Text Mining

When the Words Become the Variables

What is Text Mining?

How is Text Mining Different from Data Analytics?

4ᵀᴴ **COFFEE**

What is Data?
What is Text Data?

● Types of Data

1) Record
2) HTML/Web
3) JSON
4) Network
5) Tweets
6) etc. etc. ...

Is text data "different" from number data?

Can text data be represented as numerical data?

Structured Data

| 0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Unstructured Data

Converting Documents to Matrices (DTM)

1) Can text be represented with numbers? Yes
2) What is a DTM? (What is a TDM?)
3) What is "frequency", vs. "TF-IDF", vs. "binary"?

	intelligent	applications	creates	business	processes	bots	are	i	do	intelligence
Doc 1	2	1	1	1	1	0	0	0	0	0
Doc 2	1	1	0	0	0	1	1	0	0	0
Doc 3	0	0	0	1	0	0	0	1	1	1

4TH COFFEE
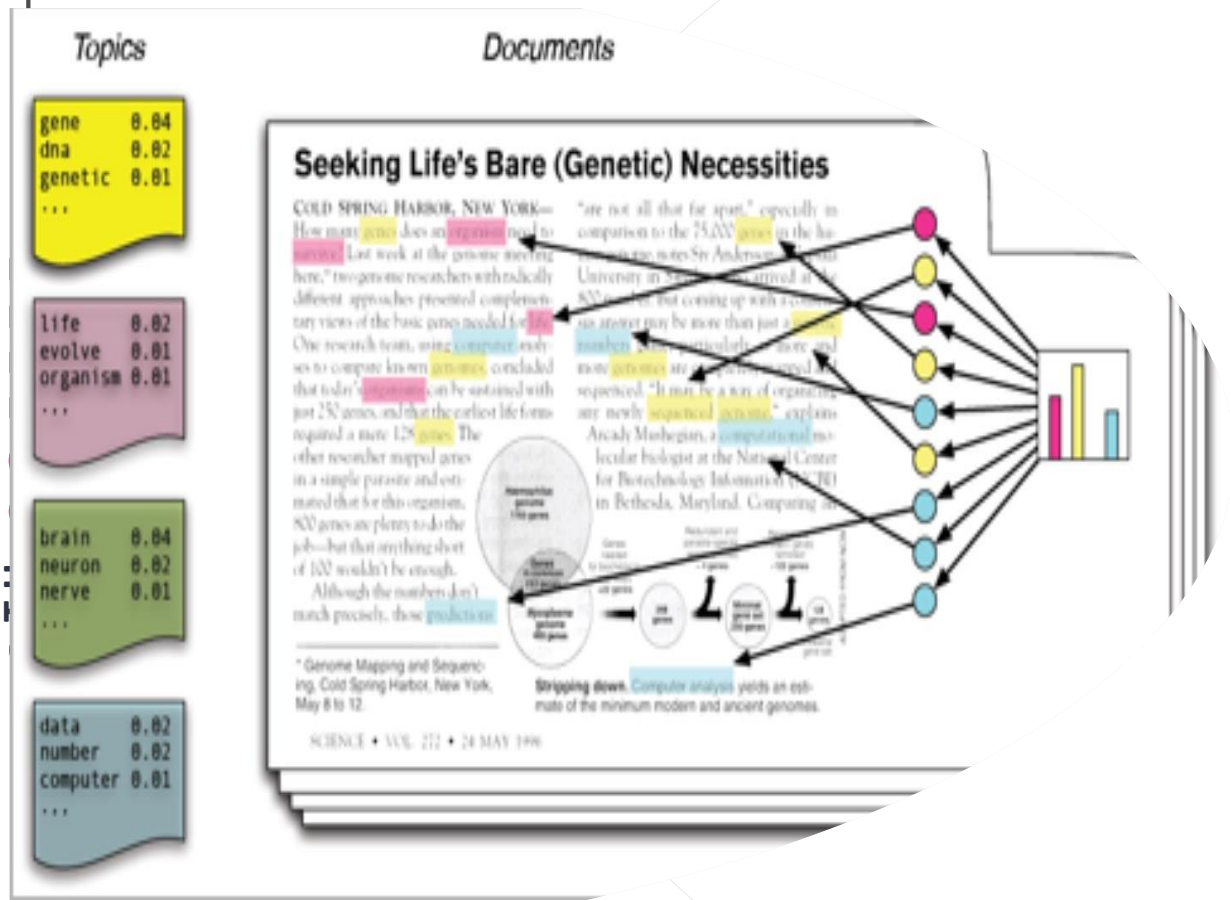
What is Topic Modeling?

What is a document?
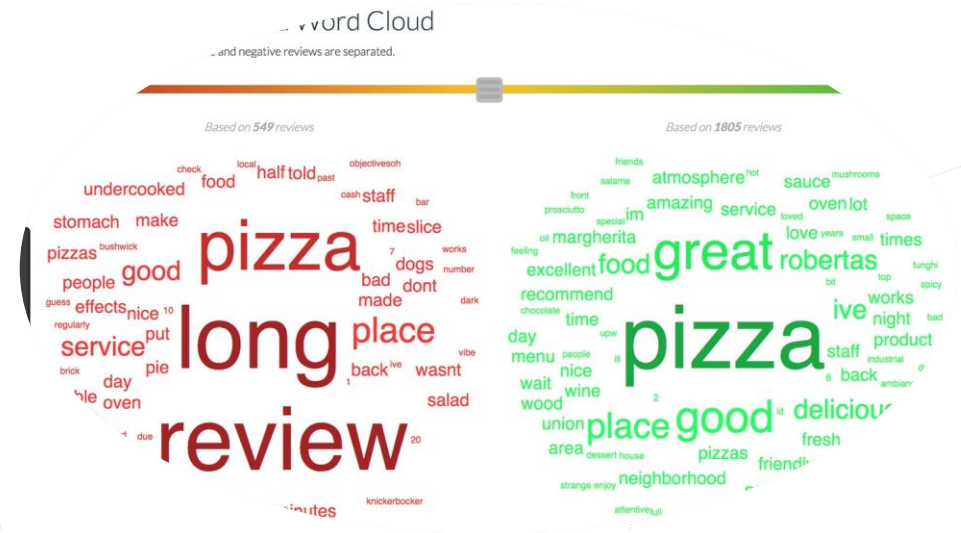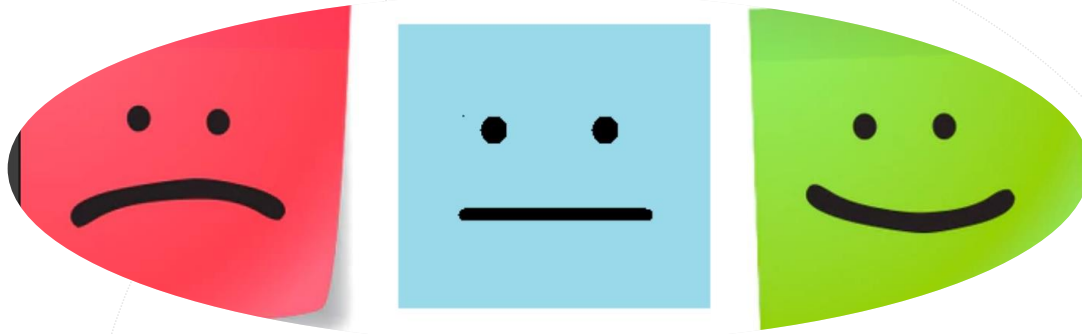
What is a Corpus?

Do documents have "topics"?

What is a topic distribution?

How can you determine the topic(s) of a document?

Word Cloud

... and negative reviews are separated.

Based on **549** reviews · Based on **1805** reviews

SENTIMENT ANALYSIS

Can Data Analytics be used to understand sentiment?

Can models be created (trained) to predict sentiment?

What are some "supervised" methods that can be used?

Can unsupervised methods be used?

 - How?

4TH COFFEE

Similarity Distance Measure
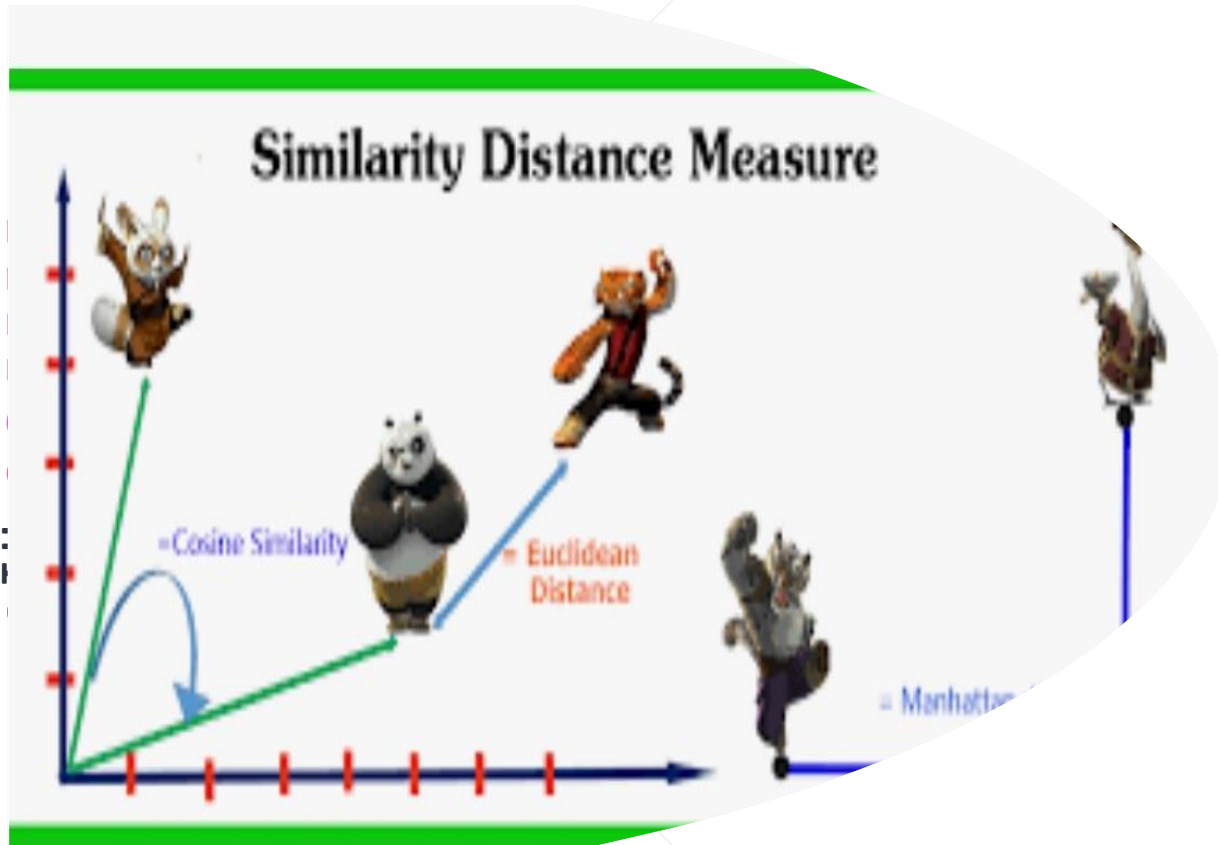
What are Distance Measures?

- Why do we need to measure distance?

- Is this the same as "similarity"?

- What does it mean for two documents to be "similar"?

Types of Distance Measures

1) Are there others?

2) What is "edit distance"?

3) How about Jaccard?

Euclidean	$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$
Squared Euclidean	$d(x, y) = \sum (x_i - y_i)^2$
Manhattan	$d(x, y) = \sum |(x_i - y_i)|$
Canberra	$d(x, y) = \sum \dfrac{|x_i - y_i|}{|x_i + y_i|}$
Chebychev	$d(x, y) = \max(|x_i - y_i|)$
Bray Curtis	$d(x, y) = \dfrac{\sum |x_i - y_i|}{\sum x_i + y_i}$
Cosine Correlation	$d(x, y) = \dfrac{\sum (x_i y_i)}{\sqrt{\sum (x_i)^2 \sum (y_i)^2}}$
Pearson Correlation	$d(x, y) = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (y_i - \overline{y})^2} \sqrt{\sum (y_i - \overline{y})^2}}$
Uncentered Peason Correlation	$d(x, y) = \dfrac{\sum x_i y_i}{\sqrt{\sum (y_i - \overline{y})^2} \sqrt{\sum (y_i - \overline{y})^2}}$

4TH COFFEE