PREVIEW OF TEXT ANALYSIS

· *Jesse Lecy* ·

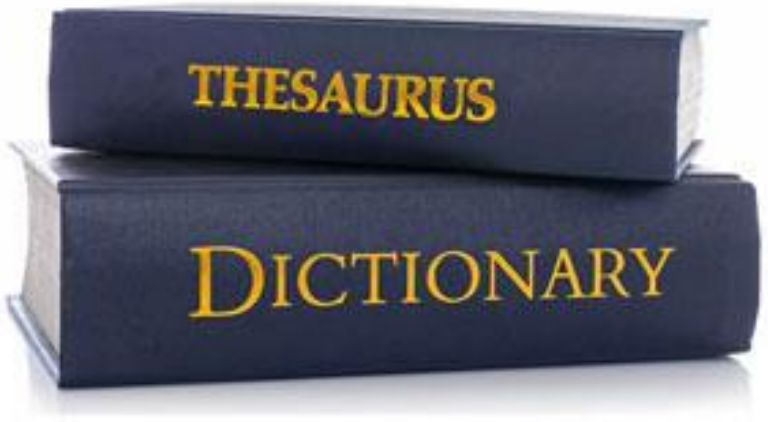
Text as Data

1. pre-processing
2. tokenization
3. feature selection
4. modeling

THE CORPORATION'S SPECIFIC PURPOSE IS TO SUPPORTS AFFORDABLE HOUSING, COMMUNITY DEVELOPMENT AND ECONOMIC DEVELOPMENT OF THE CITY AND COUNTY OF SAN FRANCISCO'S ECONOMICALLY DISADVANTAGED INDIVIDUALS AND COMMUNITIES, BY LENDING TO, INVESTING IN, AND DIRECTLY ACQUIRING SUCH AFFORDABLE HOUSING AND RELATED COMMUNITY DEVELOPMENT REAL ESTATE ASSETS.

~~the~~ corporation specific purpose ~~is to~~ support AFFORDABLE_HOUSING, community development ~~and~~ ECONOMIC_DEVELOPMENT ~~of the~~ city and county of SAN_FRANCISCO economically disadvantaged individuals and communities by lending ~~to~~ investing ~~in and~~ directly acquiring ~~such~~ AFFORDABLE_HOUSING ~~and~~ related community development REAL_ESTATE assets

1. Remove punctuation
2. Delete words with little information value ("stop words" in quanteda)
3. Identify compound constructs (apply "dictionary")



DISAMBIGUATION

Examples of N-GRAMS

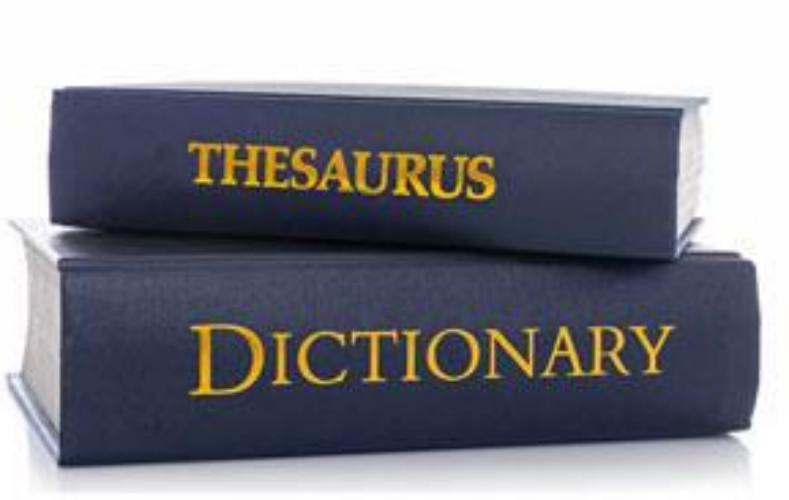
two-word or three-word groups that map to a single concept

George W. Bush

George Bush Jr.

President Bush

→ GW_BUSH



DISAMBIGUATION

Other mappings that you might want to do within the dictionary to disambiguate the term “New York” using context.

New York City → PLACE

THING ← New York University

New Yorker → STATE OF MIND

STATE ← New York (ambiguous – can also be the city)

STEMMING

LEND	RELATE
LEND~~ing~~	RELATE~~d~~

Convert all words to their respective **word roots** to standardize the data.

~~the~~ corporation specific purpose ~~is to~~ support AFFORDABLE_HOUSING, community development ~~and~~ ECONOMIC_DEVELOPMENT ~~of the~~ city and county of SAN_FRANCISCO economically disadvantaged individuals and communities by lending ~~to~~ investing ~~in and~~ directly acquiring ~~such~~ AFFORDABLE_HOUSING ~~and~~ related community development REAL_ESTATE assets

1. Remove punctuation
2. Delete words with little information value ("stop words" in quanteda)
3. Identify compound constructs (apply "dictionary")

DOCUMENT FREQUENCY MATRIX (DFM):

final output of pre-processing steps in quanteda

Terms	Documents													
	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	M 12	M 13	M 14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0