4TH COFFEE

Text Mining

When the Words Become the Variables

What is Text Mining?

How is Text Mining Different from Data Analytics?

4TH COFFEE

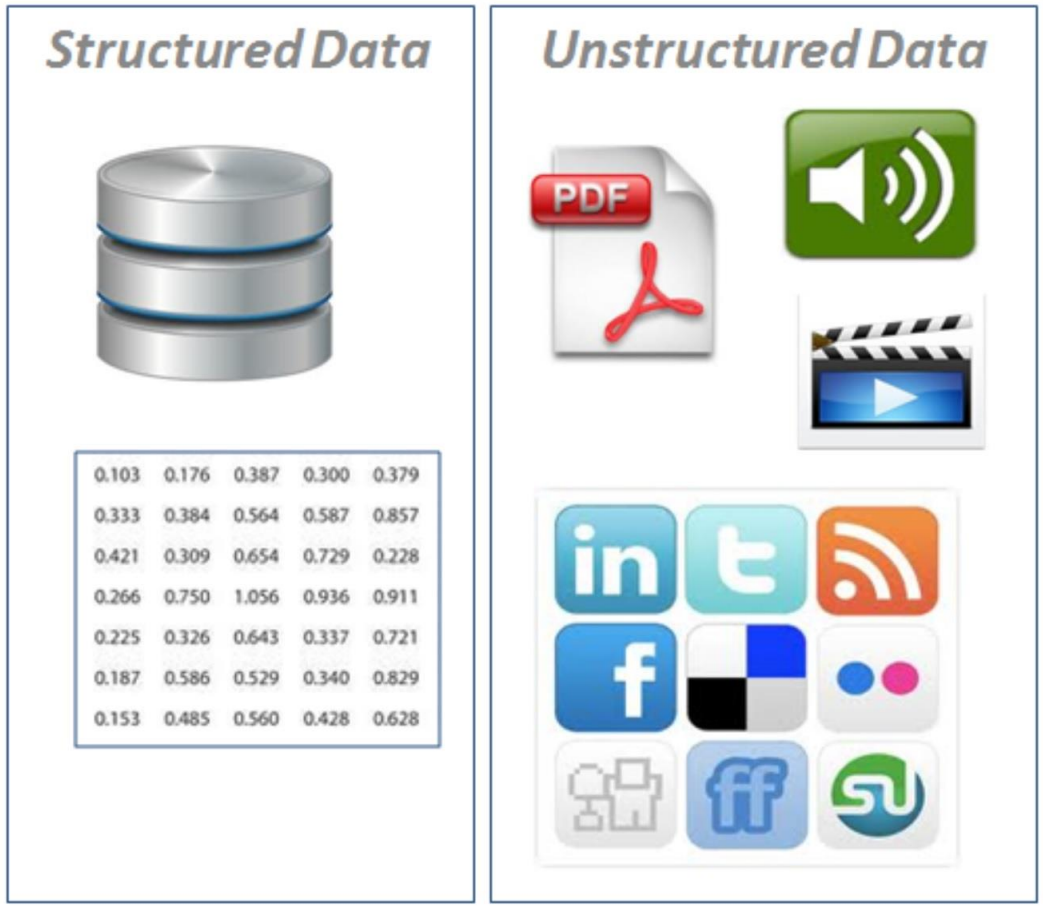
What is Data?
What is Text Data?

Types of Data

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

Is text data “different” from number data?

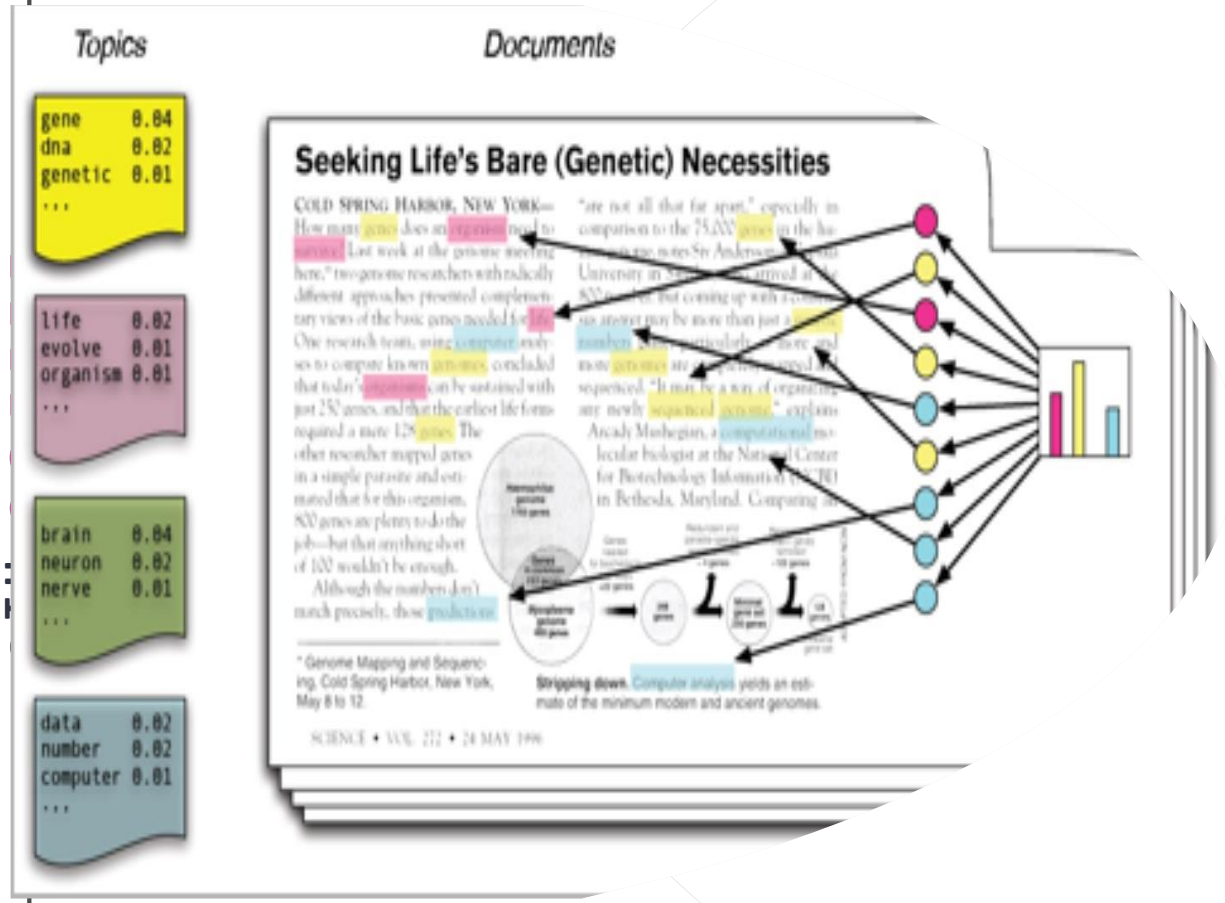
Can text data be represented as numerical data?



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



What is Topic Modeling?

What is a document?

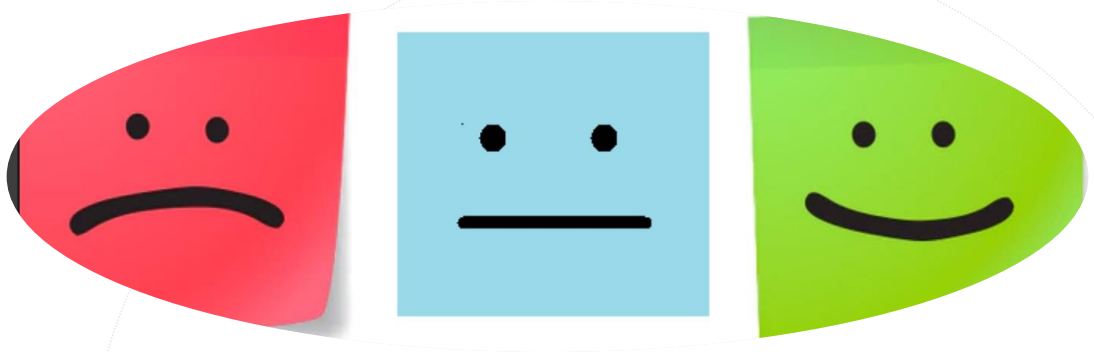
What is a Corpus?

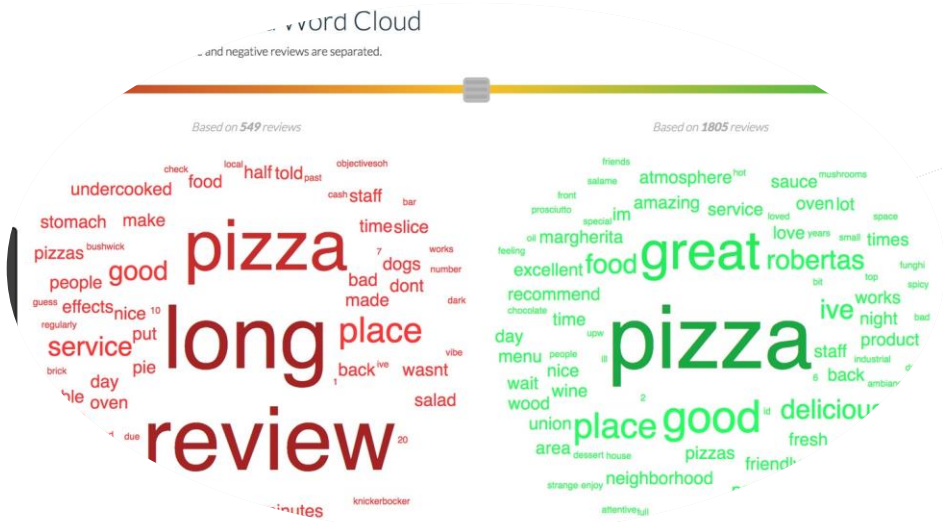
Do documents have “topics”?

What is a topic distribution?

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

4TH COFFEE





SENTIMENT ANALYSIS

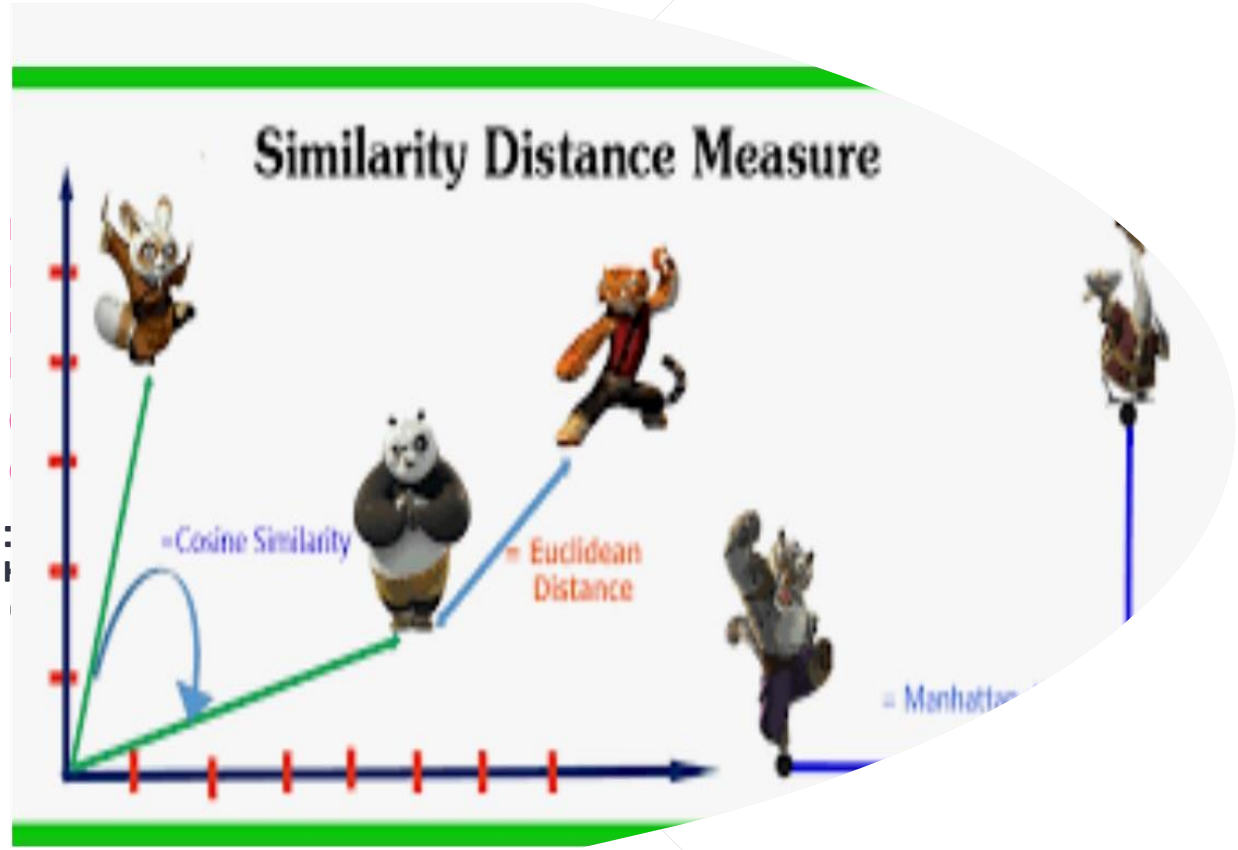
Can Data Analytics be used to understand sentiment?

Can models be created (trained) to predict sentiment?

What are some “supervised” methods that can be used?

Can unsupervised methods be used?

- How?



What are Distance Measures?

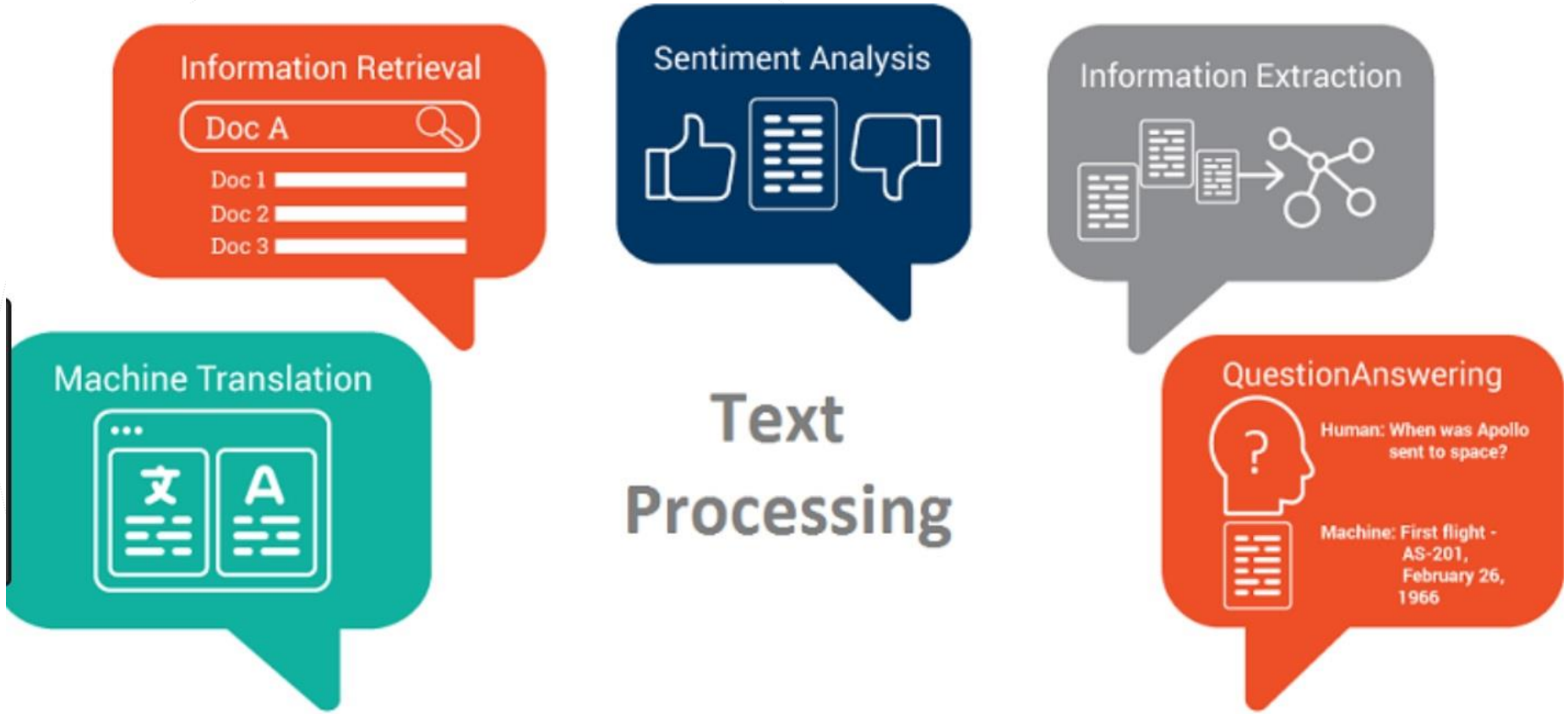
- Why do we need to measure distance?
- Is this the same as “similarity”?
- What does it mean for two documents to be “similar”?

4TH COFFEE

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	$d(x,y) = \sum \lvert (x_i - y_i) \rvert$
Canberra	$d(x,y) = \sum \frac{\lvert x_i - y_i \rvert}{\lvert x_i + y_i \rvert}$
Chebychev	$d(x,y) = \max(\lvert x_i - y_i \rvert)$
Bray Curtis	$d(x,y) = \frac{\sum \lvert x_i - y_i \rvert}{\sum x_i + y_i}$
Cosine Correlation	$d(x,y) = \frac{\sum (x_i y_i)}{\sqrt{\sum (x_i)^2 \sum (y_i)^2}}$
Pearson Correlation	$d(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2} \sqrt{\sum (y_i - \bar{y})^2}}$
Uncentered Peason Correlation	$d(x,y) = \frac{\sum x_i y_i}{\sqrt{\sum (y_i - \bar{y})^2} \sqrt{\sum (y_i - \bar{y})^2}}$


Information Retrieval
Doc A
Doc 1
Doc 2
Doc 3
Sentiment Analysis
Information Extraction
Machine Translation
Text Processing
QuestionAnswering
Human: When was Apollo sent to space?
Machine: First flight - AS-201, February 26, 1966