Document Similarities of National Constitutions

Introduction: Hypothesis

- Systems of governments do not develop in isolation; countries borrow concepts, foundational laws, and institutional blueprints from other countries. We wanted to see which countries influenced one another in terms of the basic government structures
- Our hypothesis was that countries that were geographically close, and those that came from similar colonial backgrounds would have more similar systems of government compared to a random pairing of countries

Introduction: Data Used

- Looking at a country's constitution is a good way to get a high level picture of the government system. Thus, we decided to compare the constitutions of 192 countries to see how similar they were to one anothe
- Once the constructions were cleaned and normalized, we could compare all of them in R Studio. It was our goal to develop groups of constitution that were similar. Based on the countries in these groupings, we can see our hypothesis was correct
- We collected html copies of each constitution from the Constitute Projec www.constituteproject.org

LINEAR ALGEBRA CONCEPTS

1. Term Frequency and Inverse Document Frequency (TF-IDF)
2. Latent Semantic Analysis - Singular Value Decomposition
3. Cosine similarity

Data Preprocessing Pipeline

```
constitue_df[countries=='united_states_of_america',]
```

```
185  Try a new topic or search term. We the People of the United States, in Order to form a more perfect
Union, establish Justice, insure domestic Tranquility, provide for the common defense, promote the genera
```

Tokenize texts, lower case the tokens, remove stopwords, and perform stemming on the tokens

```
 [1] "tri"         "new"         "topic"       "search"      "term"        "peopl"
 [7] "unit"        "state"       "order"       "form"        "perfect"     "union"
[13] "establish"   "justic"      "insur"       "domest"      "tranquil"    "provid"
```

Feature Engineering

Add bigrams to our feature matrix

```
library(quanteda)
corpus_tokens <- tokens_ngrams(corpus_tokens, n = 1:2)
```

```
[1] "constitut_met"      "met_pursuant"      "pursuant_actual"     "actual_guarante"
[5] "guarante_thereof"   "thereof_mechan"    "mechan_state"        "state_articl"
[9] "articl_shall"       "shall_put"         "put_practic"         "practic_three"
```

Term Frequency and Inverse Document Frequency (TF-IDF)

```
library(quanteda)
corpus_tokens.dfm.tfidf <- dfm_tfidf(corpus_tokens.dfm)
```

$$TF\text{-}IDF(t,d) = TF(t,d) * IDF(t)$$

$$TF(t,d) = \frac{freq(t,d)}{\sum_i^n freq(t_i,d)}$$

$$IDF(t) = \log\left(\frac{N}{count(t)}\right)$$

Terms	Word Count	TF-IDF
unit_state	88.00	103.10
congress	62.00	44.48
senat	50.00	22.65
Total count	7,855.00	

Countries	tri	new	topic	search	term
afganistan	0	0	0	0	0
albania	0	0	0	0	0
algeria	0	0	0	0	0
andorra	0	0	0	0	0
angola	0	0	0	0	0

Latent Semantic Analysis - SVD

Center the data

```
corpus_tokens.tfidf.colmean <- apply(corpus_tokens.tfidf,2,mean)
corpus_tokens.tfidf.centered <- corpus_tokens.tfidf - corpus_tokens.tfidf.colmean
```

Use model explained variance to determine number of dimensions

```
pca <- prcomp(corpus_tokens.tfidf.centered)
```

	PC73	PC74	PC75
Standard deviation	23.93406	23.49189	23.2898
Proportion of Variance	0.00232	0.00224	0.0022
Cumulative Proportion	0.89593	0.89817	0.9004

Latent Semantic Analysis - SVD

$$SVD\ of\ X = X = U\sum V^T$$

```
library(irlba)
corpus_irlba <- irlba(t(corpus_tokens.tfidf.centered), nv = 75, maxit = 600)
```

col = 75

row = 392,924 Terms

U	X1	X2	X3	X4	X5
1	8.80E-05	0.00018445	5.187E-05	-5.75E-05	-1.46E-04
2	1.35E-04	0.00023343	0.00028512	-4.37E-05	-4.83E-05
3	1.22E-04	0.00038296	0.00029662	-6.38E-04	-7.48E-05
4	2.61E-04	0.00039637	0.00021921	-1.45E-04	-1.63E-04
5	2.35E-04	0.00039232	0.00016084	-2.50E-05	-1.57E-04

row = 75

col =75

∑	X1	X2	X3	X4	X5
X1	2399.00	0.00	0.00	0.00	0.00
X2	0.00	1766.00	0.00	0.00	0.00
X3	0.00	0.00	1616.00	0.00	0.00
X4	0.00	0.00	0.00	1577.00	0.00
X5	0.00	0.00	0.00	0.00	1508.00

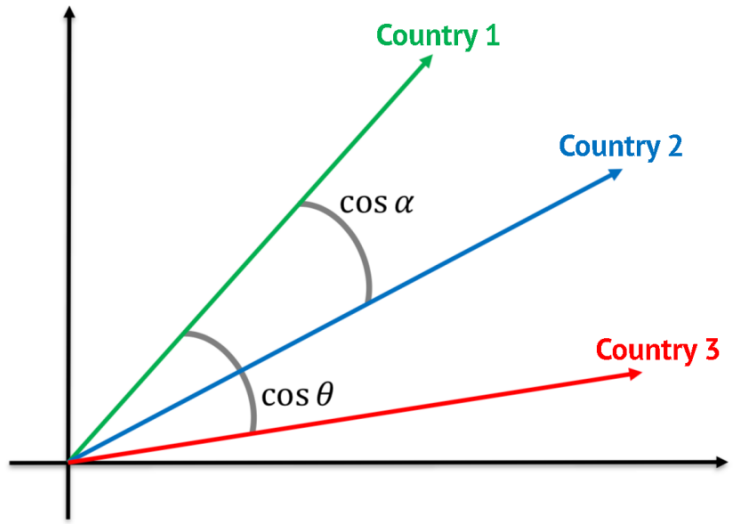
row = 75

row = 193 Countries

V^T	X1	X2	X3	X4	X5
1	-7.26E-05	-0.00091516	-4.89E-05	3.51E-05	0.0001394
2	-1.37E-03	-0.00094177	-8.69E-04	1.10E-03	0.00184677
3	-3.17E-04	-0.00059831	-2.96E-04	8.24E-04	0.00087813
4	1.07E-04	-0.00018017	-8.32E-05	1.84E-04	0.00071524
5	-9.36E-04	-0.00071014	-6.23E-04	1.18E-03	0.00110722

Cosine similarity

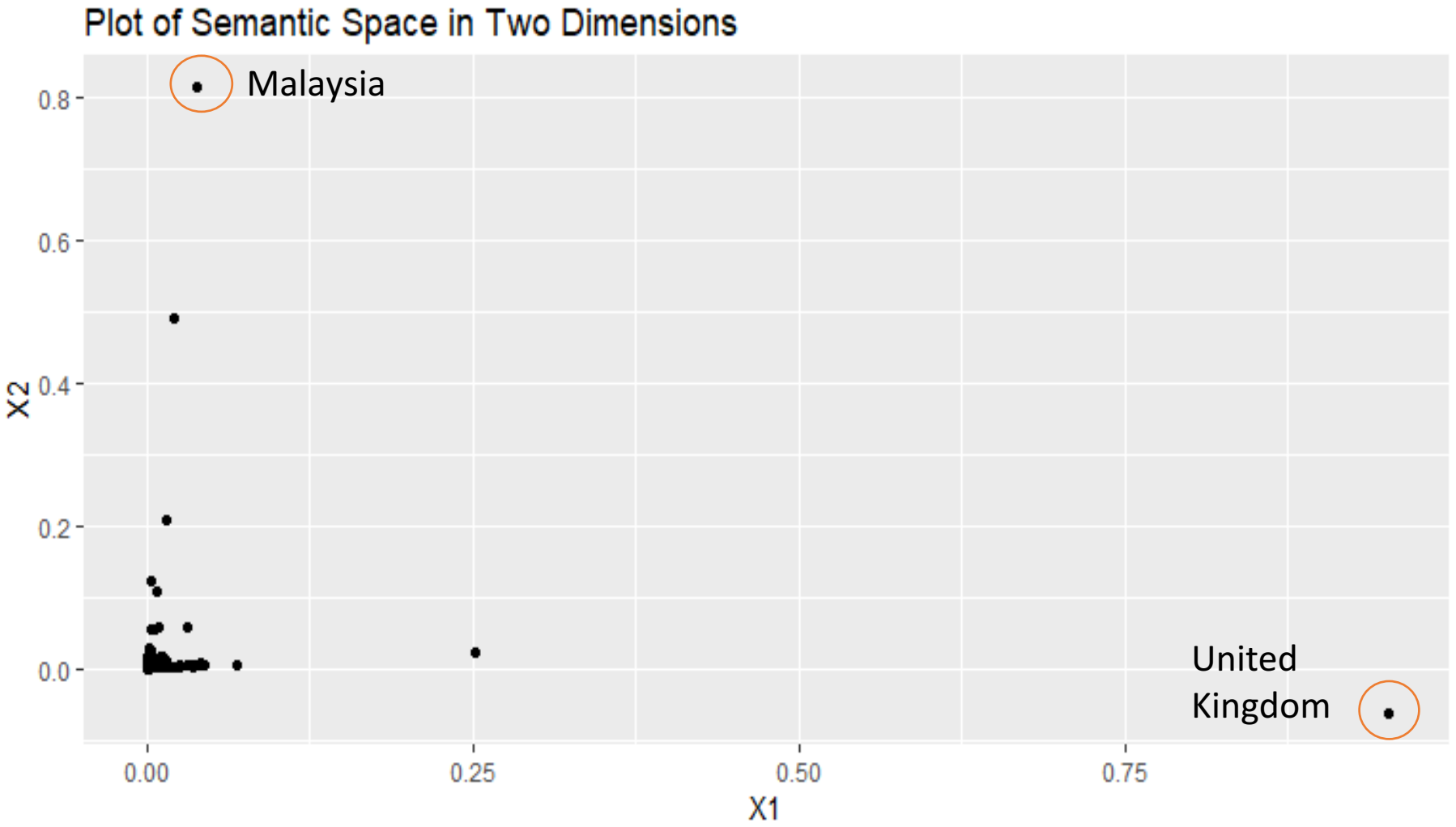
$$\cos\theta = \frac{\mathrm{A} \cdot \mathrm{B}}{\|\mathrm{A}\|_2 \|\mathrm{B}\|_2}$$




```
corpus_similarities <- cosine(t(as.matrix(corpus_svd[,-1])))
```

1	1.0000000	0.6331318	0.92751377	0.31741539	0.6556371
2	0.6331318	1.0000000	0.87656756	-0.53304983	0.9995669
3	0.9275138	0.8765676	1.00000000	-0.06005194	0.8903516
4	0.3174154	-0.5330498	-0.06005194	1.00000000	-0.5079192
5	0.6556371	0.9995669	0.89035162	-0.50791924	1.0000000

VISUALIZATION OF DIMENSION REDUCTION



Top Words (post normalization)

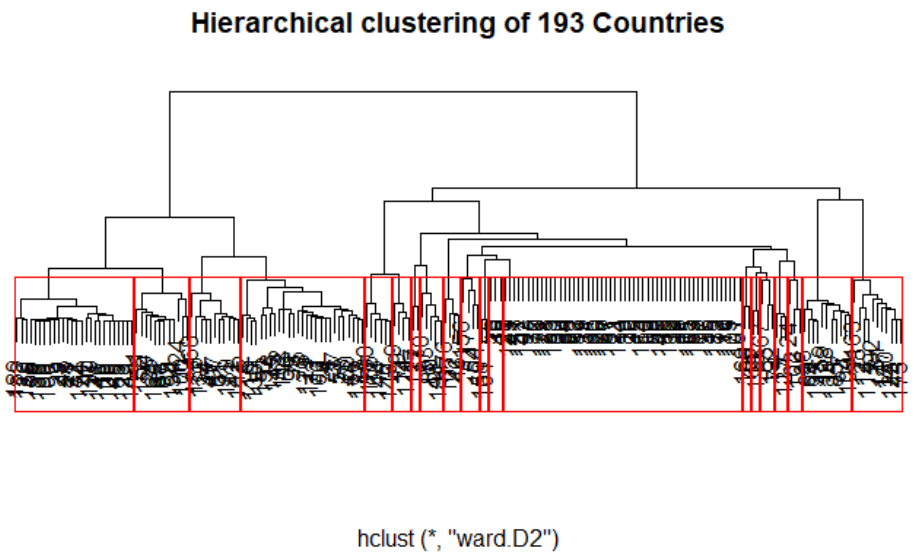
Malaysia:

- yang
- di
- yang di
- agong
- pertuan agong

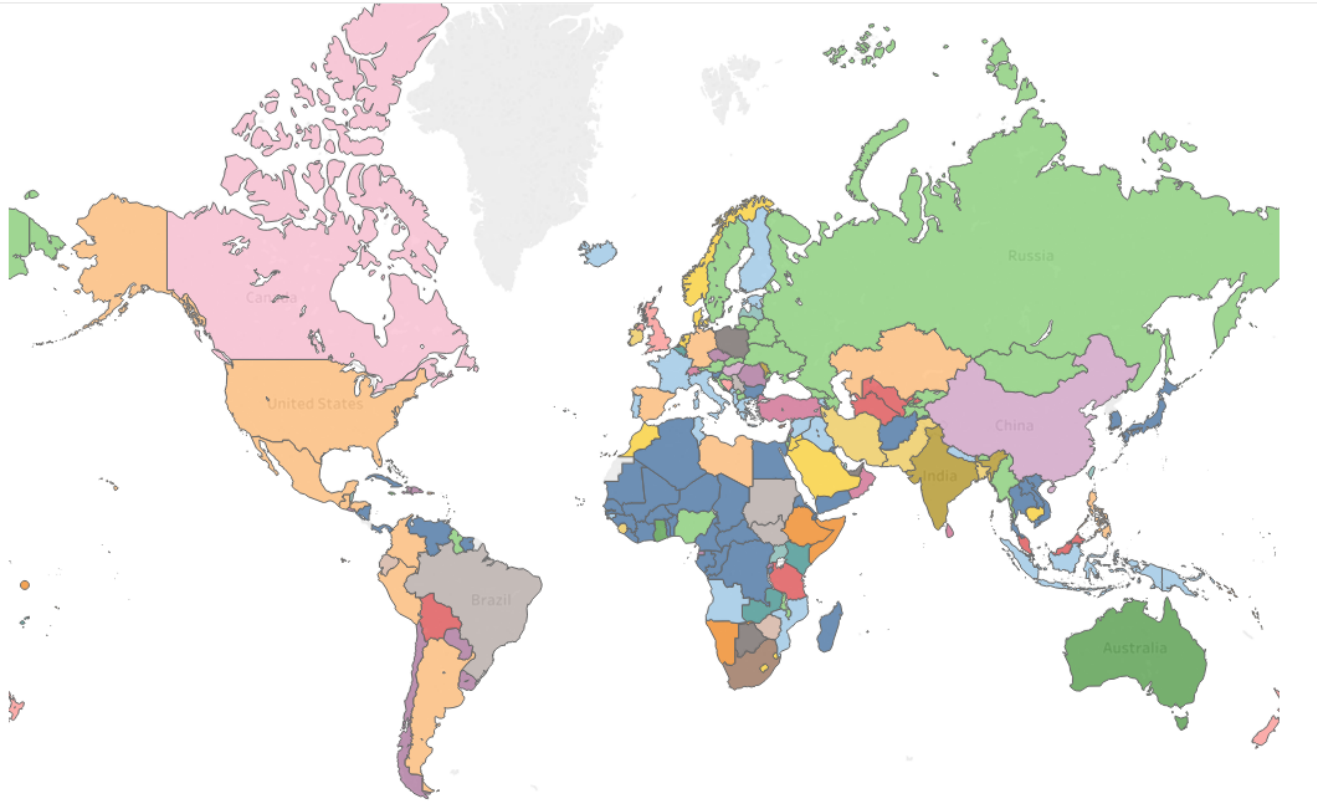
United Kingdom:

- subsect
- welsh
- lord
- welsh Minist
- lord Chancellor

RESULTS - DOCUMENT SIMILARITY GROUPED INTO 35 CLUSTERS




```
hc <- hclust(cdist, "ward.D2")
clustering <- cutree(hc, 20)
plot(hc, main = "Hierarchical clustering of 193 Countries",
     ylab = "", xlab = "", yaxt = "n")
rect.hclust(hc, 20, border = "red")
```

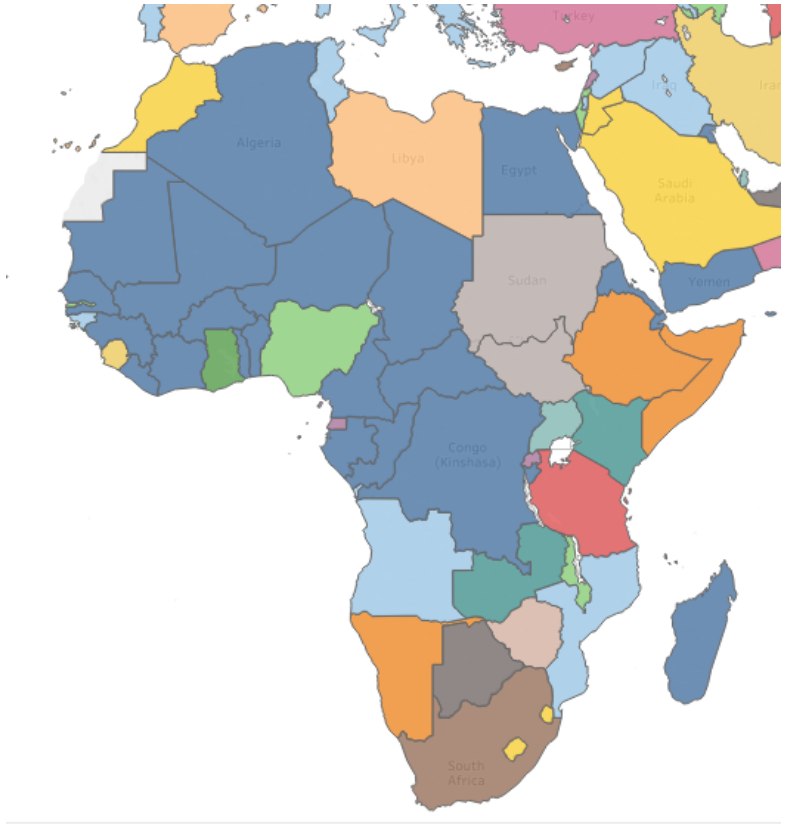


RESULTS – GEOPOLITICAL INFLUENCES

Historical Context of Azerbaijan

- Contested territory between Russia and Iran in the 1800s
- Annexed into Soviet Union in 1920
- Following collapse of Soviet Union in 1991, Azerbaijan gained independence
- Azerbaijan constitution written in 1995

RESULTS – GEOGRAPHIC CLUSTERING



West Africa

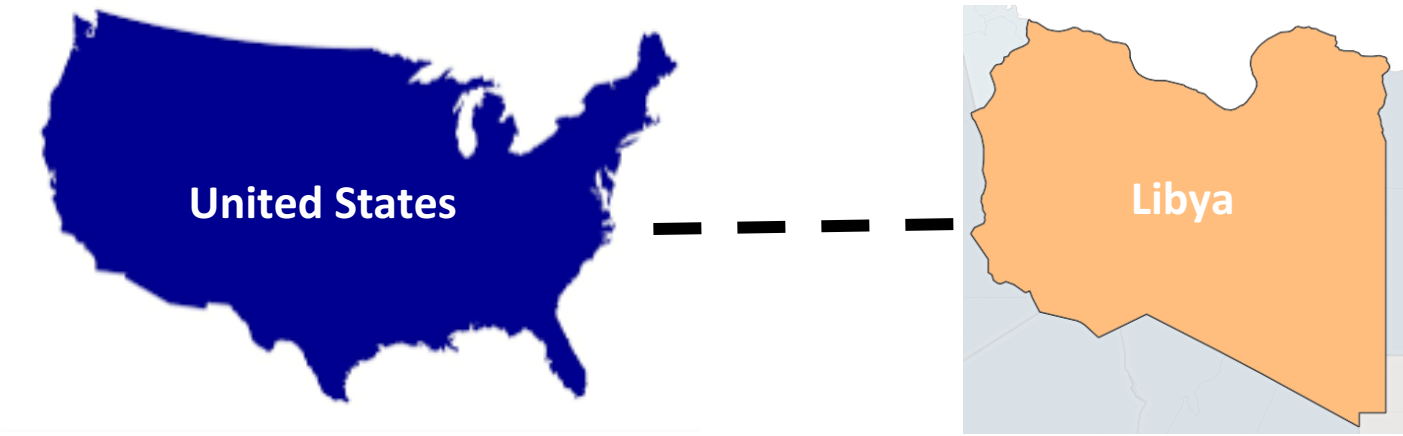
Eastern Europe/Eurasia

EXPECTATION VS OUTCOME

Top Words
1. unit state
2. congress
3. sever state
4. senat

…

6. congress_shall



Top Words
1. transit_council
2. nation_transit
3. libyan

…

6. nation_congress
8. congress

Potential explanation of variance

- Constitution similarity may not be the best predictor of geopolitical history or geographical position
- Information loss through translation
- Parameters not fully optimized (i.e. dimension reduction or cluster size)
- Two seemingly different countries can derive similar ways to govern

Future Consideration

- In addition to the documents themselves, we could have added additional descriptive variables describing the type of governmental system used in each country which would have led to more accurate grouping
- Certain countries often included religious language in their contusions; maybe foundation demographic variables could have been included to capture this variation
- A different clustering system would produce different results from what we received; these additional test can help further confirm our findings