

Data Science in Business: Text Analytics

Dr. Peter Molnar

Sarah Zeis

Carly Wieting

Course Overview

Class 1: Introduction to Machine Learning and Set-Up Python

Class 2: Data Exploration

Class 3: Machine Learning Models (Decision Tree and KNN)

Class 4: Text Analytics

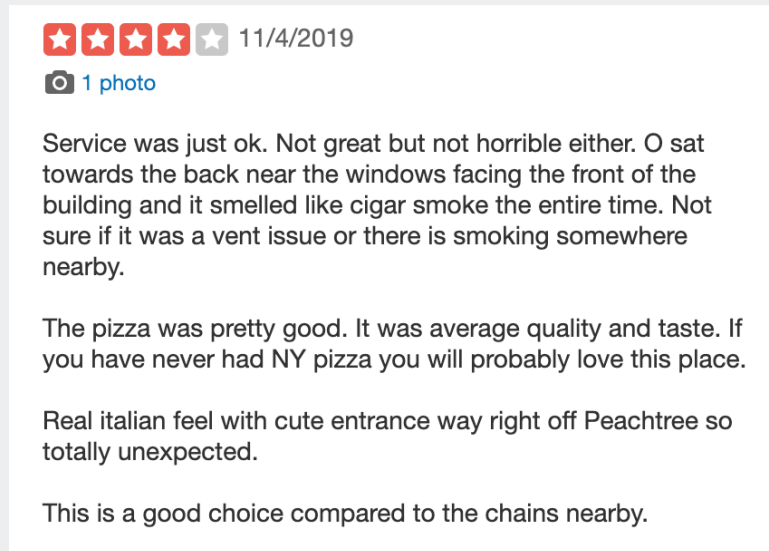
Class 5: Forecasting with Facebook Prophet

Goals and Takeaways

- Understand Natural Language Processing & data preprocessing requirements
- Topic Modeling Process & Applications
- Sentiment Analysis & Applications
- Python Notebook Lab

Natural Language Processing

Natural Language Processing (**NLP**) is the extraction of **meaning** and **information** from text documents.



A screenshot of a Yelp review. At the top, there are five red stars, four of which are filled, followed by the date '11/4/2019' and a camera icon with the text '1 photo'. The review text is as follows: 'Service was just ok. Not great but not horrible either. O sat towards the back near the windows facing the front of the building and it smelled like cigar smoke the entire time. Not sure if it was a vent issue or there is smoking somewhere nearby.' followed by 'The pizza was pretty good. It was average quality and taste. If you have never had NY pizza you will probably love this place.' followed by 'Real italian feel with cute entrance way right off Peachtree so totally unexpected.' and finally 'This is a good choice compared to the chains nearby.'

Information: words as a vector, # of times words were used

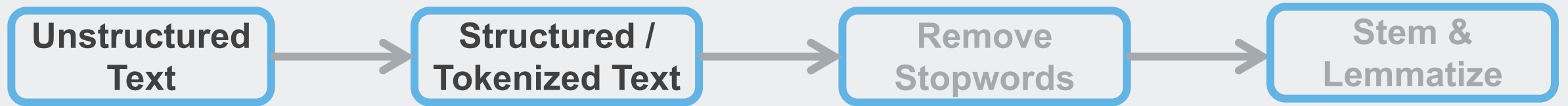
Meaning: Positive, joy, anticipation

NLP Pre-processing

NLP Requires Structured Data

Text is aggregated and stored as unstructured data.

NLP requires data pre-processing to turn blocks of text into a structure.



NLP pre-processing is typically initiated with by adding structure with **tokenization**.

The Tokenization Process

Tokenization breaks a sentence (or block of text) into its individual components.

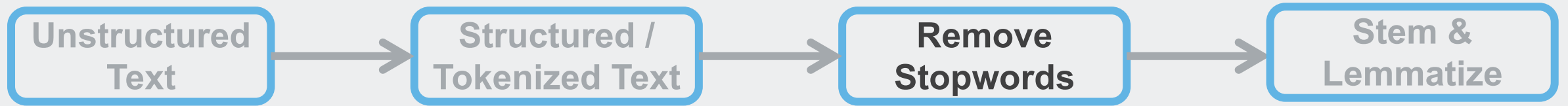


Use sentence breaks to turn an **unstructured** block of words into a **list**

['Tokenization', 'breaks', 'a', 'sentence', 'or', 'block', 'of', 'text', 'into', 'its', 'individual', 'components']

Now that we have a list, we have added structure to our data.

Removing Stopwords



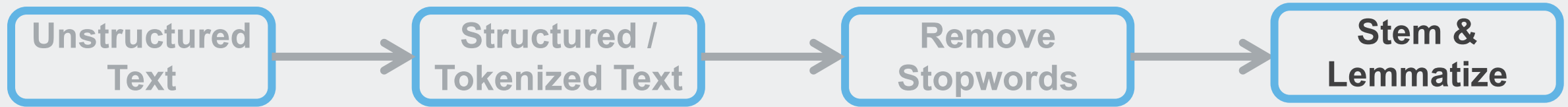
Removing stopwords removes general language terms that do not add value to the information in your text.

Stopwords in the above sentence: ['do', 'that', 'to', 'the', 'in', 'your']

Sentence after removing stopwords: Removing stopwords removes general language terms not add value information text

Stopwords can be identified by creating a list or using an existing package that can provide a list of stopwords.

Stemming & Lemmatization



Stemming

Removing common endings from words by defining rules

Ex: ly, ing, er, es, s, ed

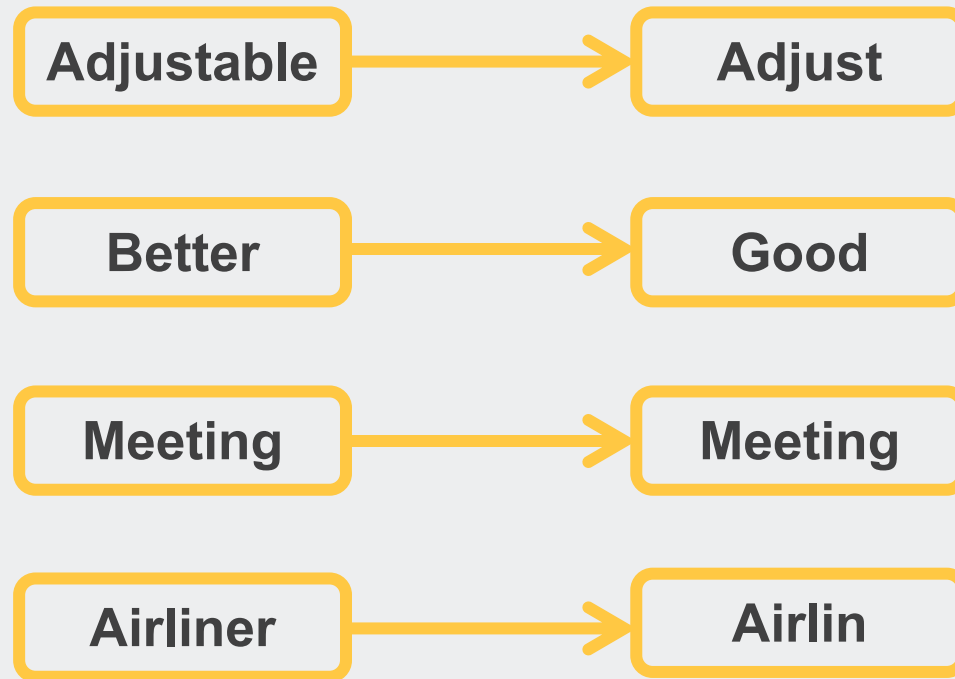
Lemmatization

Using language & grammar rules to identify the roots of words

Ex: best -> good, lying -> lie

Stemming & Lemmatization

Which transformation was used in the below word examples:



Bag of Words

Bag of Words: Converting pre-processed text into a dictionary that counts how often a word occurs.

Example

Sarah likes to run. Carly also likes to run. They run in the city.



Tokenize & remove stopwords

[sarah, like, run, carly, like, run, they, run city]



Transform into bag of words

{"sarah": 1, "carly": 1, "run": 3, "like": 2, "city": 1}



Machine Learning Models

Topic Modeling

Topic Modeling is a method of statistical modeling for identifying the general topics within a document. Topic modeling is unsupervised because the topics in the documents are unknown.

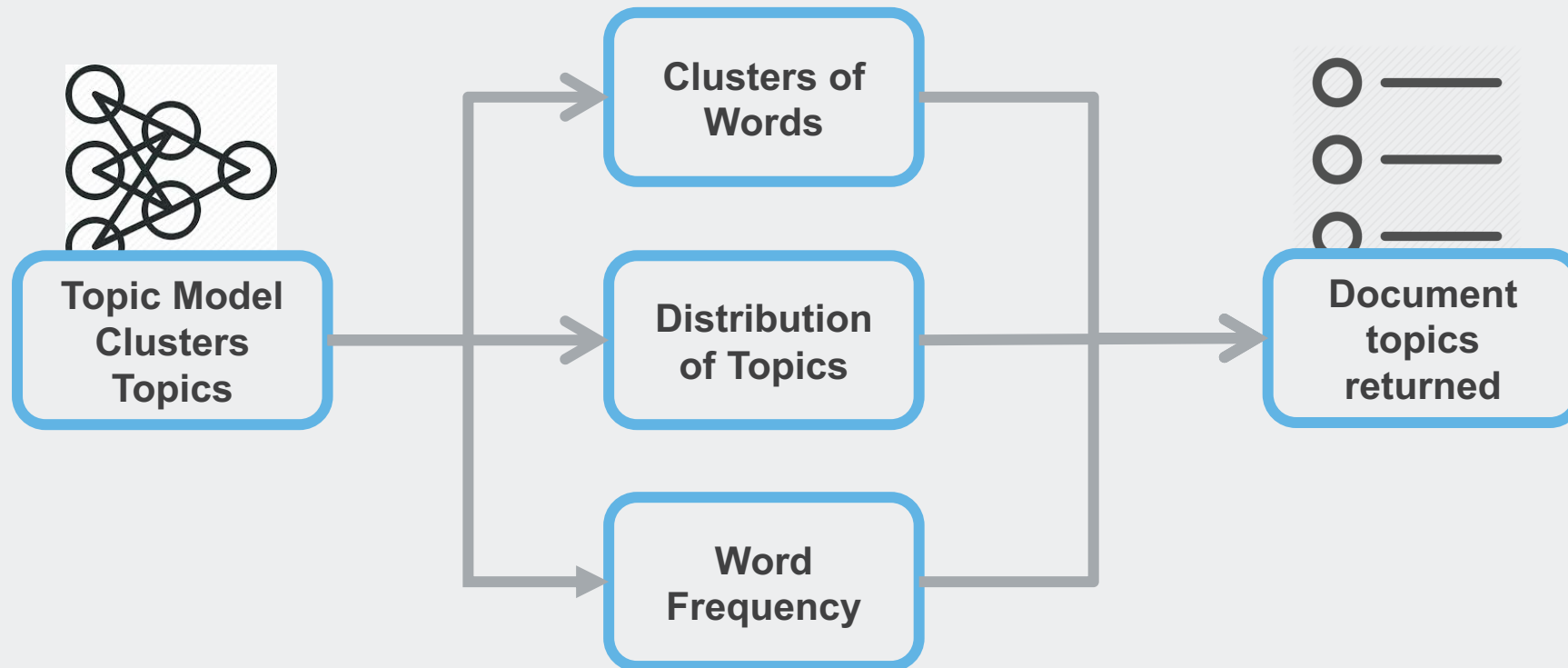
Example: I am looking for a common theme across multiple text documents for a legal case.



Topic Modeling Process

Modeling Overview & Outputs

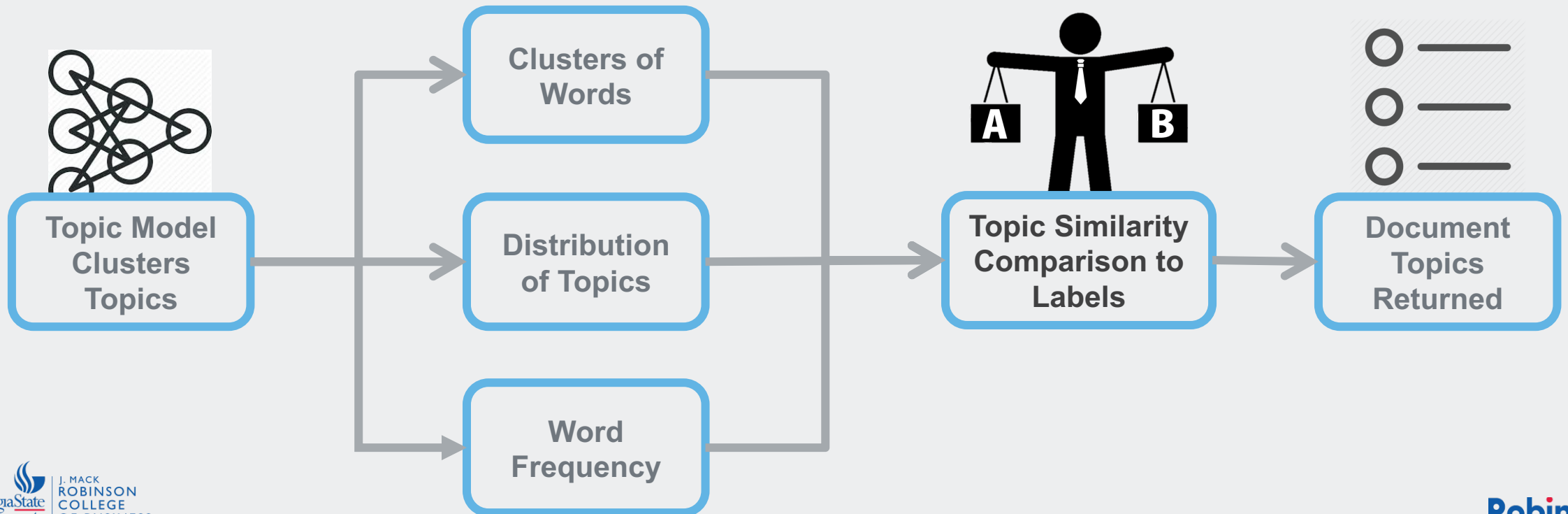
Model identifies topics present in a document (unstructured text) by deriving patterns in the document.



Supervised Topic Modeling

Supervised Topic Modeling is referred to as Topic Classification

Topic modeling can be trained with labels as well. This allows the user to have the algorithm pre-identify clusters or distributions of topics present in each document.



Sentiment Analysis

Determining whether a response or statement is **positive**, **negative** or **neutral**.

Sentiment can be difficult for humans to detect.



Robbie Collin ✓ @robbiereviews · 18m

Glad to report that **Cats** is everything you'd hoped for and more: a mesmerisingly ugly fiasco that makes you feel like your brain is being eaten by a parasite. A viewing experience so stressful that it honestly brought on a migraine.



Ge Wang
@gewang

Having now seen the movie, I get the inherent difficulty in reviewing it. The reason to see, or to not see "Cats" is one and the same: it is experiential (thus defying description) and it's bafflingly awful and kind of awesome at the same time. (cont)



jen yamato ✓
@jenyamato

True story: I got home after seeing CATS and couldn't look my own cat in the face for a good hour or two

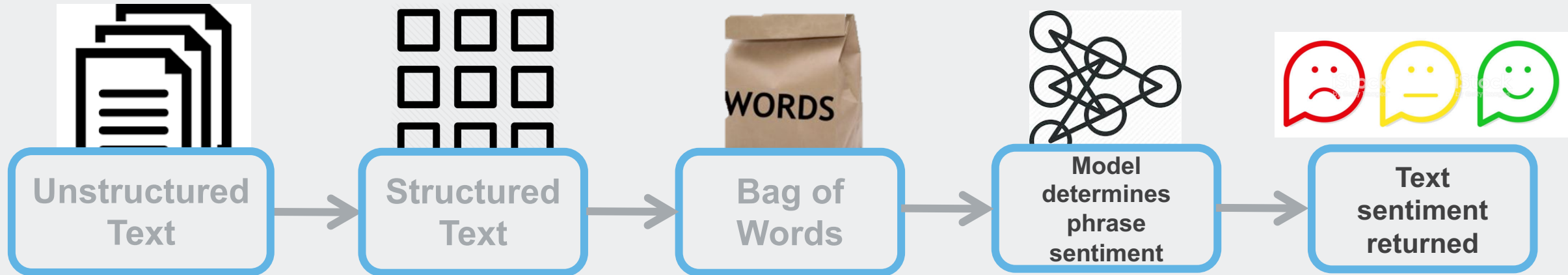
4:48 PM · Dec 18, 2019 · [Twitter for iPhone](#)

2.2K Retweets 17.5K Likes

Sentiment Analysis Process

Sentiment is typically classified by using one of two approaches:

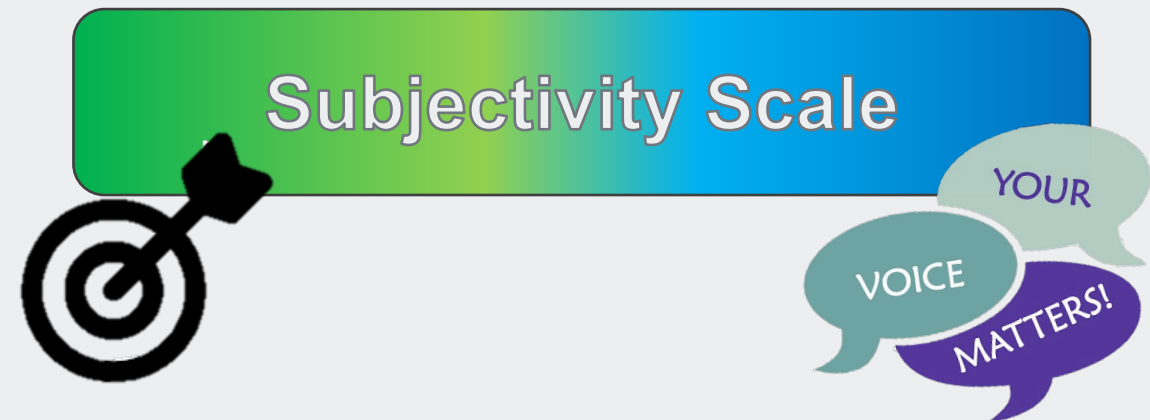
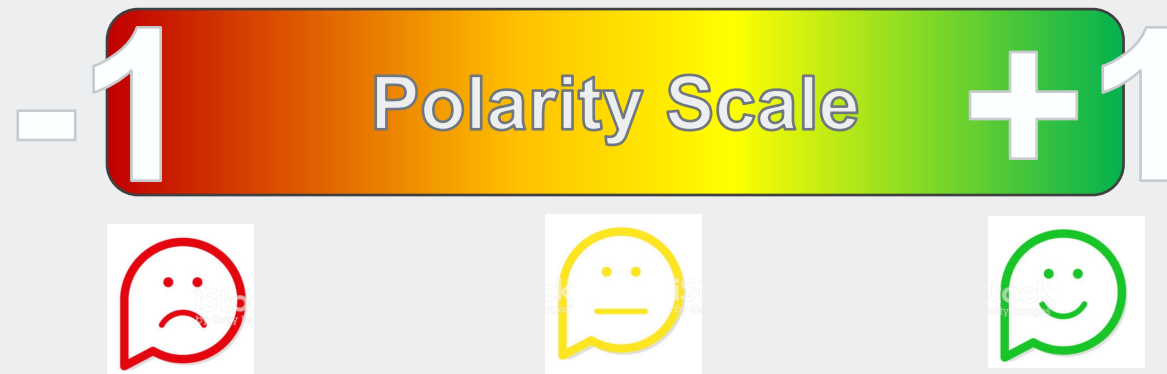
1. A supervised approach with a labeled dataset (ie. Yelp review text w/ a star rating)
2. An unsupervised approach with a pretrained statistical package (ie. Textblob)



Generating Sentiment Analysis

Sentiment analysis can be supervised or unsupervised. The **textblob** sentiment analysis package will return two values (Polarity, Subjectivity):

1. **Polarity**: score ranges from -1:1. -1 is a negative sentiment, 0 is neutral, 1 is positive
2. **Subjectivity**: score ranges from 0:1. 0 is an objective score; 1 is a subjective score



Natural Language Processing Applications

Voice of the Customer (Customer reviews, customer sentiment)

Legal Analytics (information from lengthy legal documents)

E-mail classification (eg: spam, promotions, social)

Topic Modeling in Pythom

Navigate to the bootcamp Github:

<https://github.com/institute4insight/Data-Science-in-Business-Bootcamp>

And download today's notebook!