**Research Design**

**Problem Statement:**

Since the end of the cold war, non-democratic national governments have attempted to do a better job of disguising their authoritarian traits. This has led to the rise of competitive authoritarian governments. These governments have elements of democracies, including frequent elections, that allow the government to claim to be democratic. However, they violate critical democratic elements including, free and fair elections, civil liberties, & voting access (nearly the whole adult population can vote). Many of these rules and concepts are secured through constitutions which define the scope and role of institutions in these new governments. The modern world has a harder time than ever determining if a nations regime is truly a democracy or a competitive authoritarian government. We want to know if there is a natural language difference in the constitutions of democracies and competitive authoritarian governments. Do democratic nations constitutions contain different natural language elements than competitive authoritarian governments? And if they do, can we use natural language processing (NLP) techniques to predict the regime of a nation based on their constitution? This project will undertake an initial exploration into whether NLP of constitutions can predict regime type. We suspect that a highly focused NLP analysis would be able to identify differences, possibly through relationship extraction or sentiment analysis, between the regimes. However for this initial, relatively low complexity NLP analysis, we hypothesize that democratic and competitive authoritarian constitutions will be extremely similar and may lack a natural language difference.

**Research Plan:**

**Corpus:**

Our corpus, or dataset, will be composed of the constitutions of ten nations. Each constitution will be translated into english, our data comes from the comparative constitutions project. This will be split into two groups of five democracies (D) and five competitive authoritarian (CA) governments. We have choses nations based on their relatively clear status as either a democracy or competitive authoritarian government. We tried to select a representative sample of nations in many regions of the world. We chose the following nations:

* CA governments: Turkey, China, Russia, Egypt, Venezuela
* D governments: US, France, Germany, Australia, Japan

**Analysis:**

We plan to use a combination of micro and macro level analysis to investigate our research question. The micro level analysis will analyze features at the sentence level for each document in our corpus. The macro level analysis will investigate the similarities and differences between the constitutions of democracies and competitive authoritarian governments. We will examine the constitutions of each nation individually at sentence level. This micro level analysis will inform us of the characteristics and key elements of each individual nation. The micro level analysis look at features for attributes such as 1,2,3 word pair frequency, number of words, frequency of words, topic modeling (LDA), and possibly a relationship extraction. Then we will compare the micro analysis of the democratic constitutions with the competitive authoritarian constitutions to see if there are macro level differences. We will use the microfeatures as our IVs and the constitution type (Democratic vs Competitive authoritarian) as a binary DV to train our model.

**NLP pipeline:**

1. Original Corpus
2. Data exploration
3. Cleaning & Preprocessing (lower -> contractions -> stop words -> stemming -> lemma -> tokenization -> POS tag)
4. Feature extraction (Word count, Sentence count, Paragraph count, Top\_Bigrams, Top\_Trigrams, Top\_Quadgrams, POS\_Counts, NER\_Counts)
5. Modeling (Machine Learning or logistic model)
6. Evaluation

**Evaluation:**

We will likely use a test-train split to cross-validate our model and prevent overfitting, however, we fear our sample size of 10 documents may be too small to effectively use the method. Therefore, we may simply undertake a logistic regression or similar model as an initial exploratory analysis to determine if there are sufficient differences in their natural language. This initial exploratory analysis could act as a test case that will allow researchers to later capture a much larger corpus of constitutions and test prediction ML models.  
If we use a logistic regression, we will evaluate the model based on the substantive and statistical significance of our variables and the overall model fit (AIC and adj. R sq). If we use a machine learning model and test-train split, we will evaluate our model using a confusion matrix and area under the ROC curve. This will allow us to evaluate the precision, recall, & accuracy of our model. These metrics will determine the effectiveness of our model at predicting if the constitution of a nation belongs to a democracy or competitive authoritarian government.