Clustering Text

Clustering requires vectors..

What we did last time

Sepal Length	Sepal Width	Petal Length	Petal Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2

Distance is easy, when you know how to "describe" your response numerically.

But some things are not easily described numerically..

Text is not a vector…

How do I make the following a vector?

“When in the Course of human events it becomes necessary for one people to dissolve the political bands which have connected them with another and to assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation.

We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. — That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, — That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed, will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn that mankind are more disposed to suffer, while evils are sufferable than to right themselves by abolishing the forms to which they are accustomed.”

What are we looking for?

There are many ways we can “break up” a sentence automatically.

1. Bag of words.
   - Treat the document as just a bunch of words -> analyze by frequency.
2. Sentiment
   - Try to find certain words that have a specific meaning. “positive”, “negative”, “angry”, etc.
3. Grammar
   - Ideal but very hard.

The first two are what we are going to focus on, but ideally we should express grammar too (WAY OUTSIDE OF THE SCOPE OF THIS COURSE).

Let’s eat grandma.

Let’s eat, grandma.

Bag of words

Review:

1. We simply count the number of times a word appears in a document.
2. Comparing documents from a cosine distance is easy, other methodologies more difficult -> need to normalize the number of words per document
3. Let’s look at file 1 on Moodel

Sentiment Analysis

Sentiment are words that express a specific idea.

To do a sentiment analysis sentiment databases, we first need to have a sentiment database:

Two “flavors”

- Bag of Word Flavor
- Numerical Value to express opposites for an idea love +1, hate -1, joy +1, sad -1, ebullient +2 ect.

Let’s look at the code 2:

Questions

Now we have started clustering REAL DATA, but it is text. What do we need to do to think about mixing numeric and text data?

What should we do when we have a ton of data?