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:

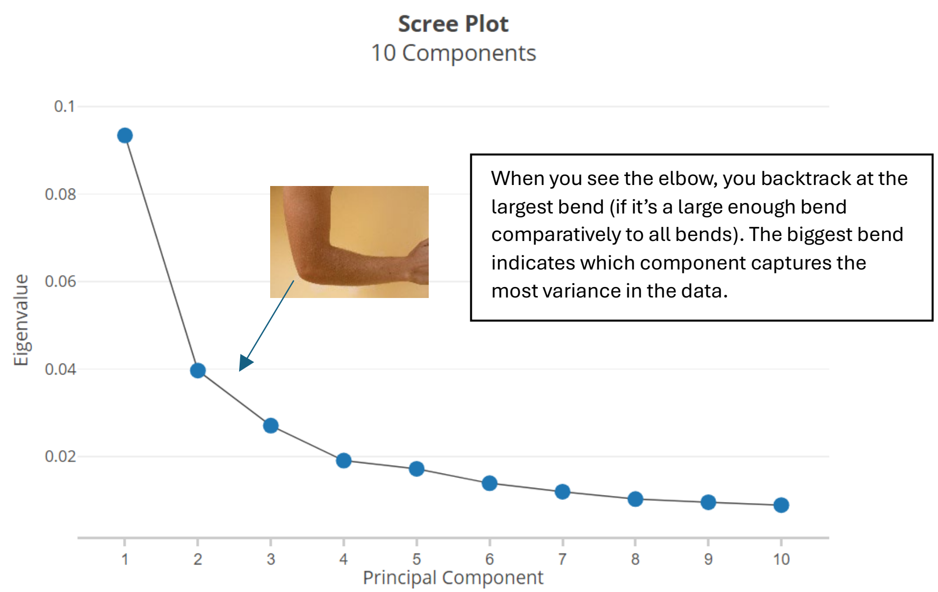
1. **Split** "majesti" into all possible 5-character sequences (substrings):
   a. "**majes**"
   b. "**ajest**"
   c. "**jesti**"
2. **Compare Substrings**: We compare these 5-character sequences with the 5-character sequences from another text.
   a. e.g., if the other treaty text also has "majesty", it would have the same substrings:
      i. "**majes**"
      ii. "**ajest**"
      iii. "**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.

Ultimately, 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:

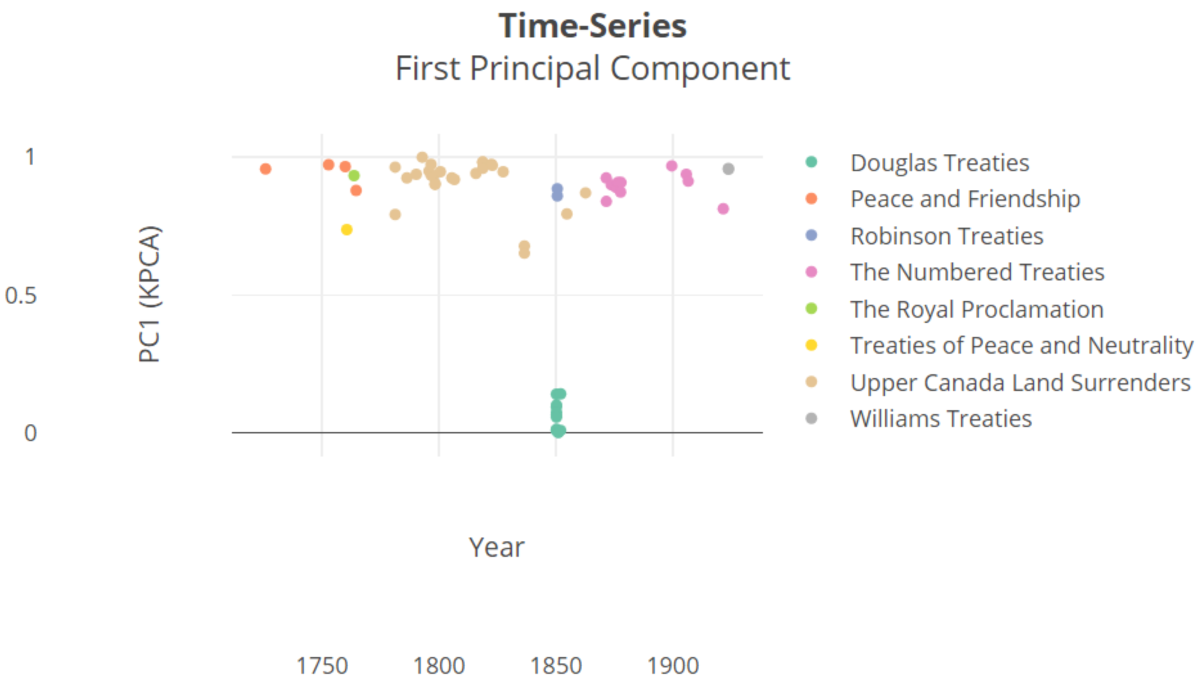
Arguably, we are looking for a **1st** and a **2nd component:**

- **Eigenvalues**: indicates the amount of variance in the data that is explained by its corresponding principal component.
  - Higher eigenvalues mean that the principal component captures more variance from the data.



Component 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).

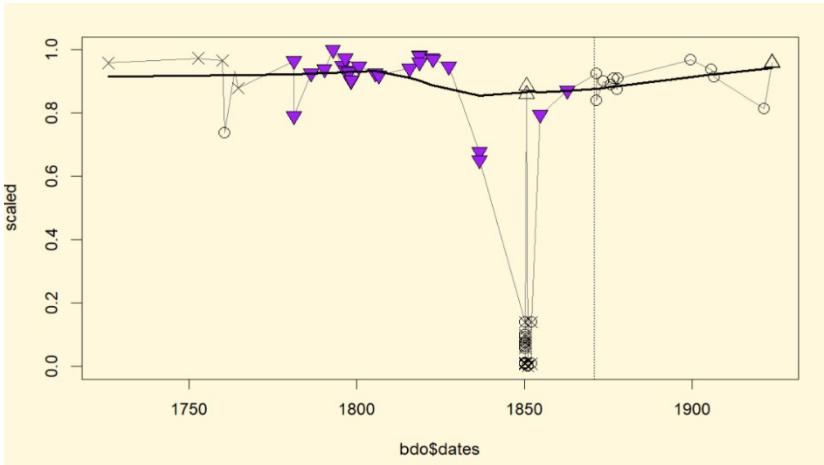


There's some validation of this "main thing" trend in work by **Feir et al. (2023)** who did sentiment analysis on a sample of Canadian treaty texts.

- *"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."*

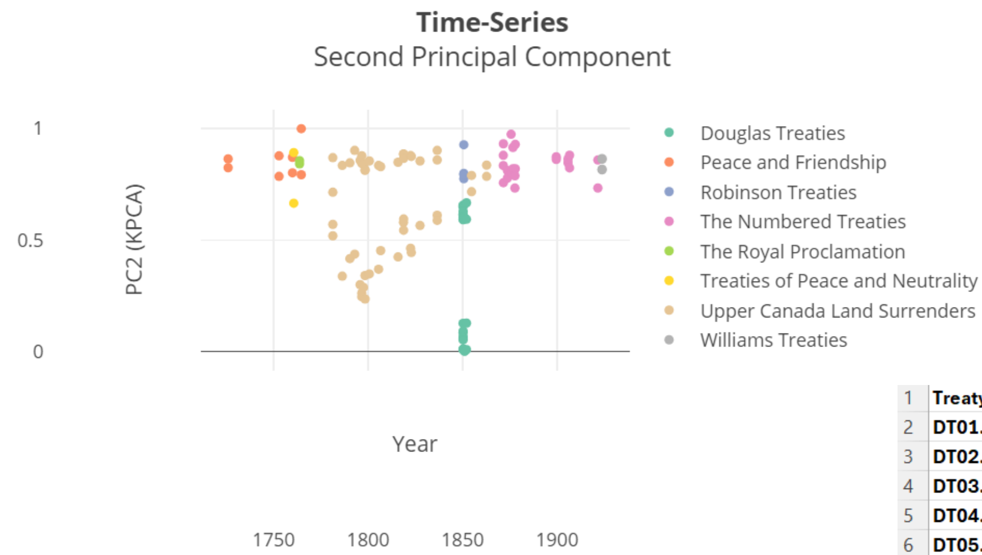
Arguably, when looking at word importance (below) we're seeing consistency in language, possibly due to the fact that Canada remained attached to the British (vs. American independence). This 1st component could reflect crown involvement ("surrend"ing to "majesti"; "becom"ing "subject"s; "commission"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

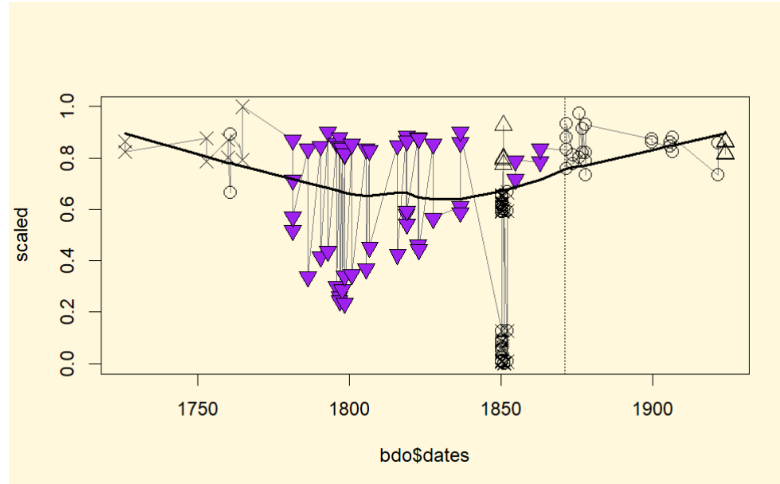
- This is our **Messy Graph 1** (mirrors Spirling's visual) with a trendline.
  - We'll be working on adjusting these graphs in a way that can clearly distinguish each individual treaty in a more visually appealing manner.



Component 2 Graph – the "secondary thing":

- 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.

1	Treaty	KPC1 Score	KPC2 Score
2	DT01.txt	-1.805939295	-0.757759895
3	DT02.txt	-1.071957768	-0.588192455
4	DT03.txt	-1.660291693	-0.542594024
5	DT04.txt	-2.114739514	-0.780190148
6	DT05.txt	-1.664173292	-0.479384933
7	DT06.txt	-2.316421396	-0.781109101
8	DT07.txt	-1.809882909	-0.829460649
9	DT08.txt	-2.0779433	-0.866227488
10	DT09.txt	-2.285623464	-0.864577987
11	DT10.txt	-2.698021539	-0.827409912
12	DT11.txt	-1.933846074	-0.271300189
13	DT12.txt	-2.94369106	-0.742106516
14	DT13.txt	-3.02916917	-0.72253333
15	PF01.txt	2.80444319	1.330688558
16	PF02.txt	2.783729121	1.350246077
17	PF03.txt	1.96052126	1.219802883
18	PF04.txt	2.035523878	1.763186146
19	RT01.txt	2.333115289	0.674901634
20	RT02.txt	2.06597273	0.373810102
21	NT1001.txt	2.077765088	2.137342434
22	NT101.txt	2.099417879	1.732886955
23	NT1101.txt	2.002294882	1.850719755
24	NT201.txt	2.297879254	1.38725355
25	NT301.txt	2.003271098	1.415544125
26	NT401.txt	2.03522479	1.159501644
27	NT501.txt	2.035583326	0.89999058
28	NT601.txt	1.893508784	1.035726057
29	NT701.txt	1.586139384	1.451852376
30	NT801.txt	1.136616258	2.510567299
31	NT901.txt	1.083016182	2.008435989
32	RP01.txt	0.814985275	1.056943555
33	TPN01.txt	-0.07808634	1.798424343
34	UCLS01.txt	0.180702587	-0.507366387
35	UCLS02.txt	1.205600699	-0.855465498
36	UCLS03.txt	0.856365642	-2.141594265
37	UCLS04.txt	0.843057714	-1.550028808
38	UCLS05.txt	1.151524218	-1.38706171
39	UCLS06.txt	0.727962902	-2.351893106
40	UCLS07.txt	0.505677978	-2.421121864
41	UCLS08.txt	0.21574232	-2.018717748
42	UCLS09.txt	0.594539368	-2.569337535
43	UCLS10.txt	0.246342411	-2.662392678
44	UCLS11.txt	-0.053964966	-2.712946623
45	UCLS12.txt	0.131702736	-1.877420133
46	UCLS13.txt	-0.110170069	-1.70523296
47	UCLS14.txt	-0.244868996	-1.079909157
48	UCLS15.txt	-0.197212933	-1.265845056
49	UCLS16.txt	-0.022725131	-0.120466124
50	UCLS17.txt	-0.133823501	-0.358277834
51	UCLS18.txt	-0.363767622	0.027117047
52	UCLS19.txt	-0.3697459	-0.897702799
53	UCLS20.txt	-0.492446814	-1.017909268
54	UCLS21.txt	-0.737304235	-0.120741353
55	UCLS22.txt	-2.59121508	2.027620146
56	UCLS23.txt	-2.85533955	2.354146889
57	UCLS24.txt	-2.021616009	1.566145754
58	UCLS25.txt	-1.62217179	1.92189358
59	WT01.txt	-1.145200535	1.790566166
60	WT02.txt	-1.2568663	1.828963885

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.

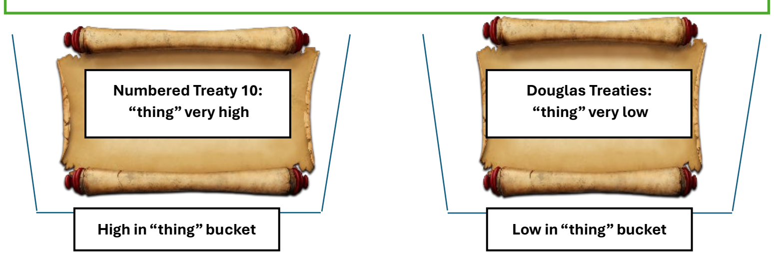
- Once we’ve stemmed, removed common words, etc. we feed them to two algorithms – **Random Forest** (Spirling) and **xgboost** (new age extremely powerful competition-winning predictive algorithm that uses Random Forest and **Gradient Descent**).

Importance Algorithms:

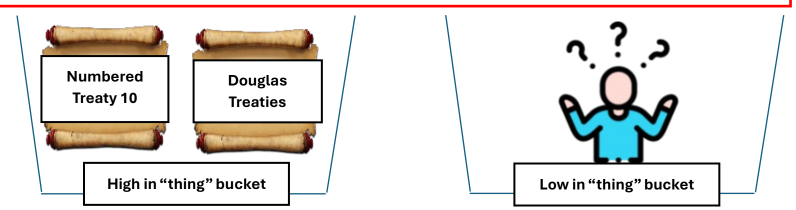
(1) **Random Forest:** you can think of the entire sample of the treaties as the “forest”. The algorithm plants 500-2000 “trees” at random points (word stems) in each treaty. Each tree makes a predictive “bucketing” decision (simplified below) extending throughout the entire corpus before the algorithm 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:



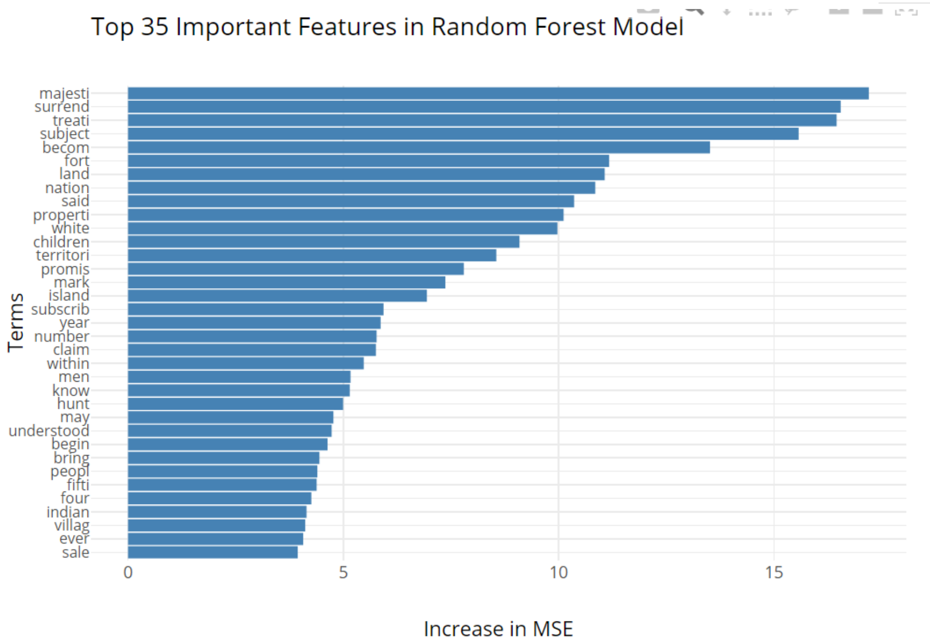
(2) XGBoost: Extreme Gradient Boosting builds a model in a stage-wise fashion, where each new tree corrects errors made by previous trees.

- 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.

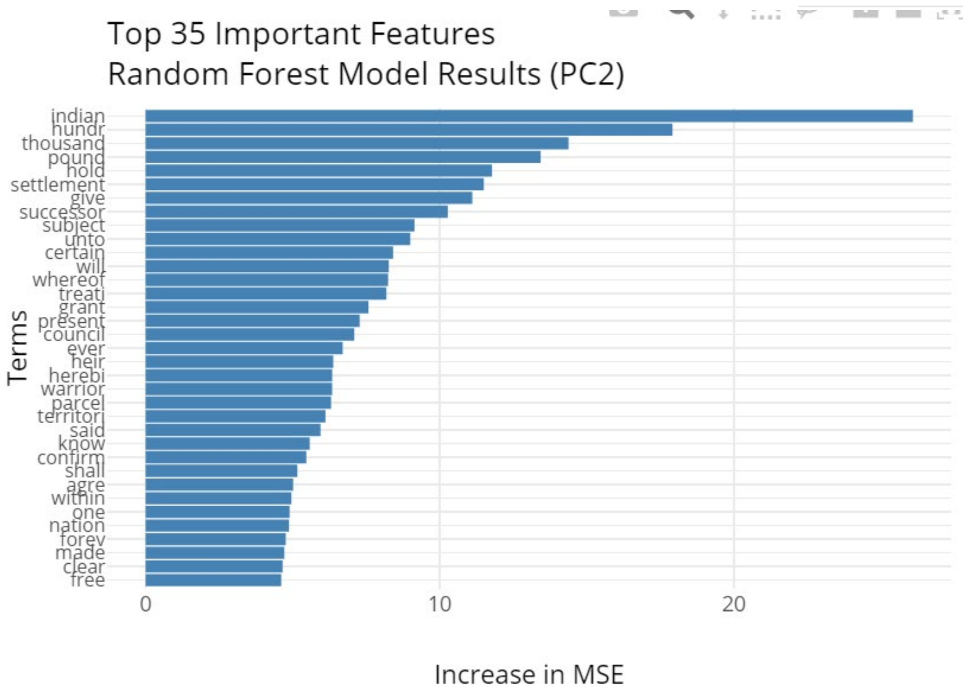
Algorithm Results – KPC1 ("main thing"):

- 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.



Random Forest - KPC1:	Term	Importance	xgboost - KPC1:	Feature	Gain	Cover	Frequency
	majesti	17.20697585		majesti	0.470726	0.041197	0.020045
	surrend	16.5538414		surrend	0.174665	0.038444	0.020045
	treati	16.45841218		subject	0.159209	0.017313	0.011136
	subject	15.57668265		treati	0.065014	0.021309	0.013363
	becom	13.51821743		advantag	0.045189	0.007103	0.011136
	fort	11.17504978		begin	0.019643	0.007458	0.008909
	land	11.07196381		becom	0.011008	0.028323	0.013363
	nation	10.85303403		among	0.010526	0.005416	0.006682
	said	10.3613294		acknowle	0.006334	0.006037	0.017817
	properti	10.11668294		west	0.004329	0.00728	0.011136
	white	9.973048806		lead	0.004017	0.003196	0.004454
	children	9.092199914		war	0.003254	0.001687	0.002227
	territori	8.554677878		abus	0.002759	0.001953	0.01559
	promis	7.799111152		island	0.002355	0.015182	0.020045
	mark	7.370902179		absolv	0.002324	0.005327	0.026726
	island	6.941061562		acadi	0.002164	0.115067	0.069042
	subscrib	5.934418155		arm	0.002066	0.007547	0.013363
	year	5.867650169		along	0.00179	0.002308	0.008909
	number	5.772806223		children	0.001785	0.004972	0.006682
	claim	5.757193419		april	0.001737	0.001598	0.002227
	within	5.475084359		deliveri	0.001254	0.001332	0.002227
	men	5.168293178		wit	0.000755	0.001598	0.002227
	know	5.150215297		includ	0.000609	0.003463	0.006682
	hunt	4.995431091		case	0.000518	0.011897	0.006682
	may	4.769946451		fifti	0.000449	0.001243	0.002227
	understood	4.729358811		divis	0.000439	0.003285	0.004454
	begin	4.634295618		confirm	0.000435	0.004084	0.004454
	bring	4.444877021		benefit	0.000343	0.003463	0.006682
	peopl	4.397882078		four	0.000307	0.012519	0.008909
	fifti	4.380869802		bring	0.0003	0.008435	0.004454
	four	4.257711054		albert	0.000296	0.000622	0.002227
	indian	4.146229586		abid	0.000294	0.002308	0.011136
	villag	4.115932613		fort	0.000277	0.020154	0.013363
	ever	4.068352044		accru	0.000274	0.001421	0.004454

Algorithm Results – KPC2 (“secondary thing”):



Random Forest - KPC2:	Term	Importance	xgboost - KPC2:	Feature	Gain	Cover	Frequency
	indian	26.08844314		indian	0.691937	0.075631	0.03856
	hundr	17.91526816		successor	0.056646	0.006577	0.005141
	thousand	14.38026684		parcel	0.053136	0.007465	0.007712
	pound	13.43267643		hundr	0.041199	0.026218	0.012853
	hold	11.77752796		angl	0.033272	0.002222	0.002571
	settlement	11.50018322		resid	0.026144	0.017686	0.017995
	give	11.10844294		agre	0.018576	0.00471	0.007712
	successor	10.27532906		confirm	0.011996	0.008798	0.007712
	subject	9.143401219		herebi	0.009707	0.003644	0.005141
	unto	8.999725288		come	0.007571	0.009509	0.007712
	certain	8.416015242		present	0.006358	0.008443	0.007712
	will	8.270434861		within	0.006091	0.003733	0.005141
	whereof	8.246522196		begin	0.006068	0.01502	0.015424
	treati	8.188283219		degre	0.005024	0.002666	0.005141
	grant	7.579728496		adjut	0.002878	0.001955	0.002571
	present	7.279748213		thousand	0.002044	0.021685	0.017995
	council	7.09764383		affair	0.001618	0.001777	0.012853
	ever	6.703050922		abid	0.001605	0.004621	0.048843
	heir	6.381392474		whereof	0.001498	0.003999	0.005141
	herebi	6.345505471		enabl	0.001371	0.01182	0.015424
	warrior	6.3439248		unto	0.001184	0.001511	0.002571
	parcel	6.30755173		civil	0.001156	0.001866	0.002571
	territori	6.117104943		assist	0.001104	0.021507	0.030848
	said	5.947730047		certain	0.001094	0.00391	0.007712
	know	5.582131019		articl	0.00106	0.002844	0.005141
	confirm	5.466342867		abovenan	0.001059	0.003377	0.025707
	shall	5.162242571		five	0.001007	0.001333	0.002571
	agre	5.0191569		heir	0.000971	0.00871	0.005141
	within	4.959517698		majesti	0.000935	0.003288	0.005141
	one	4.898348406		alexand	0.000879	0.123889	0.07455
	nation	4.875862893		caus	0.000687	0.001333	0.002571
	forev	4.76695098		northwest	0.000635	0.0016	0.002571
	made	4.718605158		acknowle	0.000578	0.005688	0.012853
	clear	4.658949901		abandon	0.000504	0.0008	0.005141

- 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.

KPC1:

Correlations - KPC1:	Positive Term	Frequency	Correlation		Negative Term	Frequency	Correlation
	majesti	42	0.75032493		lie	33	-0.5484079
	subject	23	0.60185895		consent	30	-0.5536586
	promis	28	0.5733319		former	23	-0.5602938
	subscrib	16	0.55606724		surrend	40	-0.5675753
	indian	45	0.55445765		small	28	-0.5873255
	observ	17	0.55082493		condit	26	-0.5882028
	conduct	15	0.54632025		deed	19	-0.5887561
	treati	22	0.54610993		except	32	-0.5893003
	conclud	21	0.53462359		proper	28	-0.6145289
	gracious	16	0.5209063		eight	35	-0.6293328
	obtain	16	0.50829347		committe	21	-0.6303388
	taken	14	0.50816445		howev	25	-0.6330398
	cede	17	0.50321271		token	14	-0.6399935
	perform	15	0.50107137		agent	24	-0.6422898
	stipul	13	0.47663875		dougla	13	-0.645177
	deliber	11	0.47540714		sale	25	-0.6457287
	right	33	0.474763		deputi	29	-0.6526822
	bounti	11	0.4697672		understand	16	-0.6565673
	deal	11	0.46719		kept	19	-0.6577462
	behav	11	0.46600849		field	16	-0.6721316
	strict	17	0.4657789		unoccupi	13	-0.6742749
	proport	14	0.46491511		villag	21	-0.6764396
	behaviour	10	0.45646489		fisheri	17	-0.6784546
	immigr	10	0.44796383		white	32	-0.6812513
	obey	10	0.44660683		survey	25	-0.6866723
	school	10	0.4369626		understood	22	-0.6930203
	solemn	15	0.42870286		properti	39	-0.7057746
	assur	11	0.42469754		children	30	-0.7096801
	accept	15	0.42457366		land	53	-0.7353574
	dominion	14	0.36452071		becom	32	-0.7443686

KPC2:

Correlations - KPC2:	Positive Term	Frequency	Correlation		Negative Term	Frequency	Correlation
	indian	45	0.73625333		appurten	14	-0.4066091
	treati	22	0.6003013		languag	9	-0.4110095
	govern	19	0.57026647		certain	32	-0.4348608
	commission	13	0.55899732		absolv	4	-0.4351349
	council	23	0.55440436		princip	26	-0.4422135
	subject	23	0.54270485		dispos	21	-0.4457174
	proceed	15	0.53347524		consider	36	-0.4691596
	punish	13	0.53154849		execut	21	-0.4718961
	obtain	16	0.53098622		deputi	29	-0.4720265
	travel	12	0.52135228		deliveri	9	-0.4774989
	conduct	15	0.51494313		clear	12	-0.4864373
	upon	27	0.5125373		seven	24	-0.4885747
	meet	12	0.50906024		present	42	-0.4892227
	minist	10	0.50759944		rehears	5	-0.5042469
	territori	18	0.50437313		situat	38	-0.5051159
	infring	12	0.49922643		forev	30	-0.5096283
	assum	13	0.49919554		descend	10	-0.5181152
	matter	13	0.49362828		begin	27	-0.5320179
	defin	12	0.49314839		pretend	6	-0.5355966
	assur	11	0.49196637		renounc	6	-0.5355966
	report	10	0.49152494		instrument	13	-0.5512711
	possibl	11	0.49048868		grant	29	-0.5518602
	interfer	14	0.48868337		whereof	46	-0.5623836
	dominion	14	0.48852598		emolu	7	-0.5631035
	request	18	0.48746437		greet	8	-0.5671626
	offend	13	0.48725998		parcel	19	-0.5720966
	negoti	13	0.48190129		warrior	10	-0.5812067
	advis	10	0.47961758		nation	26	-0.6003893
	appoint	18	0.47625332		receipt	14	-0.6115586
	notifi	13	0.47618779		successor	34	-0.6136265
	observ	17	0.47335551		confirm	27	-0.6140772
	school	10	0.46710603		heir	29	-0.6252556
	perform	15	0.46287027		unto	27	-0.6782567
	requir	16	0.46188875		hundr	54	-0.6921874

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.