Using NLP to Analyze Constitutional Preambles

Nicholas Archambault

1 April 2021

Introduction

Scholars have posited that the United States Constitution has exerted a profound influence on the language of the founding documents of nations across the globe, either through verbatim quotation or paraphrasing. Studies have attempted to measure constitutional influence across global founding documents by comparing the constitutional rights they share, and how such relationships evolve over time.

This project will explore a different approach predicated on using natural language processing (NLP) to assess the textual similarity of constitutional preambles, which typically illuminate the guiding purpose and principles for the rest of the document.

Variables and their descriptions can be found in the following table.

Variable	Description
country	country name
maps_country	country name compatible with R's `maps` package
year	year constitution was created
preamble	raw English text of constitutional preamble

1. Preprocessing and Data Visualization

Prior to conducting any analysis, we must preprocess the textual data into a form more conducive to manipulation. There are a number of useful packages for managing textual data, especially `tm`, which provides various NLPtools and techniques. Some of these functions include eliminating extraneous words or whitespace, and removing stopwords, the most commonly used words in a language, such as 'a' and 'the.'

One of the most important NLP functions of `tm` and its constituent package, `SnowballC`, is *stemming*, which prunes a root word of prefixes and suffixes so that various forms of that word can be recognized. The stemmed form of 'government' and 'governing' is 'govern.'

Commonly used functions to preprocess raw texts are listed in the table below.

Function	Description
tolower()	transform to lower case
stripWhitespace()	remove white space
removePunctuation()	remove punctuation
removeNumbers()	remove numbers
removeWords()	remove specified words
stemDocument()	stem the words in a document for specified language

```r
# Function for printing with line breaks
printlb <- function(x) {
  cat(paste(strwrap(paste(x, collapse = " ")), collapse = "\n"))
}

# Set seed
set.seed(12)

# Load packages
suppressPackageStartupMessages(library(igraph))
suppressPackageStartupMessages(library(SnowballC))
suppressPackageStartupMessages(library(tm))
suppressPackageStartupMessages(library(wordcloud))
suppressPackageStartupMessages(library(maps))

# Load data
constitutions <- read.csv("constitution.csv", stringsAsFactors = FALSE)
mapnames <- constitutions$maps_country
constitutions <- constitutions[, -1]
```

The `tm_map()` function enables NLP operations on the corpus, the body of raw text documents. From the corpus we can create the *document-term matrix*, a rectangular array with rows representing founding documents and columns representing unique word terms. The document-term matrix counts the frequency of each term listed in each document — in this case, each individual constitutional preamble. The (i, j) element of the matrix lists the count of the jth column term within the ith document.

Using the distribution of *term frequency* within a document allows us to infer topics found in the text. The downside of analyzing term frequency is that a term's importance loses value if that term appears frequently across all documents in the corpus. The *term frequency-inverse document frequency* (tf-idf) downweights terms that appear frequently across all documents in order to remedy this problem. For a document d and term w, the tf-idf(w, d) is defined as

$$tfidf(w, d) = tf(w, d) \times idf(w)$$

.

The expression tf(w, d) represents the number of occurrences of term w in document d; it equals zero when w never occurs in d.

The *inverse document frequency* is defined as

$$idf(w) = log\left(\frac{N}{df(w)}\right)$$

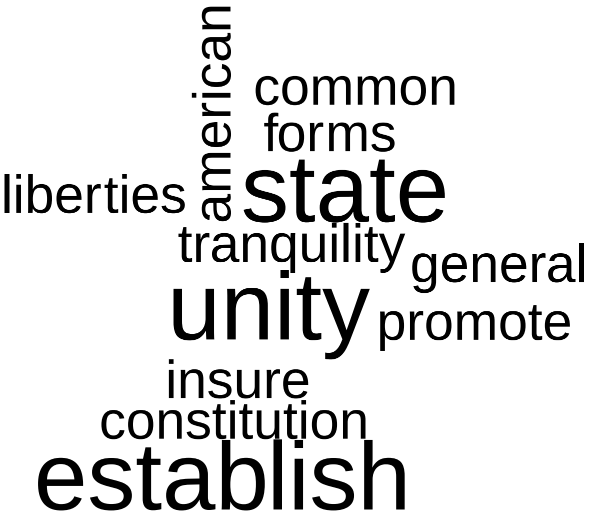
,

where N is the number of documents in the corpus and df(w) is the number of documents that contain term w. Below, we compute the tf-idf metric with the `weightTfIdf()` function, which takes as an input the document-term matrix output from the `DocumentTermMatrix()` function. We activate the `normalize` argument to divide the term frequency of each term by the total number of terms in a given document.

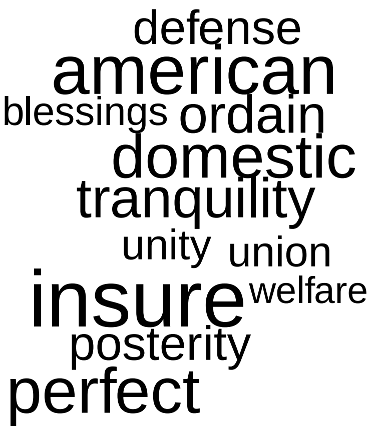
```r
# Prepare corpus
corpus <- VCorpus(VectorSource(constitutions$preamble))
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removeWords, stopwords("english"))
corpus <- tm_map(corpus, stripWhitespace)
corpus_unstemmed <- corpus
corpus <- tm_map(corpus, stemDocument)

# Make raw and tfidf term matrices
term_mat_tm <- DocumentTermMatrix(corpus)
term_mat <- as.matrix(term_mat_tm)
term_mat_tfidf <- as.matrix(weightTfIdf(term_mat_tm, normalize = TRUE))
rownames(term_mat) <- rownames(term_mat_tfidf) <- constitutions$country
```

With preprocessing complete, we visualize the results of topic inference with a wordcloud, a graphic that lists more frequent terms in larger fonts. Inputs of the `wordcloud()` function include a vector of words — the preamble text — and a vector of their frequencies. We limit the number of words shown to 12.

```
# Make word clouds
terms_unstemmed <- stemCompletion(colnames(term_mat), corpus_unstemmed)

# Raw
wordcloud(terms_unstemmed,
          term_mat["united_states_of_america", ],
          max.words = 12, scale = c(3, 0.25))
```



```
# Tf-Idf
wordcloud(terms_unstemmed,
          term_mat_tfidf["united_states_of_america", ],
          max.words = 12, scale = c(2.5,0.25))
```



Wordclouds generated by the raw document-term and tf-idf matrices for the United States Constitution are quite similar, though the wordcloud formed from the raw matrix of terms fails to capture the unique essence and tone of the famous U.S. preamble. The most common word stems in the raw analysis include 'state', 'unit', and 'establish', while the adjusted analysis of the tf-idf matrix reveals idiosyncratic terms that set the U.S. preamble apart and appear less frequently in the founding documents of other countries: 'insure', 'america', 'perfect', 'domestic', and 'tranquility.'

2. Clustering Similar Preambles with K-Means

The k-means algorithm is an iterative algorithm in which operations are repeatedly performed until no discernible difference in results is produced. Frequently employed as a nonlinear technique in machine learning classification problems, k-means attempts to split data into k similar groups, each associated with a centroid, a point equal to the within-group mean. The algorithm requires prior choosing of the value of k. It first assigns each observation to its closest cluster, then recomputes cluster centroids based on the new assignments. This process repeats until changes to the structure of the data no longer surpass a minimal threshold of alteration.

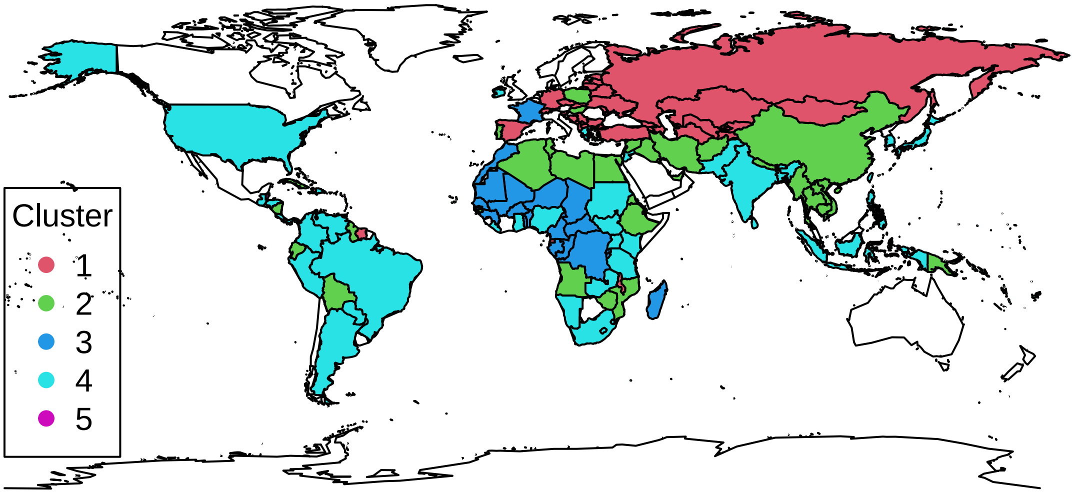
In R, the base function `kmeans()` provides a straightforward method for performing k-means clustering. We group countries into five clusters based on the linguistic similarity of their constitutional preambles, first normalizing the row vectors to eliminate the effects of varying preamble lengths. The clustering results can be printed and visualized with a world map from the `maps` package.

```r
# Normalize `term_mat_tfidf`
row_vector_lengths <- sqrt(rowSums(term_mat_tfidf^2))
term_mat_normalized <- term_mat_tfidf / row_vector_lengths

# Run k-means clustering
kmeans_results <- kmeans(term_mat_normalized, centers = 5)
for (i in 1:5) {
  cat("Cluster ", i, ":\n")
  printlb(constitutions$country[kmeans_results$cluster == i])
  cat("\n\n")
}
```

```
## Cluster  1 :
## albania armenia azerbaijan belarus bosnia_and_herzegovina bulgaria
## czech_republic equatorial_guinea estonia georgia germany haiti
## kazakhstan kosovo kyrgyzstan latvia lithuania
## macedonia_the_former_yugoslav_republic_of malawi moldova_republic_of
## mongolia montenegro russian_federation serbia slovakia slovenia spain
## suriname switzerland tajikistan turkey turkmenistan ukraine uzbekistan
```

```
##
## Cluster  2 :
## afghanistan algeria andorra angola bahrain
## bolivia_plurinational_state_of cambodia cape_verde china cuba ecuador
## egypt eritrea ethiopia gambia guinea-bissau guyana hungary
## iran_islamic_republic_of iraq korea_democratic_people's_republic_of
## kuwait lao_people's_democratic_republic libya mozambique myanmar nepal
## nicaragua papua_new_guinea poland portugal sao_tome_and_principe
## syrian_arab_republic thailand timor-leste tunisia united_arab_emirates
## viet_nam zimbabwe
##
## Cluster  3 :
## benin burkina_faso burundi cameroon central_african_republic chad
## comoros congo_democratic_republic_of_the congo djibouti france gabon
## guinea madagascar mali mauritania morocco niger rwanda senegal togo
##
## Cluster  4 :
## argentina bahamas bangladesh bhutan brazil brunei_darussalam colombia
## costa_rica dominican_republic fiji ghana greece guatemala honduras
## india indonesia ireland japan jordan kenya kiribati korea_republic_of
## liberia liechtenstein marshall_islands micronesia_federated_states_of
## namibia nauru nigeria pakistan palau panama paraguay peru philippines
## saint_kitts_and_nevis saint_vincent_and_the_grenadines samoa seychelles
## solomon_islands south_africa south_sudan sri_lanka sudan swaziland
## taiwan tonga tuvalu uganda united_republic_of_tanzania
## united_states_of_america vanuatu venezuela_bolivarian_republic_of
## zambia
##
## Cluster  5 :
## antigua_and_barbuda barbados belize dominica grenada saint_lucia
## trinidad_and_tobago
```

```r
# Plot countries by their cluster
map("world")
for (i in 1:nrow(constitutions)) {
  if (mapnames[i] != "") {
    map(database = "world", regions = mapnames[i],
        col = unname(kmeans_results$cluster[i]) + 1,
        fill = TRUE, add = TRUE)
  }
}
legend(-180, 20, title = "Cluster", legend = 1:5, col = 2:6, pch = rep(19, 5))
```



The clusters quite clearly demarcate global influence and the vestiges of colonial rule or historical interference. We see, for example, that the cyan Cluster 4 encompasses not only the United States, but also much of South America, reinforcing our conception of the lingering influence of American thought and political meddling in the Western Hemisphere.

Additionally, we notice a red Cluster 1 containing former Eastern Bloc states that extends into Eastern Europe and Central Asia. A Middle East-centered green Cluster 2 spreads into North Africa, enfolding countries with similar geographies, economies, and political histories. Finally, we observe a blue Cluster 3 that encompasses France and its former colonies in northern and central Africa, underscoring the lasting effects of colonialism and its influence on the founding philosophies of formerly subject states.

One downside of using the k-means algorithm, however, is that in cases like this it is highly unstable and dependent on the randomly chosen initial cluster centroids. We set the randomizing seed at the beginning of this script in order to replicate the same results over multiple trials, but we would assuredly generate different clustered groupings if this seed were not fixed.

The instability of the algorithm means that while it is interesting to superficially analyze which nations fall into which clusters, we should carefully interpret the results and avoid reading too much into these groupings.

3. Similarity of Foreign Constitutions to the U.S. Constitution

Having preprocessed and explored the constitutional data, we turn to the main purpose of this project: identifying the influence of the U.S. Constitution on the founding documents of other nations.

To compare the similarity of documents found in the document-term matrix, we'll define a function to assess *cosine similarity*, the angle θ between two corresponding n-dimensional vectors $a = (a_1, a_2, ..., a_n)$ and $b = (b_1, b_2, ..., b_n)$. In its full form, this similarity is defined as

$$cos\theta = \frac{a \cdot b}{||a|| \times ||b||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

.

The numerator of this expression represents the dot product of vectors a and b, while the denominator is the product of their lengths. Cosine similarity values range from -1, where documents are most dissimilar and the vectors lie in opposite directions, to 1, where documents are identical and the vectors completely overlap.

```
# Cosine similarity function from exercise
cosine <- function(a, b) {
  numer <- apply(a * t(b), 2, sum)
  denom <- sqrt(sum(a^2)) * sqrt(apply(b^2, 1, sum))
```

```r
  return(numer / denom)
}

# Index of USA
us <- (rownames(term_mat) == "united_states_of_america")

# Get similarity with US using `term_mat_tfidf`
us_similarity_tfidf <- cosine(term_mat_tfidf[us, ], term_mat_tfidf[!us, ])
sort(us_similarity_tfidf, decreasing = TRUE)[1:5]
```

```
##   argentina philippines       ghana      brazil      taiwan
##  0.26250584  0.17965664  0.09919651  0.09733481  0.09551709
```

```r
# Raw similarity
us_similarity_raw <- cosine(term_mat[us, ], term_mat[!us, ])
sort(us_similarity_raw, decreasing = TRUE)[1:5]
```

```
##   argentina  azerbaijan philippines       spain      latvia
##   0.4304580   0.3309113   0.3214030   0.2703721   0.2641644
```

Applying the cosine similarity function to each row in the matrix, we can gauge which constitutions are most similar to the vector representing the U.S. Constitution.

The top five most similar constitutions have moderate similarity scores, though results vary depending on whether we use the raw or tf-idf weighted matrix. Argentina and the Philippines appear on both lists. The presence of the Philippines is unsurprising, given the long history of U.S. influence on the islands. Argentina is a bit puzzling, but the Argentine War of Independence occurred roughly 20 years after the crafting of the U.S. Constitution. It's possible that the U.S. Constitution served as inspiration for the newly-liberated Argentine revolutionaries.

The presence of other countries on this list is less intuitive. We can print the preambles of the Ghanaian and Azerbaijani constitutions, two of the most notable surprises, to read how their texts compare to the preamble of the United States.

```r
printlb(constitutions$preamble[constitutions$country == "ghana"])
```

```
## IN THE NAME OF THE ALMIGHTY GOD We the People of Ghana, IN EXERCISE of
## our natural and inalienable right to establish a framework of
## government which shall secure for ourselves and posterity the blessings
## of liberty, equality of opportunity and prosperity; IN A SPIRIT of
## friendship and peace with all peoples of the world; AND IN SOLEMN
## declaration and affirmation of our commitment to; Freedom, Justice,
## Probity and Accountability, The Principle that all powers of Government
## spring from the Sovereign Will of the People; The Principle of
## Universal Adult Suffrage; The Rule of Law; The protection and
## preservation of Fundamental Human Rights and Freedoms, Unity and
## Stability for our Nation; DO HEREBY ADOPT, ENACT AND GIVE TO OURSELVES
## THIS CONSTITUTION.
```

```r
printlb(constitutions$preamble[constitutions$country == "azerbaijan"])
```

```
## The Azerbaijan people, continuing the traditions of many centuries of
## their Statehood, guided by the principles which are reflected in the
## Constitutional Act on the State Independence of the Republic of
## Azerbaijan, wishing to provide welfare for all and everyone, and to
## establish justice, freedom, security, and being aware of their
## responsibility before past, present, and future generations, exercise
## their sovereign right by solemnly declaring the following goals: to
## protect the independence, sovereignty and the territorial integrity of
## the Republic of Azerbaijan; to guarantee the democratic system within
## the framework of the Constitution; to achieve the realization of a
## civil society; to establish a law-governed, secular state which assures
## the supremacy of the law as an expression of the will of the people; to
## assure to all a decent level of life in accordance with a just economic
## and social order; to live under conditions of friendship, peace and
## safety with other peoples, maintaining a commitment to general human
## values and to implement a mutually beneficial cooperation for these
## purposes. For the sake of the above stated high intentions, this
## Constitution shall be adopted through a nationwide referendum.
```

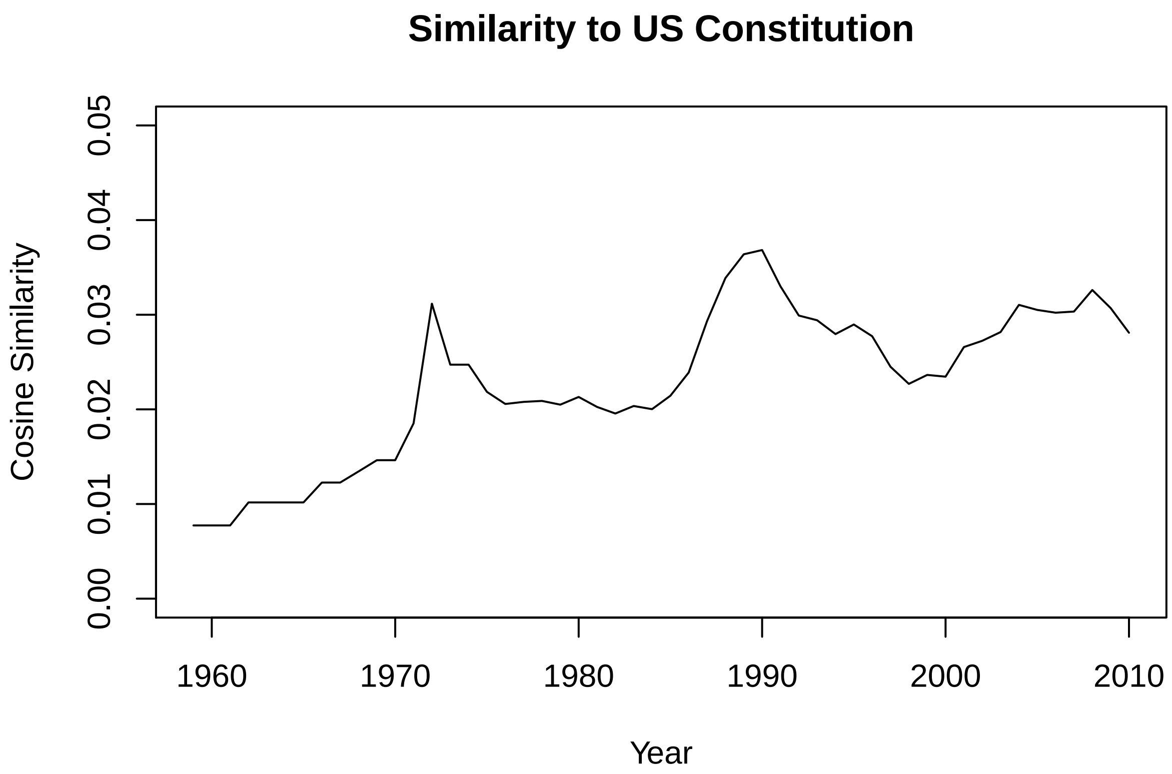
There are some clear similarities between the preambles. For example, the passage "secure the Blessings of Liberty to ourselves and our Posterity" in the American preamble is too similar to the passage "secure for ourselves and posterity the blessings of liberty" in the Ghanaian preamble to be a coincidence. The Azerbaijani preamble uses many words that appear in the U.S. preamble, but there are no passages that are clearly taken from the American document.

4. Examining U.S. Constitutional Influence Chronologically

To better explore the original research question — whether the influence of the U.S. Constitution on other countries' founding documents has waned over time — it would be worthwhile to examine cosine similarity chronologically. We can conduct this analysis with two approaches that are related but distinct.

The first method defines the cosine similarity for a given year as the average similarity across the previous ten years. The value for 2010, for example, would be the average cosine similarity with the U.S. Constitution of all constitutions created between 2000 and 2009.

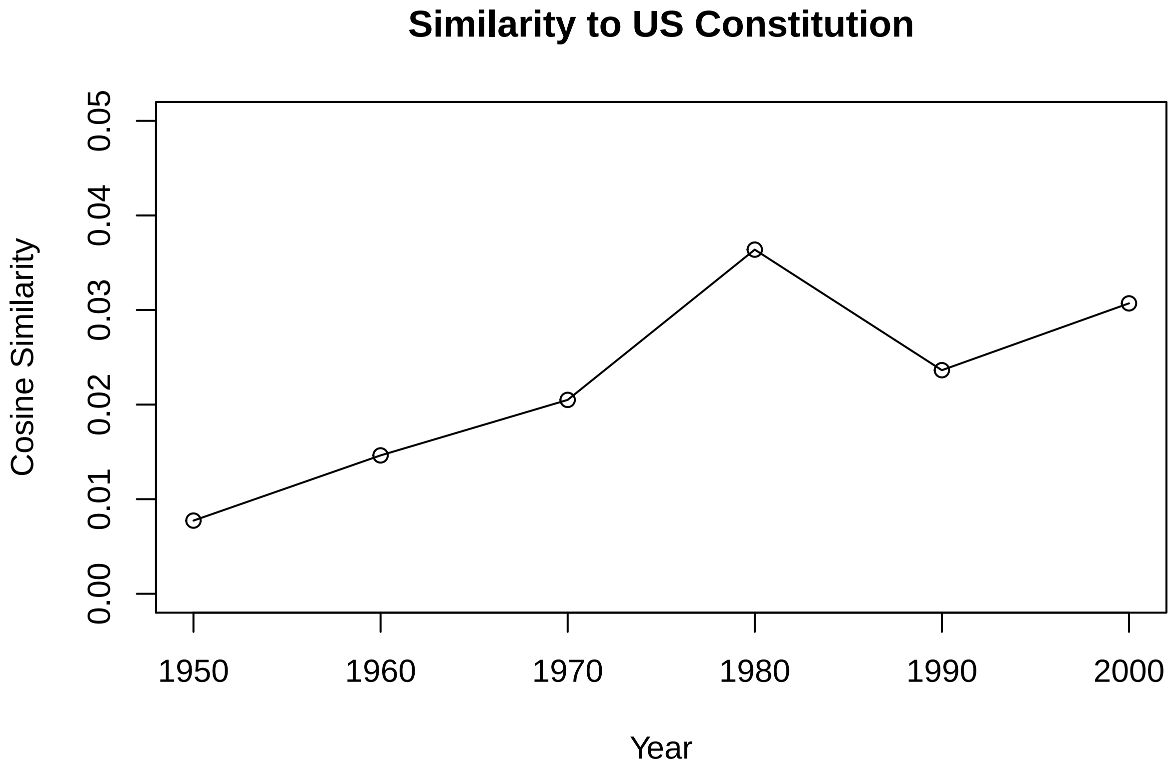
```r
years <- 1959:2010

# Function for computing moving similarity average over a given time interval
moving_average <- sapply(years, function(x) {
  mean(cosine(term_mat_tfidf[us, ],
              term_mat_tfidf[constitutions$year %in% (x - 9):x, ]))
})

# Plot annual average similarity for preceding ten years
plot(years, moving_average,
     type = "l", ylim = c(0, 0.05),
     main = "Similarity to US Constitution",
     xlab = "Year", ylab = "Cosine Similarity")
```



Average annual similarity has generally increased over time between 1960 and 2010. There are two notable spikes at the beginning of the 1970's and 1990's corresponding to leaps in average similarity throughout both the 1960's and 1980's. These spikes reflect major historical movements and realignments of geopolitical power: throughout the 1960's a number of former British and French states liberated themselves of colonial rule, and the 1980's saw protests and liberalization sweep through the Eastern Bloc, ultimately resulting in the fall of the Berlin Wall in 1989.

The second method of chronological analysis computes the average cosine similarity by decade. On the plot below, the similarity value for 1980 corresponds to the average similarity of the U.S. Constitution to all constitutions created between 1980 and 1990.

```r
# Plot decadal average similarity
decades <- seq(1950, 2000, 10)
plot(decades, moving_average[(years - 9) %in% decades],
     type = "l", ylim = c(0, 0.05),
     main = "Similarity to US Constitution",
     xlab = "Year", ylab = "Cosine Similarity")
points(decades, moving_average[(years - 9) %in% decades])
```



Average decadal similarity has generally increased since the end of World War II. The spikes in this plot are less pronounced, but they corroborate the previous finding that average similarity with the U.S. Constitution generally rose through the 1960's and 1970's, spiked strikingly in the 1980's, and dipped during the 1990's. The dip is possibly the result of the vast number of former Soviet states that created constitutions in the nineties and sought to maintain a certain degree of alignment with Soviet values and systems of governance.

Taken together, the findings of both approaches indicate that U.S. Constitutional influence has not diminished across the globe. Aside from a brief dip, its average textual similarity to other founding documents steadily grows, more than two centuries after the Constitutional Convention of 1787, as other nations continue to model the language of their preambles — and likely their entire constitutional documents — after those of the United States.

5. Constructing a Network of Influence

Cosine similarity and clustering are two methods of analysis which consider the relationships of constitutional preambles in isolation. We now turn to analyzing constitutional influence as a network of relationships between and among preambles.

Central to network analysis is the *adjacency matrix*, whose entries indicate the existence or absence of a relationship between two preambles The network we construct will be *directed*, meaning it contains directionality between senders and receivers of constitutional influence. The adjacency matrix of a directed network is not symmetric. This makes sense intuitively — if the constitution of the United States guided the creation and language of the constitution of the Philippines, that does not necessarily mean the effect was reciprocated. Influence propagated in one direction only.

In this analysis, the (i, j)th entry of the adjacency matrix represents the cosine similarity between the ith and jth constitutional preambles, where the ith constitution was created in the same year or after the jth constitution. The entry is zero of the ith constitution was created before the jth constitution, since influence can only propagate forward in time.

To fully explore the relationship network of all constitutional preambles, we will apply the *PageRank* algorithm using the `page.rank()` function in base R. The algorithm is an iterative measure of *centrality*, the extent to which each *node*, or unit, within a network plays a role in that network. PageRank was developed by Sergey Brin and Larry Page, the co-founders of Google, to optimize the ranking of websites within search engine outcomes. It is based on the idea that nodes with greater numbers of incoming edges — or lines of influence directed at them — are more crucial to the network. More intuitively, we can think of incoming edges as 'votes of support': constitutional preambles with the most incoming edges are the most influential, and they receive the most votes of support for their importance within the network. If a node has an incoming edge from another node with a large number of incoming edges, it results in a greater PageRank value than if it has an incoming edge from a node with fewer incoming edges. For example, if the United States has an incoming node from France, which has influenced many other constitutions across the globe, the PageRank value of the United States would be higher than if that edge came from Estonia, which has influenced far fewer constitutions.

The PageRank algorithm assigns a set of initial values to all nodes, then updates that value at each iteration using the formula

$$PageRank_j = \frac{1-d}{n} + d \times \sum_{i=1}^{n} \frac{A_{ij}}{outdegree_i}$$

.

In this equation, A_{ij} is the (i, j)th element of the adjacency matrix indicating whether or not an edge connects node i to node j; d is a constant typically set to 0.85; n is the number of nodes in the network. PageRank for a given node j equals the sum of 'votes' from other nodes that have an incoming edge into node j. If there is no edge from node i to node j, then $A_{ij} = 0$, and therefore no vote is given to node j from node i. However, if $A_{ij} = 1$, then a vote from node i to node j is equal to the PageRank value of node i divided by node i's *outdegree*, the number of edges protruding from it. Each node must equally allocate its PageRank value across all other nodes to which it has outgoing edges. For example, if a node has a PageRank value of 0.1 with two outgoing edges, then each receiver obtains 0.05 from this node. The algorithm's iteration stops when the PageRank values for all nodes no longer change. The sum of all PageRank values is 1.

```r
# Apply the `cosine` function to each row of matrix
cosine_mat <- apply(term_mat_tfidf, 1, cosine, term_mat_tfidf)

# Set cells where row constitution is older than column to zero
cosine_mat[sapply(constitutions$year, `>`, constitutions$year)] <- 0

# Make graph
cosine_graph <- graph.adjacency(cosine_mat, mode = "directed",
                                weighted = TRUE, diag = FALSE)
# Get page rank
sort(page.rank(cosine_graph)$vector, decreasing = TRUE)[1:5]
```

```
## united_states_of_america                argentina                   latvia
##               0.11733124               0.06360799               0.03352647
##                    tonga                  ireland
##               0.03309379               0.02947772
```

The top five most important constitutions, according to their centrality within the network, are those of the United States, Argentina, Latvia, Tonga, and Ireland. The score for the U.S. is clearly the largest among all countries, nearly double that of second-place Argentina.

Based on previous results, we should not be surprised at the suggestion that the U.S. Constitution is a keystone within the network of constitutional influence. We previously observed that newly-created constitutions generally became more similar to the U.S. Constitution throughout the twentieth century. The conspicuously high PageRank score of the U.S. Constitution validates prior suggestions that the ideals to which it aspires exert considerable sway on the architecture of constitutions worldwide.

We must treat these results carefully, however, since the PageRank approach grants more edges to older constitutions. It is not a coincidence that the top two most influential constitutions, those of the United States and Argentina, are the two oldest found in the data set (ratified in 1789 and 1853, respectively). The other three Constitutions in the top five are also fairly old — Latvia's was created in 1922, Tonga's in 1875, and Ireland's in 1937.

Without doubt, older constitutions exert more influence across the world, as they are the beneficiaries of age and concomitant renown. The PageRank algorithm, however, artificially inflates their influence to a certain, unknown degree. Rather than focusing on the place and influence of the U.S. Constitution within the broader global network, it may be shrewder to examine the causal effect of the U.S. Constitution's creation. This is a much more difficult question to address.

Conclusion

In this project, we employed a number of natural language processing tools and techniques to examine the influence of the preamble of the United States Constitution on the preambles of other nations' founding documents. After substantial preprocessing, we visualized the Amreican preamble with wordclouds to understand how terminology and term frequency vary between the raw and tf-idf weighted matrices. We next used the k-means algorithm to group global constitutions into five clusters based on similar topics and sentiments, then implemented a cosine similarity function to identify the five constitutions most similar to that of the United States. Two of the most similar, Argentina and the Philippines, were logical and to some degree expected. Others, such as Ghana and Azerbaijan, were initially sources of puzzlement; examining the text of those countries' preambles, however, revealed language that was heavily paraphrased from the U.S. preamble.

By evaluating rolling ten-year and decadal averages, we determined that the influence of the U.S. Constitution on other founding documents crafted throughout the back half of the twentieth century has grown consistently across all but one decade. Examining the data as a network through the lens of Google's PageRank algorithm revealed that the constitutions of the United States and Argentina are two of the most crucial influencers within the global network. These results, however, should be taken with caution, as these two constitutions are among the world's oldest, and the nature of the directed network we constructed artificially inflated the importance of older documents within it.

This exercise was based in part on the 2012 work of David S. Law and Mila Versteeg: "The declining influence of the United States Constitution," which appeared in the *New York University Law Review*, vol. 87, no. 3, pp. 762-858, as well as the 2012 rebuttal of Tom Ginsburg and James Melton: "Comments on Law and Versteeg's 'The declining influence of the United States Constitution,' ", which appeared in the *New York University Law Review*, vol. 87, no. 6, pp. 2088-2101.