Texts as Data:

A Short Introduction to Vectors, Matrices, Pandas

What is text?

- Stream of characters
- Means little to a computer
- We need to give the computer some idea of the relationships we want to track
- The typical way to do this is with vectors and matrices

First decision: which relationships?

- We tend to think of texts in words, documents, and, sometimes, corpora, i.e., collections of related texts
- So we could have the following relationships
  - Word to word
  - Document to document
  - Corpus to corpus
  - Word to document
  - Word to corpus
  - Document to corpus

Word to document relationship

- Relationships:
  - Is word in document (binary/incidence: 0 or 1)?
  - How many in document (frequency: integers)?
- The result is a term-document vector
- A vector is simply a series of values related to each other somehow

Term-document vector

- Document 1: “the dog bit the man”
- Binary term-document vector:

the	dog	bit	man	bat	hit	ball
1	1	1	1	0	0	0

- Frequency term-document vector:

the	dog	bit	man	bat	hit	ball
2	1	1	1	0	0	0

Term-document matrix

- Document 2: “the bat hit the ball”

the	dog	bit	man	bat	hit	ball
1	0	0	0	1	1	1

the	dog	bit	man	bat	hit	ball
2	0	0	0	1	1	1

- Term-document matrix:

	the	dog	bit	man	bat	hit	ball
Doc 1	1	1	1	1	0	0	0
Doc 2	1	0	0	0	1	1	1

	the	dog	bit	man	bat	hit	ball
Doc 1	2	1	1	1	0	0	0
Doc 2	2	0	0	0	1	1	1

Term-document matrix

- Used to compare documents with each other
  - e.g., authorship attribution, genre identification, document topic recognition
- Assumes independence of the entities
  - i.e., the occurrence of each entity is independent of the occurrence of any other entity (Naive Bayes)
- For language, this is a bad assumption
  - the occurrence of a word will depend heavily on the words around it

Word to word relationships

- Typical question: How often do two words occur together in the same context (span, sentence, paragraph, document, etc.)?
- E.g., how often in the same sentence?

Binary word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the							
dog							
bit							
man							
bat							
hit							
ball							

Binary word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	1	1	1	1	1	1	1
dog	1	0	1	1	0	0	0
bit	1	1	0	1	0	0	0
man	1	1	1	0	0	0	0
bat	1	0	0	0	0	1	1
hit	1	0	0	0	1	0	1
ball	1	0	0	0	1	1	0

Frequency word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the							
dog							
bit							
man							
bat							
hit							
ball							

Frequency word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	1						
dog							
bit							
man							
bat							
hit							
ball							

Frequency word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	1	2					
dog	2						
bit							
man							
bat							
hit							
ball							

Frequency word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	1	2	2				
dog	2						
bit	2						
man							
bat							
hit							
ball							

Frequency word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	2	2	2				
dog	2						
bit	2						
man							
bat							
hit							
ball							

Frequency word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	2	2	2	2			
dog	2						
bit	2						
man	2						
bat							
hit							
ball							

Frequency word-word in sentence

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	4	2	2	2	2	2	2
dog	2	0	1	1	0	0	0
bit	2	1	0	1	0	0	0
man	2	1	1	0	0	0	0
bat	2	0	0	0	0	1	1
hit	2	0	0	0	1	0	1
ball	2	0	0	0	1	1	0