**CAN Treaties – Methods Overview**

**Using String Kernels:** When we use string kernels to measure the similarity between two texts, we look at common sequences of characters. Using the term “**majesti**” as an example, with a specified length of 5 characters:

1. **Split** “majesti” into all possible 5-character sequences (substrings):
   1. “**majes**”
   2. “**ajest**”
   3. “**jesti**”
2. **Compare Substrings**: We compare these 5-character sequences with the 5-character sequences from another text.
   1. e.g., if the other treaty text also has “majesty”, it would have the same substrings:
      1. “**majes**”
      2. “**ajest**”
      3. “**jesty**” (which shares “jesti” with “majesti”)
3. **Count Common Substrings**: We count how many of these 5-character sequences are common between the two texts. The more common sequences they have, the more similar the texts are considered to be.

A graph with blue dots and numbers

Description automatically generatedUltimately, this is how we get the graph(s) of the “thing” we want to look at – the overall theme or commonality across all the treaties – because we’ve computed the **Kernel Principal Component Analysis (KPCA)**. For Spirling this was harshness.

**How many “things” are there? – we can argue for 1 or 2:**

Arguably, we are looking for a **1st** and a **2nd component:**

* **Eigenvalues**: indicates the amount of variance in the data that is explained by its corresponding principal component.

When you see the elbow, you backtrack at the largest bend (if it’s a large enough bend comparatively to all bends). The biggest bend indicates which component captures the most variance in the data.

* + Higher eigenvalues mean that the principal component captures more variance from the data.

**A graph with colorful dots

Description automatically generatedComponent 1 Graph – “the main thing”:**

* If you draw a trend line across each of the treaties is fairly stable outside of the Douglas Treaties (slight dip around 1850s).

There’s some validation of this “main thing” trend in work by **Feir et al. (2023)** who did sentiment analysis on a sample of Canadian treaty texts.

* “*The length of treaty texts increased over two centuries of historical treaty-making, while the average sentiment in the treaty texts remained relatively constant, contrasting the changing sentiment in agreements between Indigenous nations and the United States during this same period.*”

**A graph with lines and dots

Description automatically generated**Arguably, when looking at word importance (below) we’re seeing consistency in language, possibly due to the fact that Canada remained attached to the British (vs. American independence). This 1st component could reflect crown involvement (“surrend”ing to “majesti”; “becom”ing “subject”s; “commission” involvement) – although interestingly, “white” is an important distinguisher.

* This is our **Messy Graph 1** (mirrors Spirling’s visual) with a trendline.
  + We’ll be working on adjusting these graphs in a way that can clearly distinguish each individual treaty in a more visually appealing manner.

**A diagram of a graph

Description automatically generatedComponent 2 Graph – the “secondary thing”:**

* A screenshot of a computer

  Description automatically generatedA screenshot of a computer

  Description automatically generated**A graph with colorful dots

  Description automatically generated**This component is more of a mystery (see *word importance* below).

Here’s the “**loadings**” of each treaty on each of the “things”.

* e.g., Referring back to the 1st graph (p. 2) the Douglas Treaties (DT01-DT10) are *all* very low on KPC 1 (“main thing”).
  + Hence the green dots appear at the bottom of that graph.

**How do we figure out what the “things” are? – we combine 2 methods:**

1. **Vector-space analysis: interpret string kernels directly**
   * We’ve done this already to determine the **KPCA**, but we could analyze the string kernels directly. However, academics add on a 2nd method to interpret string-kernels because it provides more nuance. Thus, we add in…
2. **Term Document Matrix conversion (stemming, etc.)**
   * Take the words and break them up into root components.
   * Eliminate common words (said, the, etc.)
   * Remove sparse terms (in our case, if the words don’t appear in 90% of documents they aren’t considered).
   * Remove punctuation.

The argument is that using string kernels gives a more accurate *overall* representation of what the things are – because when you use **string kernels** the algorithm *preserves and considers the order that word appear in*. However, it’s easier to add on a **TDM-IDF** process to assess each word’s importance in a vacuum.

* Once we’ve stemmed, removed common words, etc. we feed them to two algorithms – **Random Forest** (Spirling) and **xgboost** (new age extremely powerful competition-winning predictive algorithm that uses Random Forest and **Gradient Descent**).

**Importance Algorithms:**

1. **Random Forest:** you can think of the entire sample of the treaties as the “forest”. The algorithm plants 500-2000 “trees” at random points (word stems) in each treaty. Each tree makes a predictive “bucketing” decision (simplified below) extending throughout the entire corpus before the algorithm aggregates its calculations.
   * e.g., Planting a “**punish**” tree: In Numbered Treaty 10, “punish” might occur frequently, indicating it is high in the “thing.” Conversely, the Douglas Treaties might have few occurrences of “punish,” indicating they are low in the same “thing.”
     + The algorithm then asks: “If we split the treaties based on the occurrence of ‘punish,’ do they fall into the correct buckets?”

Numbered Treaty 10 **SHOULD** go into the bucket of “treaties high in thing,” and the Douglas Treaties would go into the bucket of “low in thing.”

If **punish** is **important**, we should see proper separation = punish is important:

If **punish** is **not important**, we might see inaccurate separation; we already know Douglas Treaties are low in thing, but “punish” is putting them in the wrong bucket = punish not important:

A scrolled paper with red trim

Description automatically generatedA person with arms spread out

Description automatically generated

**Douglas Treaties:  
 “thing” very low**

**Numbered Treaty 10:  
 “thing” very high**



[This Photo](https://www.pngall.com/scroll-png/) by Unknown Author is licensed under [CC BY-NC](https://creativecommons.org/licenses/by-nc/3.0/)

**Numbered Treaty 10**



[This Photo](https://www.pngall.com/scroll-png/) by Unknown Author is licensed under [CC BY-NC](https://creativecommons.org/licenses/by-nc/3.0/)



[This Photo](https://www.pngall.com/scroll-png/) by Unknown Author is licensed under [CC BY-NC](https://creativecommons.org/licenses/by-nc/3.0/)

**Douglas Treaties**

**Low in “thing” bucket**

**High in “thing” bucket**

**High in “thing” bucket**

**Low in “thing” bucket**

1. **XGBoost:** Extreme Gradient Boosting builds a model in a stage-wise fashion, where each new tree corrects errors made by previous trees.
   * XGBoost starts with a simple initial model, perhaps predicting the average value of “thing” across all treaties.
   * It calculates the difference (residual) between the actual values of “thing” (e.g., occurrences of important terms) and the predictions made by the initial model for each treaty.
   * A new decision tree is created to predict these residuals.

* e.g., Planting a “**punish**” tree:
  + If “punish” appears often in a treaty, the tree might predict a higher residual (indicating the initial model underestimated the “thing” for this treaty).
  + If “punish” is rare, the tree might predict a lower residual (indicating the initial model overestimated the “thing”).
* The predictions from this new tree are added to the initial model to improve its accuracy.
  + This process of calculating residuals, creating new trees, and updating the model continues iteratively, with each new tree focusing on the remaining errors from the previous iteration.
  + The final model is an ensemble of all the trees, where each tree contributes to refining the predictions.

**Algorithm Results – KPC1 (“main thing”):**

* **A graph with blue lines

  Description automatically generatedA screenshot of a computer

  Description automatically generated**Here we have a graph of the most important features from KPC1 (the words that best separate the treaties into the correct “main thing” buckets), and the corresponding algorithms scores.

**A screenshot of a computer

Description automatically generatedA graph with a bar

Description automatically generated with medium confidenceAlgorithm Results – KPC2 (“secondary thing”):**

* Here we have a graph of the most important features from KPC2 (the words that best separate the treaties into the correct “secondary thing” buckets), and the corresponding algorithms scores.

**Note on Term Importance:** we spent weeks verifying that the calculations are accurate, double checking Spirling’s methodology.

* Happily, **Feir et al. (2023)** – the sentiment analysis paper – also calculated term importance scores on Canadian treaties.
  + I only wish I had found the paper sooner – it would’ve saved hours of late nights and incessant worry.

**Component (“thing”) Correlations:**

When **stemming** words, we erase word suffixes to obtain the “root” of the word.



**Word stems with a high positive correlation** appear more frequently/are more prominent in treaties that have high scores on the principal component.

* These stems are characteristic of whatever “thing” the KPC is capturing when it scores high.

**Word stems with a high negative correlation** appear more frequently or are more prominent in treaties that have low scores on the principal component.

* These stems are characteristic of treaties that represent the opposite end of whatever the “thing” we’ve captured is.

**Word Stem correlations can be treated as Kernel Principal Component Loadings** – how much each word affects the “thing(s)”.

There are 3677 terms captured in the documents (which we have the correlations for) so I’ll just include the most important ones for each of KPC1 & KPC2.

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedKPC1: KPC2:**

**e.g., Assessing the Treaties with Highest & Lowest Treaties values of KPC1:**

**Highest: Numbered Treaty 10 (1906)**

* **Content Focus**: Extensive detailing of land cession, rights, and specific provisions for the Indigenous populations involved.
* **Nature of Agreement**: Specific commitments on both sides.
  + Outlines detailed rights to hunting, trapping, fishing, and the setup of reserves, as well as educational and agricultural support.
* **Rights and Compensations**: Clear definitions of compensations and rights, including annual payments and provisions for chiefs and headmen.

**Lowest: Douglas Treaties (1850)**

* **Content Focus**: Direct cession of land with fewer detailed rights or compensations.
* **Nature of Agreement**: Simplistic, primarily focusing on the surrender of lands with minimal protections or guarantees for the Indigenous populations beyond retaining village sites and some fishing and hunting rights.
* **Rights and Compensations**: Limited to one-time payment with no ongoing support or detailed rights enumerated for the future.

**Note:** Our search function web app that accesses where the word stems fall in treaties is operational.