**Spirling (2012) – U.S. Treaty Making with American Indians: Institutional Change and Relative Power**

**Unsupervised Machine Learning** – using R (text-data packages – e.g., Quant).

Problems of just comparing document to document (**bag of words approach – comparing if 2 bags of words have the same things in them**).

* 1. Vast majority of words in a document are likely to be the most common words (problem).
  2. There are different sentences/words that very often mean the same thing.

**Strategy**: forget about whole words, lets use part words.

**Objective**: Investigates the effect of the constitutional change in 1871 on the treaty-making process with Native American tribes. This change removed the President's ability to negotiate treaties directly, transferring this power to Congress.

* **Context**: Treaties between the U.S. government and Native American tribes are unique, with almost 600 documents signed from the Revolutionary War to the turn of the 20th century.
* **Findings**: Institutional changes in the treaty-making mechanism had little impact on the outcomes of treaties. Instead, the relative economic and military power of the United States was the main factor that led to increasingly unfavorable terms for Native Americans over the 19th century.
* **Data Collection**: Spirling digitized all treaties made with Native American tribes into a novel dataset for systematic textual analysis.
* **Textual Analysis Approach**: Unsupervised machine learning techniques (scaling methods validated with word-use information).
  + Objectively analyzes the content of the treaties without imposing preconceived categories or themes.
* **Dimensionality**: Analysis identified a single dimension characterizing the treaties, which was their "**harshness**" in terms of land and resource cession. Suggests that the treaties could be placed on a scale from less to harsher based on their textual content.
* **Why Unsupervised Machine Learning?**: Allows for the discovery of patterns or structures within the data without the need for predefined classes or categories.
  + Here, it enabled the identification of the harshness dimension in an objective manner, based solely on the text of the treaties themselves, rather than on external interpretations or classifications.
  + Novel way to quantify and analyze the terms of treaties over time, leading to insights about the changing nature of treaty terms and the factors influencing these changes.

**Methodology Specifics**:

* **Objective**: Analyze treaties (seen as contracts) to understand their underlying dimensions or themes without assuming any specific characteristics beforehand.
* **Unsupervised Learning**: Let the data—here, the treaty texts—speak for themselves and uncover patterns or dimensions naturally present in the documents.

**Relevant Terms**:

1. **Term-Document Matrix (TDM)**: Table that lists all the documents (treaties) and words (terms). Each cell in the table shows how often a word appears in a document. Essentially, transforms the text data into a numerical form that can be analyzed.
   * **Preprocessing**: Before creating the TDM, the text undergoes cleaning, like removing common but uninformative words (e.g., "the", "and") and stemming (reducing words to their root form).
   * **Limitation**: A downside of this approach is losing the order of words. For example, "no peace between us" and "peace between us" would look the same after removing common words, even though their meanings are opposite.
2. **String Kernels**: To overcome the limitation of losing word order, string kernels are used (considers the sequence of letters or characters in the texts).
   * **Substrings and p-Spectrum**: Imagine breaking down words into smaller pieces (substrings) of a certain length (p). For example, from "apartment", you get substrings like "part" or "ment". The p-spectrum is the collection of these substrings for a document.
   * **Comparing Documents**: Documents are compared based on how many substrings they have in common. This comparison includes the sequence of characters, helping to maintain some sense of the original word order.
3. **Kernel PCA for Scaling**: After preparing the data with string kernels, **Kernel Principal Component Analysis** (**PCA**) is used for scaling – finding one line that if you strung it through all the documents, the documents would be strung out the most (maximize variance – maximizing distance from center line). Helps reduce the complexity of the data by focusing on the main ways in which the documents differ from each other.
   * Once identified line – in this case **HARSHNESS** (the dimension) – look at what the actually represents.
   * **Kernel Matrix**: Before applying Kernel PCA, a kernel matrix is created from the comparisons of all document pairs – capturing the similarities between documents based on their substrings.
   * **Dimensionality Reduction**: Kernel PCA projects the documents onto a new space where the most significant patterns or dimensions become clearer.
     + Can reveal underlying themes or dimensions in the treaties without prior assumptions.

**Summary**:

Methodology employs advanced text analysis techniques to study treaties as complex texts without assuming what will be found. By turning the treaties into numerical data through the Term-Document Matrix and then focusing on the sequence of characters using string kernels, the researchers can maintain the nuanced differences in how words are used. Kernel PCA is then applied to identify the main dimensions or themes in the treaties based on these detailed textual analyses.

* As American military power increased (and Indigenous communities posed no threat) the harshness of treaties increased.

**What do the dimensions – uncovered from the scaling of treaty texts – actually represent?**

[*More method specifics/a bit redundant…*]

The aim is to understand what the dimensions, uncovered from the scaling of treaty texts, actually represent by identifying specific words or phrases that correlate with the positions of these texts on a scale (i.e., their thematic significance).

**Process**:

1. **Term Document Matrix (TDM) Creation**:
   * A TDM is constructed with words stemmed (meaning variations of the same word are treated as the same, e.g., "run" and "running"), and stop words (common words like "the", "and", etc.) plus punctuation removed.
   * Each entry in the matrix represents the term frequency-inverse document frequency (tf-idf) for a word in a treaty. Tf-idf increases with the number of times a word appears in a single document but decreases based on how common the word is across all documents. This highlights words that are uniquely frequent in some treaties but not common in all, making them potentially more significant for those treaties.
2. **Sparse Terms Removal**:
   * Words not appearing in at least 90% of the documents are removed to focus on terms that are relevant across the majority of treaties. This step reduces the dataset to 597 unique word stems.
3. **Transpose and Detrending**:
   * The matrix is transposed (denoted as X) to align rows with treaties and columns with word stems. A linear detrending is performed on the scale (Y) to account for shifts in language use over time. This ensures that the analysis can identify the significance of words/phrases beyond mere temporal variations in language.
4. **Regression via Random Forest –** machine learning technique that builds multiple decision trees and merges their results to improve accuracy and handle complex interactions between variables:
   * The detrended scale (Y) is then regressed on the transposed matrix (X) using a random forest algorithm.
   * Approach is particularly suitable here because it can handle situations where there are more variables (word stems) than observations (treaties), a common issue in text analysis.

**Outcomes**:

* The random forest algorithm provides a measure of how important each word stem is for predicting the position of a treaty on the identified scale.
  + Identified the most significant words (e.g., "land", "tract", "reserv", "relinquish", "friendship", "dollar", "boundari") that help differentiate the treaties based on their thematic content.

**[ELI5] Simplified Explanation**:

Imagine you have a bunch of treaties, and you're trying to figure out what themes or topics they cover without reading each one in detail. First, you turn these treaties into a big spreadsheet where each row is a treaty, and each column represents different words used in those treaties. But you only focus on unique words that make a treaty stand out and ignore common words that don't tell you much.

Next, you adjust this spreadsheet to account for changes in how people used words over time, ensuring that older and newer treaties can be compared fairly. Then, you use a smart algorithm (**random forest**) that can handle lots of data to figure out which words are the key to understanding the main themes of these treaties. This algorithm tells you which words help you distinguish between different types of treaties based on their themes, such as land agreements, friendships, or boundaries.

By the end, you have a list of words that, according to the algorithm, are crucial for understanding what each treaty is about without having to dig into each document manually. This method helps uncover the thematic essence of these treaties in a systematic and data-driven way.

**Results**:

* **Impact of Institutional Changes**: The shift in treaty-making authority in 1871 did not significantly alter the outcomes of treaties for Native Americans.
* **Influence of Relative Power**: The worsening terms for Native Americans in treaties over the 19th century were primarily due to the increasing economic and military power of the United States, rather than changes in the treaty-making process itself.

**Lecture Notes** – **Term Document Matrix**:

* **Spreadsheet**: each word is a row; each column is a document.
  + Cells are frequency of times that word occurs in a document.
* Rare words are often the most expressive, and words that are unique to the document probably capture the essence of the document more.
  + (ex.) If “evil” is there – and its not in the other documents – term frequency inverse document frequency (**tf-idf**) **can be used** (which is just tf/df); how often the term is in the document divided by document frequency.
    - If it was 5/1 (5 times evil appeared in the document/appeared in only one document); the rarer a word is, the more expressive the word is (thus important).