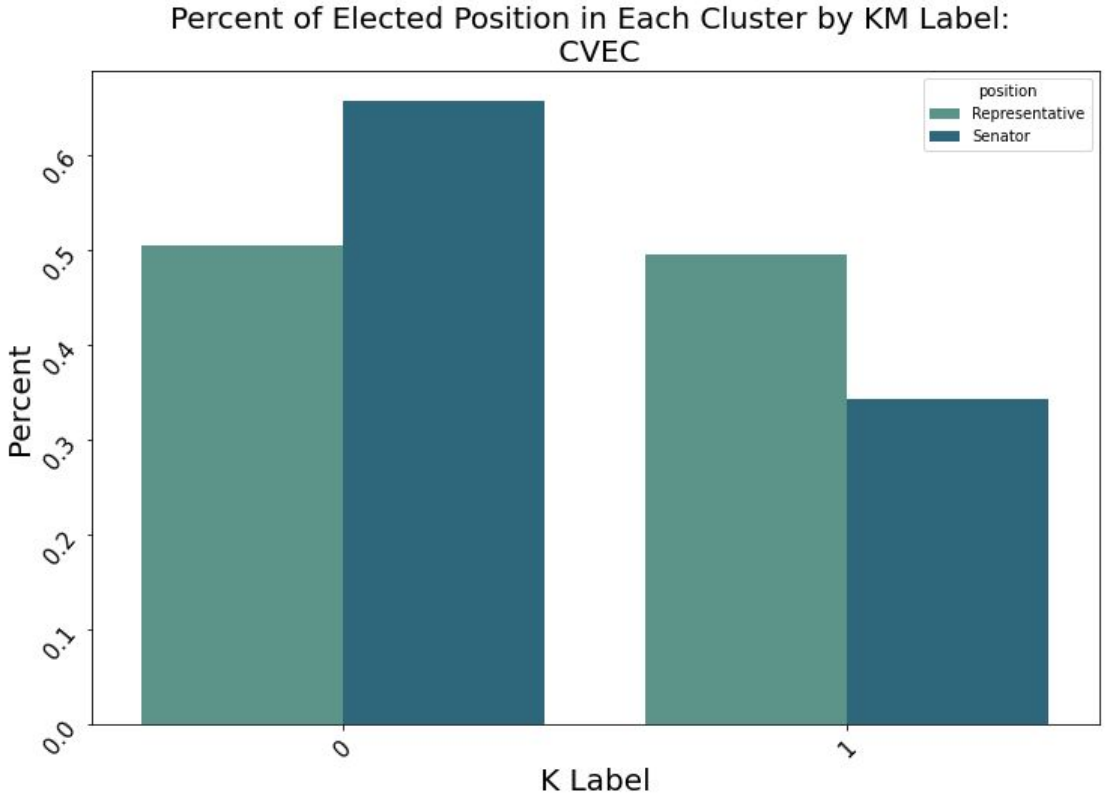

Average Sentiment Scores by K Means Cluster Label: CVEC
Sentiment
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0
1
Cluster


Distribution of Sentiments for Each K Means Cluster: CVEC
Sentiment Score
0.05
0.10
0.15
0.20
0.25
0
1
K Means Group: CVEC

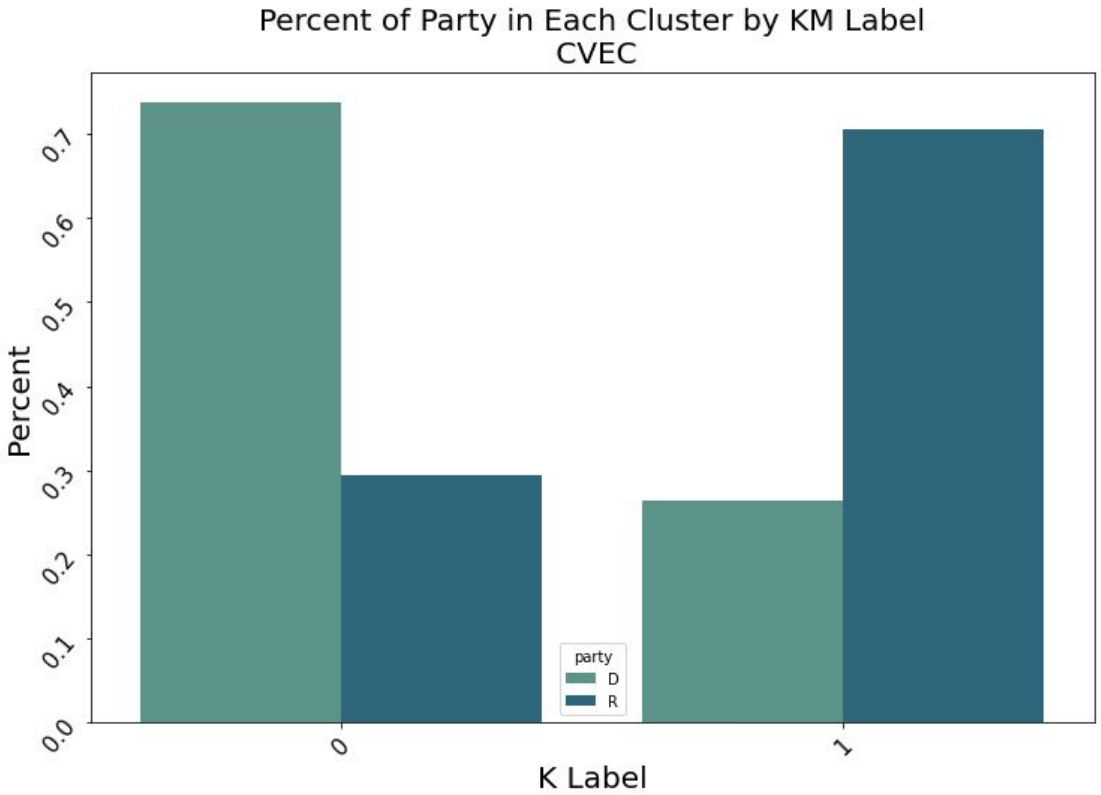
RESULTS


Percent of Tweet Count Bins in Each Cluster by KM Label:
CVEC
(-3.247, 2659.4]
(2659.4, 5308.8]
(5308.8, 7958.2]
(7958.2, 10607.6]
(10607.6, 13257.0]
Percent
0.0
0.2
0.4
0.6
0.8
1.0
0
1
K Label

RESULTS


Percent of Elected Position in Each Cluster by KM Label:
CVEC
position
Representative
Senator
Percent
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
K Label

RESULTS


Percent of Party in Each Cluster by KM Label
CVEC
Percent
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
party
D
R
0
1
K Label

RESULTS

	senduckworth
RepPressley	0.999999
NydiaVelazquez	0.999998
SenFeinstein	0.999990
RepChuyGarcia	0.999986
LeaderHoyer	0.999931

	senduckworth
RepTrey	-1.000000
RepRussFulcher	-0.999998
RepPeteStauber	-0.999996
RepMikeTurner	-0.999995
SenJohnThune	-0.999989

- Find Similarity Values!
  - Cosine Similarity
  - -1 to 1

CONCLUSIONS

- Two distinct clusters found using K-Means
  - Cluster 0 tended to write more tweets, had more Senators and more Democrats
- Mostly neutral sentiment with a very slight positive arc
- Cosine Similarity to find similar or dissimilar Congress Persons

RECOMMENDATIONS

- Collect more official Senator and Representative Twitter data
  - Replicate on Twitter data for time periods during Barack Obama's presidency
- Explore other dimensionality reduction measures
- Try other word vectorization packages
  - Gensim
  - GloVe
- Look at voting data for Senators and Reps
  - How would this look in comparison to tweet clusters?

THANKS!

What questions do you have?