# Text Classification Using Naive Bayes

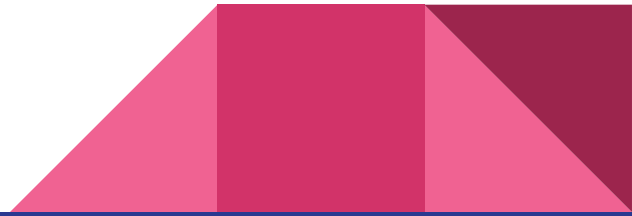
Data Science, Pandas, sci-kit learn

Prepared by Chad Valencia

# Objective

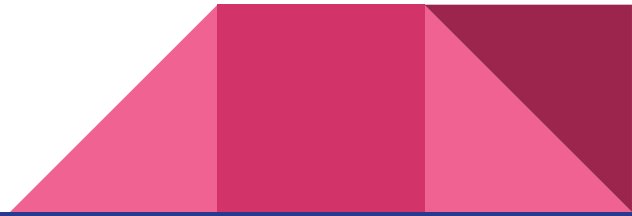
To **classify** a review as *positive* or *negative* using sample data and the **Naive Bayes classifier**.

This lesson is an introduction into the field of Machine Learning called **Natural Language Processing (NLP)**.



# Review

- Classification: Placing each row of data into a target with discrete outcomes
  - Examples: True/False, Yes/No, Hot Dog/Not Hot Dog



# Review

Supervised Learning:


| Weight (lbs) | Number of legs | Meows | Is Cat |
|---|---|---|---|
| 20 | 4 | 0 | 0 |
| 17 | 4 | 1 | 1 |
| 150 | 2 | 1 | 0 |

Dataset

| Weight | Legs | Meows | Is Cat |
|---|---|---|---|
| 19 | 4 | 1 | ? |

Prediction

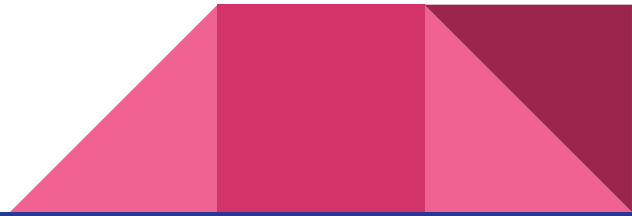
# Review

Things we've learned in the past:

Python

Pandas

Some familiarity with sci-kit learn



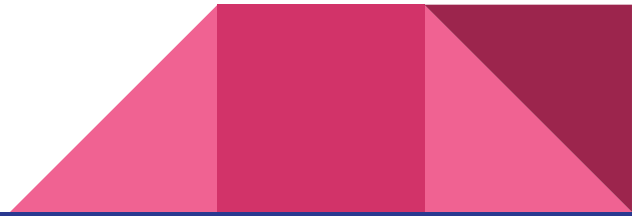
# Text Data

Common Text Data:

- Comments
- Reviews
- Forum Posts
- Reddit
- Tweets

# Naive Bayes

- An algorithm based off of Bayes Theorem, that states that the probability of an event happening is of the form:
  - Posterior = (prior * likelihood) / evidence
- Great for text classification
- Naive Assumption: Every feature is independent.
- Sci-kit learn: MultinomialNB



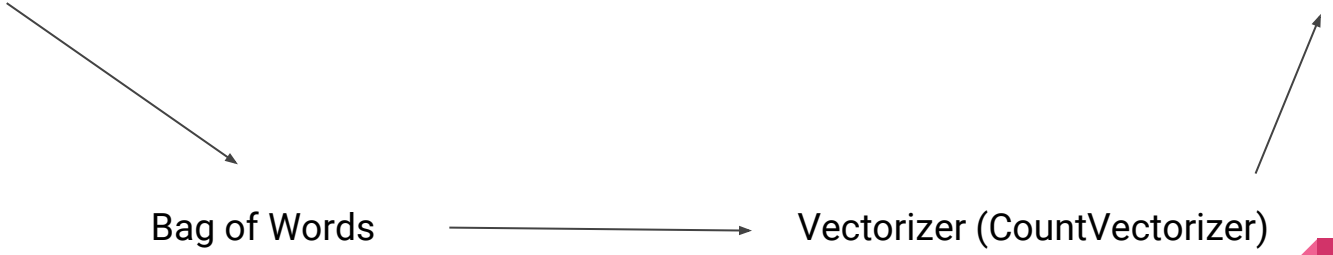
# Bag of Words

1 The fox jumps.

2 Fox jumps over.

3 The fox jumps the fox.

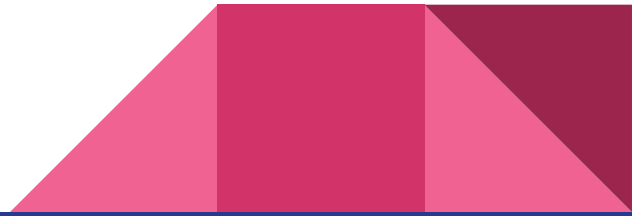
| Row | the | fox | jumps | over |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 |
| 2 | 0 | 1 | 1 | 1 |
| 3 | 2 | 2 | 1 | 0 |



# Terminology

| Data Science/Python | Natural Language Processing |
| --- | --- |
| Dataset | Corpus |
| Row of dataset | Document |
| Feature Set | Vocabulary |

# Demo



# Conclusions and Questions

- Made very simple assumptions to create a very rough model
- Model tweaks (parameters, assumptions)
- Data collection and labelling