My graduate project analyzed treaties signed with Native American tribes on the working assumption that variations in word frequency were predictable depending on certain categorical variables. From the website [www.firstpeople.us](http://www.firstpeople.us) I obtained the text of 297 treaties signed and ratified by Congress, as well as 4 treaties that were signed and unratified, 10 with Canada, 10 under the Articles of Confederation, and 9 with the Confederate States of America. An estimated 78 treaties were not included because they were not readily available in an online format.

I began by saving each treaty into a folder using a format based on the common name for each of the tribes with which the treaty was signed (taking care to use the same name each time for a tribe historically identified by one or more names, or failing that to write instructions in my code to assimilate the treaties), the year the treaty was signed, and an additional code for treaties not ratified under the Constitution to indicate a special status for the filenames.

My Python script then began by building a dictionary with the names of tribes, time periods, and categories based on the polity signing the treaties with the tribes as keys, and another dictionary mapping words to frequencies as values, to represent word-frequency vectors indexed by the invariant order in which words were encountered when reading all the files. The names of tribes, time periods, and categories such as “CSA” (to indicate a treaty with the Confederacy) or “USA” (for a treaty ratified by the United States under the Constitution) thus became the first group of categorical variables.

The years were grouped into 6 time periods based on historically plausible pivot points: the period between the American Revolution and the ratification of the Constitution; the period between the Constitution and the War of 1812; the period between the War of 1812 and 1831, when Chief Justice John Marshall ruled that the relationship of a tribe to the United States was like that of “a ward to his guardian” and Congress passed legislation to similar effect; between 1831 and the Civil War; between the Civil War and the end of treaty-making in the United States; and a final period when Canada was still making treaties but the United States was not. Later a dichotomy between pre-1831 treaties and post-1831 treaties was considered as well.

From the tribal affiliations a number of other categorical variables were generated. Each tribe was mapped to a language family and to a region (making the simplifying assumption that each tribe could be identified with a single region). Treaties that included tribes that now have reservations in Montana were grouped together, and compared with treaties that did not. Treaties that included the so-called “Five Civilized Tribes” (Cherokee, Seminole, Creek, Chickasaw, and Choctaw) were compared with treaties that did not. And treaties with tribes identified (using a Wikipedia page that cited anthropological literature published by Oxford University and Ohio State University; the Library of Congress page on kinship patterns; and powwows.com, a site maintained by the Chickasaw tribal government) as matrilineal were compared to other societies on the assumption that such societies were patrilineal.

The least reliable of these identifications were region and kinship patterns (matrilineal vs. patrilineal) due to the use of simplifying, and potentially misleading, assumptions, and reliance on unreliable literature. However, I proceeded on the assumption that the assignments would be accurate enough to reveal similarities and differences that might at least be suggestive.

For each treaty, then, based on the file name, a list was generated of categories to which that treaty applied, including that of one or more tribes, language families, and regions, as well as the binary or categorical variables identified above, and a word-frequency vector was updated for each category as the treaty was read. A separate vector for the treaty itself was also maintained in order to simplify computation later.

To remove uninformative data, I removed two-letter words (other than “ax” and “ox” which were considered potentially meaningful). I also imported a list of stop-words, supplemented with a list of my own identified via preliminary searches. Words with capital letters that were not at the beginning of sentences were also removed in order to filter out most of the names of signatories. However, I also maintained a list of words not to be removed even if found on the imported stop-words list or with a capital letter, including the tribe and state names, variations on them, the names of other political entities, etc., and updated this list by adding to it when words were added as dictionary keys.

My primary goal was to identify potential avenues for future analysis by turning up patterns that appeared interesting, and to test or validate the presumption that based on the identified categories, word-frequency vectors would differ. I was also interested in general information about all the treaties. To that end, I used three primary tools. Using Python’s linear algebra capabilities I wrote a function giving the cosine measure of similarity computed by taking the dot product of two normalized word-frequency vectors; since all values were non-negative this gave a number between 0 (no similarity) and 1 (maximal similarity) between two categories based on their respective word-frequency vectors. Another function took a list of categories and uniquely assigned a given treaty to one of them based on a multinomial naïve Bayes prediction function. I tested this function on treaties by updating (temporarily) the word-frequency vector for the target category by removing the word counts from the treaty being assigned, and to make the test realistic I only tested it for categories for which at least two treaties could be uniquely assigned. To deal with the large list of words (6,906 distinct words across all treaties) and large number of zero values, and to facilitate comparisons across relatively large numbers of categories, the additive smoothing technique was used when estimating a conditional probabiliity of a given word showing up if the treaty it was in was assigned to a particular category. Based on trial and error mostly with the language family assignments, a smoothing parameter of (1/k), where k is the number of categories, was selected. Given a list of categories, every treaty assigned uniquely to one of them could be tested and the percentage classified correctly as well as the specific misclassifications made could be determined. Finally, I wrote functions that listed the most common values for a given category and functions that listed the biggest disparities in relative frequency for a given word from one category to another. These summary statistics were revealing about the variable importance of certain concepts to the treaty-makers, and gave concrete substance to measures of similarity or dissimilarity and to the output of the classification algorithm, although the interpretation of them often required going back to the treaties in which a particular word appeared for context.

I also did preliminary work on classifying the words in the treaties into a number of categories, maintaining a list in a text file that included about 200 of the most frequent words and other words that were obvious candidates when making the list, with each one subjectively assigned to one of 15 categories (e.g., “financial,” “geographical,” or “legal”). Python read in this text file and maintained a dictionary of words by category, for each of the categorical variables. Although only roughly 1/3 of all the words for a given category were assigned to one of these groups, the relatively frequencies of certain sorts of words from one category to another were also revealing.

Across all treaties, geographical terms apparently predominated, accounting for about 30% of all classified words, compared to about 14% for “legal,” the category with the second-greatest frequency. The most frequent word was “dollars” (or the “$” sign which my script interpreted as “dollars”), with 3,076 appearances compared with 2,382 for “river” which was the second-most-frequent word. “River” was the most common word in 7 of 9 treaties with tribes with Montana reservations; “dollars” was most frequent in treaties with the Chippewa and “confederated” most frequent in the single treaty with the Confederated Salish and Kootenai Tribes. “River” was one of the 8 most frequent terms for every Montana tribe, while only 4 of the Montana tribes—the Sioux, Crow, Chippewa, and Gros Ventre—had “dollars” among the ten most frequent words.

Treaties with Montana tribes apparently had about twice as many race-related terms, but barely ¼ as many terms related to slavery, as treaties with other tribes. The cosine of the angle between the word-frequency vector for Montana tribes and other tribes was .896, and the multinomial naïve Bayes prediction function accurately classified 80.5% (58 of 72) of treaties with Montana tribes as such, and accurately classified 73.5% (197 of 268) of treaties not including Montana tribes as such.

The word “lakes” occurred 54 times in treaties with Montana tribes, compared with 5 in treaties with tribes from outside Montana, a 33:1 disparity in relative frequency. However, the words “slaves,” “courts,” “criminal,” “love,” and “elections” never appeared in treaties with Montana tribes, but appeared 72 times, 48 times, 28 times, 28 times, and 18 times respectively in treaties with other tribes.

Of Montana tribes, the cosine measure of similarity between the Cheyenne treaties and the Crow treaties was highest at .902 while that between the Confederated Salish-Kootenai tribes and the Assiniboine was lowest at .246. Similarity between the treaties with either the Cheyenne or the Sioux and any other Montana tribe was higher than .5, while the CSKT had a similarity measure lower than .5 with every other Montana tribe besides the Cheyenne and the Sioux, and the lowest similarity with every other Montana tribe.

Of the 4 Montana tribes with at least 2 treaties not shared with another tribe, a total of 43 of 66 of their treaties were accurately classified by the multinomial naïve Bayes prediction function. A large majority (32 of 40) of the treaties with the Chippewa were accurately classified as such, while 7 were misclassified as with the Crow and 1 was misclassified as with the Cheyenne. A slim majority (9 of 17) of treaties with the Sioux were accurately classified, while 7 were misclassified as with the Crow and 1 was misclassified as with the Cheyenne. 2 of 3 treaties with the Crow were misclassified as with the Cheyenne, and 5 of 6 treaties with the Cheyenne were misclassified as with the Crow.

Compared with other Montana tribes, treaties with the Sioux apparently had about 7 times as many military-related terms, while treaties with the Chippewa had nearly twice as many legal terms and less than ¼ as many military terms. Political terms occurred with about 1/3 as much frequency in treaties with the Blackfeet and Cheyenne as in treaties with other Montana tribes. Terms relating to farming appear about 1/5 as often in treaties with the Chippewa and Cree as in treaties with other Montana tribes.

The multinomial naïve Bayes prediction function had an overall success rate of 66.1% in classifying treaties among 7 selected high-population tribes (Navajo, Apache, Choctaw, Cherokee, Sioux, Iroquois, and Chippewa). The highest rate of accurate classification was 86.5% for the Chippewa, which I suspect is due to a number of treaties being signed with different Chippewa bands in rapid succession, all using the same template and linguistic conventions. 26.7% of treaties with the Choctaw were accurately classified, with a plurality misclassified as with the Cherokee. In general, when treaties were misclassified they were classified as with a tribe culturally and historically similar (in my opinion) to the target tribe; however, 9 of the 18 treaties with the Sioux were misclassified as with the Apache or Navajo (while the others were correctly classified).

The multinomial naïve Bayes prediction function accurately classified 84.2% of U.S. treaties by region, with success rates ranging from 90.3% for Eastern treaties to 79.2% for treaties in the Northwest (with 3 of 24 misclassified as Plains and 2 misclassified as Southwest).

Religious language was apparently highest in the Northwest and Southwest (about 3 times as prevalent as in the Plains) and least prevalent in the East (about 3 times less frequent than on the Plains. Terms related to slavery were about 3 times as prevalent in the Northwest and about 6 times as prevalent in the South as anywhere else in the United States, while apparently completely absent from Canadian treaties. Racial terms in general (not including those related to slavery) were most prevalent in the Southwest, occurring about 2.5 times as often there as anywhere else. Legal terms were also particularly common in the Southwest and in Canada, while financial terms were the least common in these regions. Political and geographical terms were most common in Canada, while geographical terms were least common in the Southwest. Terms related to industry were most common in the Southwest; terms related to time were most common in the Northwest. Terms related to the military and commerce were relatively uncommon in the northwest and less common still in Canada. Family-related terms were about half as prevalent in the East as anywhere else.

The cosine measure of similarity for language families ranged from .974 for the Salishan and Chimakuan languages to .270 for the Kootenai and Natchez language isolates. The multinomial naive Bayes prediction function accurately classified 63.1% of treaties uniquely associated with 1 of 8 common language families, with a success rate ranging from 87.5% of Salishan languages correctly classified (with 1 of 8 misclassified as Algonquian) to 30.0% of Uto-Aztecan languages correctly classified (with 5 of 10 misclassified as Athabaskan and 2 of 10 misclassified as Caddoan).

The cosine measure of similarity appears to militate against confirmation of the hypothesis mentioned in my topic proposal, that 1831 would represent a pivot point at which the language of the treaties would change. The cosine of the angle between the word frequency vectors from the 1812-1831 time period and the 1831-1860 time period was .829, higher for example than the similarity measured between treaties with the Blackfeet and treaties with the Sioux. Even if we restrict our observations to the “Five Civilized Tribes” that were the principal direct targets of removal efforts during this period, the similarity between treaties from each of these time periods was .800. Similarly high measures of similarity were found for treaties from before and after the War of 1812 and before and after the Civil War, and these results were not sensitive to adjustment of the endpoints (by taking the year 1832 instead of 1831 or 1860 instead of 1865, for example). The only truly large dissimilarity by this measure came when comparing treaties from before and after the ratification of the Constitution, which had a similarity measure of .641.

However, the multinomial naïve Bayes prediction function did seem to suggest that the language of a treaty depended on the time period in which it was written, accurately classifying 247 of 330 treaties into 1 of 6 time periods, and classifying 72 of the 83 misclassified treaties into one of the adjoining time periods.

Treaties under the Articles of Confederation apparently had about 5 times as many terms related to commerce and slavery, about twice as many military-related terms, and about 1/3 as many terms related to religion (relative to their total length) as treaties under the Constitution. Pre-1831 treaties had about 3 times as many commercial terms, twice as many military-related terms, and 1.5 times as many family-related terms as post-1831 treaties; post-1831 treaties had about 3 times as many terms related to religion and about twice as many related to each of the categories slavery, agriculture, and finance.

Compared with treaties with the U.S. from the same time period, Confederate treaties had about 3 times as much race-related language, twice as much language about slavery, and 1.5 times as much financial language. Confederate treaties and American treaties from the same time period had a cosine similarity measure of .711, and 100% of treaties from this period were accurately classified by the multinomial naïve Bayes prediction function as American or Confederate.

The cosine similarity between treaties with the “5 civilized tribes” and treaties with other tribes was .833. The multinomial naïve Bayes prediction function classified 91.3% of treaties with the “5 civilized tribes” and 95.9% of treaties with other tribes as such.

The cosine similarity measure between treaties with matrilineal tribes and treaties with patrilineal tribes was .885. The classification algorithm accurately classified 80.3% of treaties with matrilineal tribes and 91.7% of treaties with patrilineal tribes. In treaties with patrilineal tribes, terms related to commerce were apparently about twice as prevalent while religious language was about 1.7 times as prevalent; language about settlement and agriculture were also more frequent. Terms related to slavery showed up about 4 times as often in treaties with matrilineal tribes.

Computer-assisted analysis of variations among treaties unsurprisingly suggests that variation depends in part on the identified variables and further appears to indicate (based on low cosine similarity measures and high rates of successful classfication) that besides individual tribal affiliation, region and polity (e.g., U.S. vs. Confederacy or Constitution vs. Articles of Confederation) may be particularly good predictive variables, while other variables such as language family may be less useful in this context. It also has helped us see how in particular one category of treaties might differ from another. Further exploration is sure to reveal more, and of course the simple techniques used can be applied to other treaties and documents.