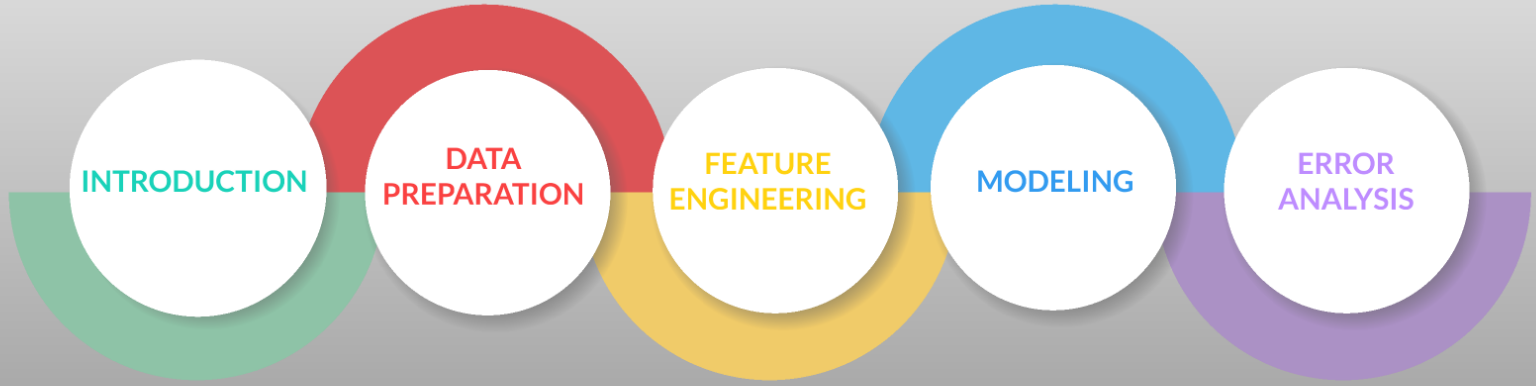
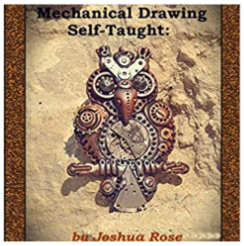
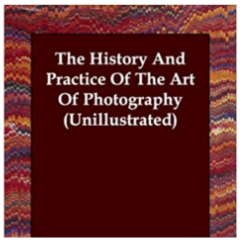
Text Classification:

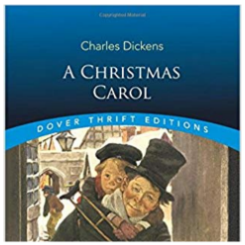


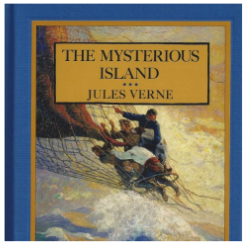
Presented by Sahand Malek
Soroush Salehi
Hossein Zinaghaji

In this project, the goal is to identify the author of several books by analyzing some words from the original books. Three different Feature selection methods for texts and five different classifying methods are being used.
The ultimate target is to classify, predict, and compare the methods. Seven different books are selected from Gutenberg Digital Books Library

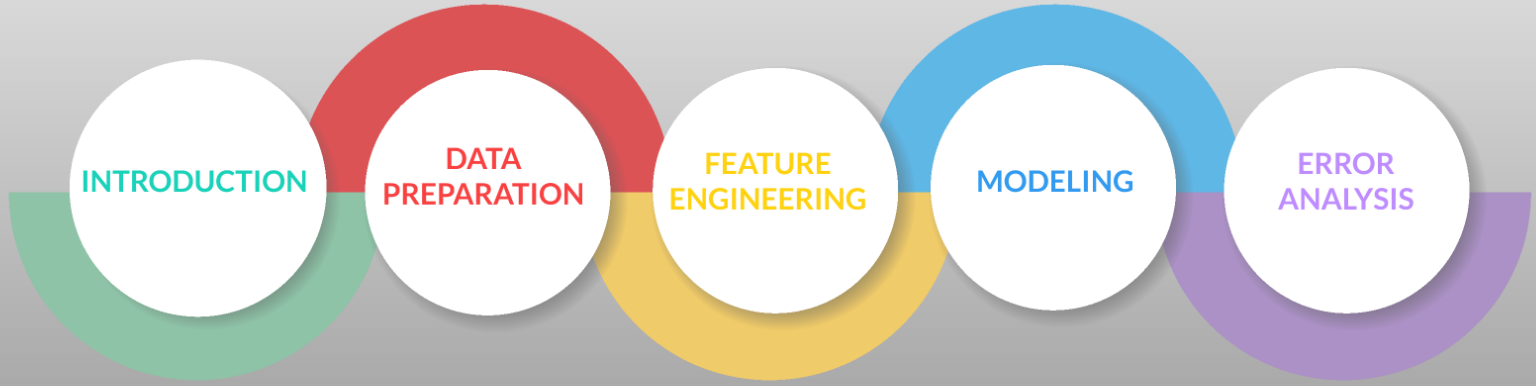




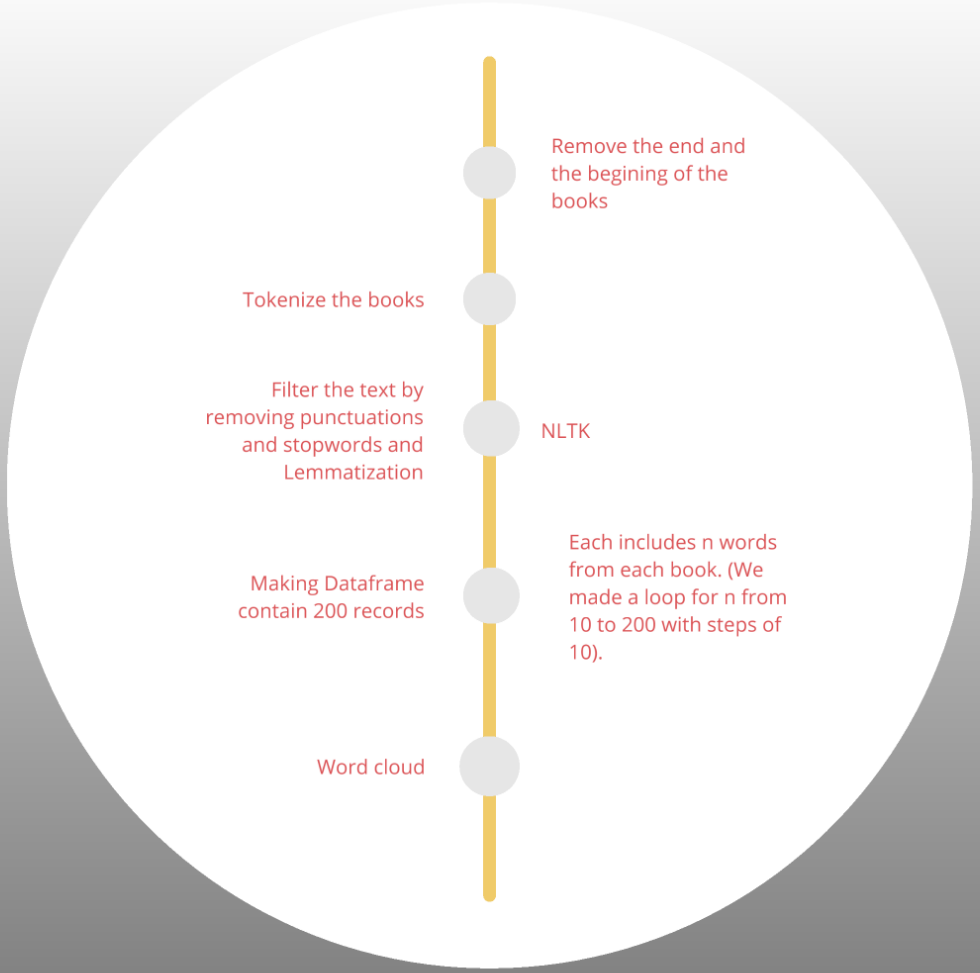


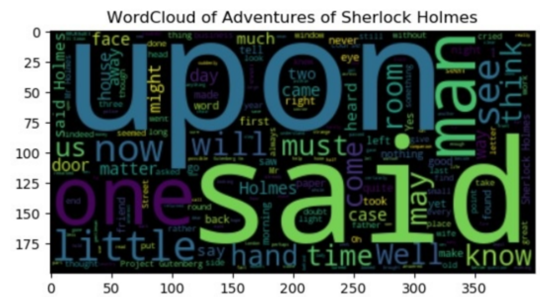


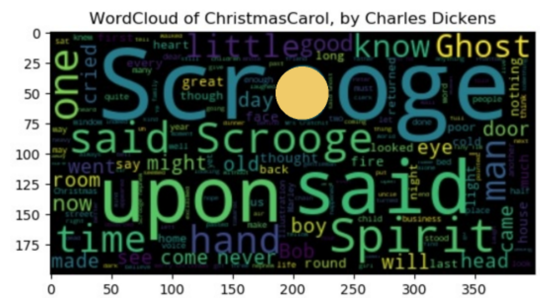
Text Classification:

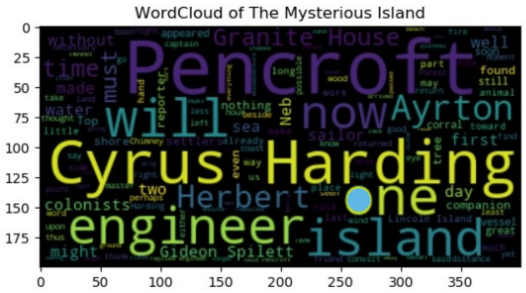


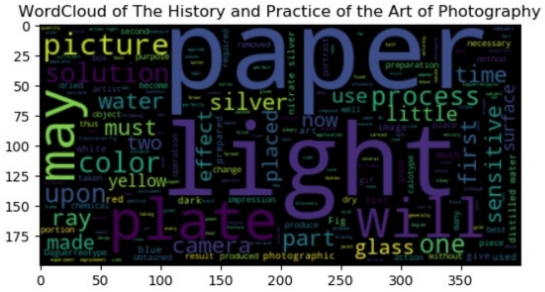
Presented by Sahand Malek
Soroush Salehi
Hossein Zinaghaji


Remove the end and the begining of the books
Tokenize the books
Filter the text by removing punctuations and stopwords and Lemmatization
NLTK
Each includes n words from each book. (We made a loop for n from 10 to 200 with steps of 10).
Making Dataframe contain 200 records
Word cloud


WordCloud of Adventures of Sherlock Holmes
0
25
50
75
100
125
150
175
0
50
100
150
200
250
300
350
upon
said
one
man
little
now
will
must
see
think
room
come
know
time
hand
Well
us
door
face
much
may
Holmes
matter
case

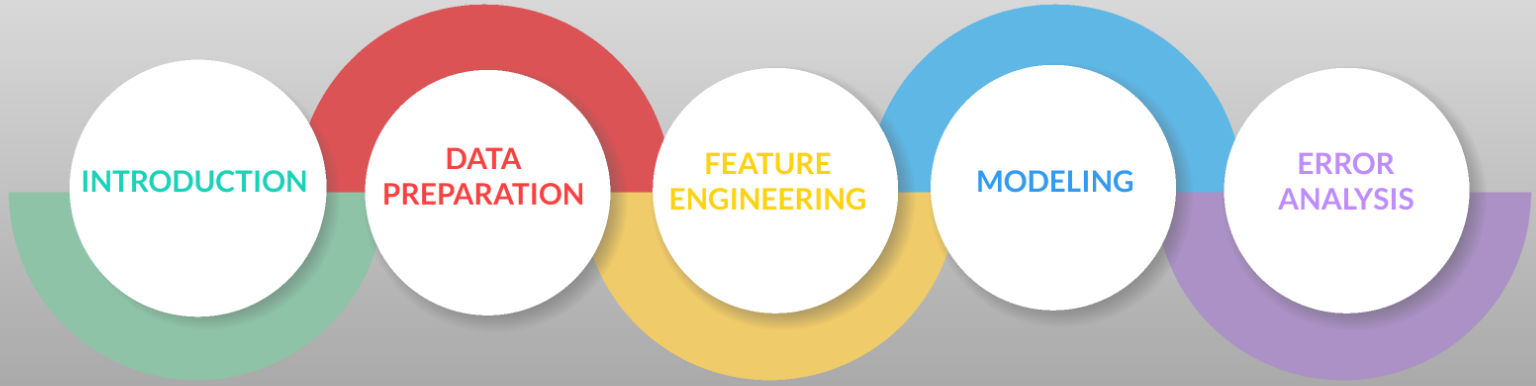

WordCloud of ChristmasCarol, by Charles Dickens
0
25
50
75
100
125
150
175
0
50
100
150
200
250
300
350
Scrooge
said Scrooge
upon
said
Spirit
time
hand
Ghost
know
little
good
one
man
door
eye
room
now
boy
will
head
made
come
never
round
went
might
old


WordCloud of The Mysterious Island
Pencroft
Granite House
Will
now
Ayrton
Cyrus Harding
Herbert
one
engineer
island
Gideon Spilett

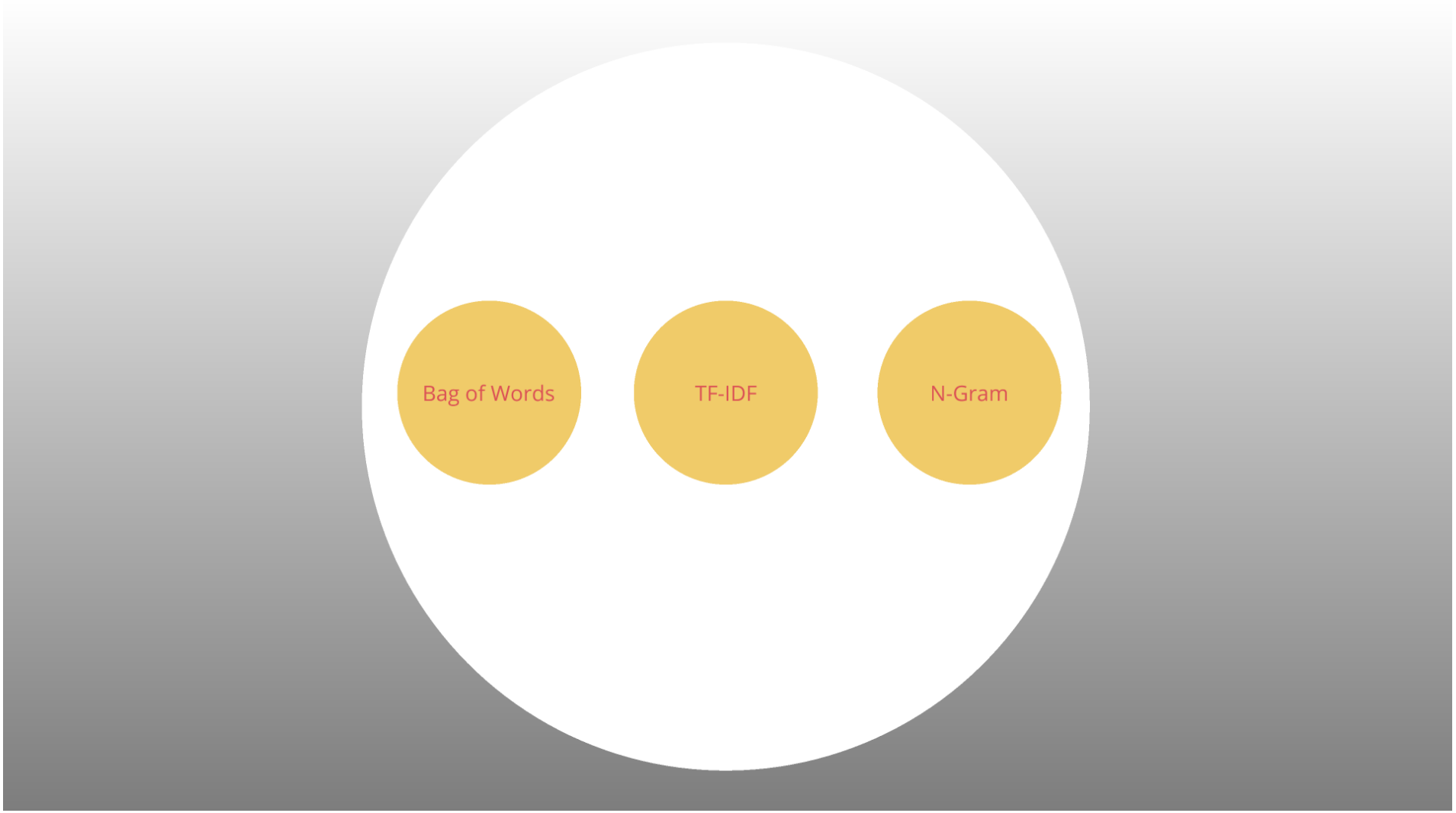

WordCloud of The History and Practice of the Art of Photography
paper
light
plate
picture
solution
water
silver
process
may
color
will
camera
glass
one


WordCloud of Mechanical Drawing Self-Taught
WordCloud of Dorothy and the Wizard in Oz
WordCloud of The Three Musketeers

Text Classification:



Presented by Sahand Malek
Soroush Salehi
Hossein Zinaghaji

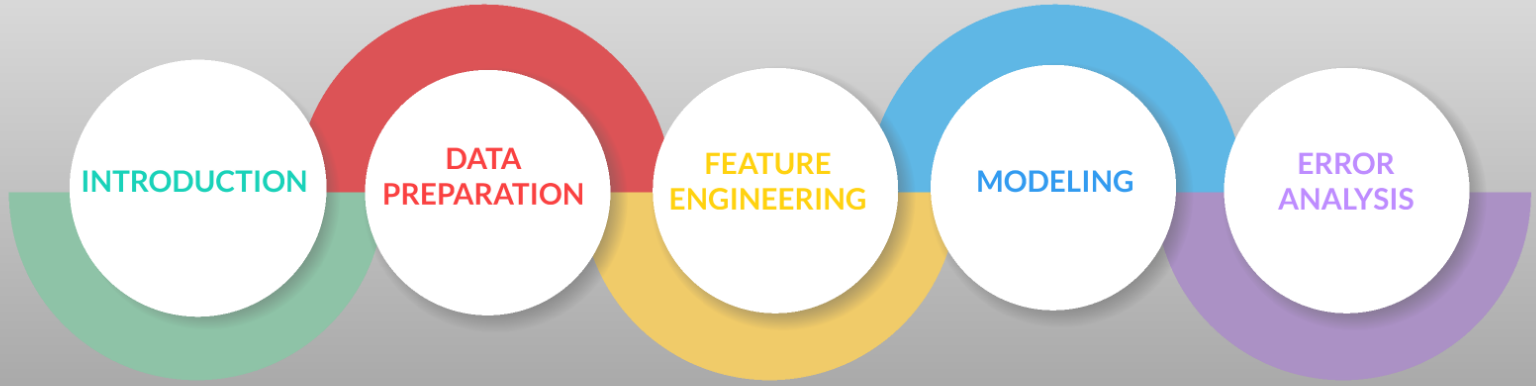

Bag of Words
TF-IDF
N-Gram

The Bag-of-words model is a simple representation used in natural language processing and information retrieval (IR). Also known as the vector space model. In this model, a text (such as a sentence or document) is displayed as a package of several sets of words, disregarding grammar and even word order. In practice, this model is mainly used as a tool for generating features. After converting the text into a bag of words, we can calculate different sizes to specify the text. The most common type of attribute derived from the model is the repetition of the phrase, that is, the number of times a phrase appears in the text.

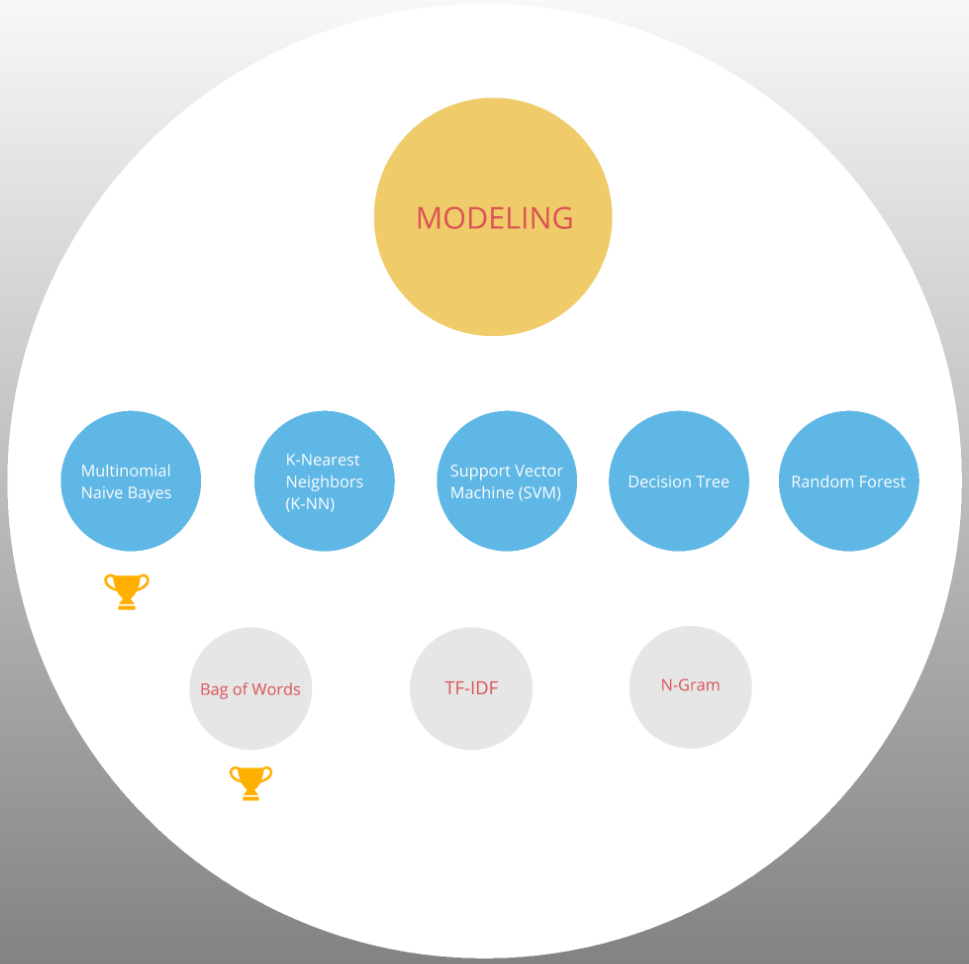
The value of TF-IDF stands for two words: TF means Term Frequency means the number of repetitions of a word in a text and IDF means Inverse Document Frequency which can be translated to the number of repetitions in the text. TF-IDF is a way to measure the importance of a word in a document. TF measures that a term is repeated several times in a document and the IDF is used to measure the significance of a term.

the n-gram is a sequence of n letters in a given sequence of text or words. Depending on their use, items can be phonemes, syllables, letters, words, or pairs. N-grams are usually collected from a textual or oral body.

Text Classification:



Presented by Sahand Malek
Soroush Salehi
Hossein Zinaghaji


MODELING
Multinomial Naive Bayes
K-Nearest Neighbors (K-NN)
Support Vector Machine (SVM)
Decision Tree
Random Forest
Bag of Words
TF-IDF
N-Gram

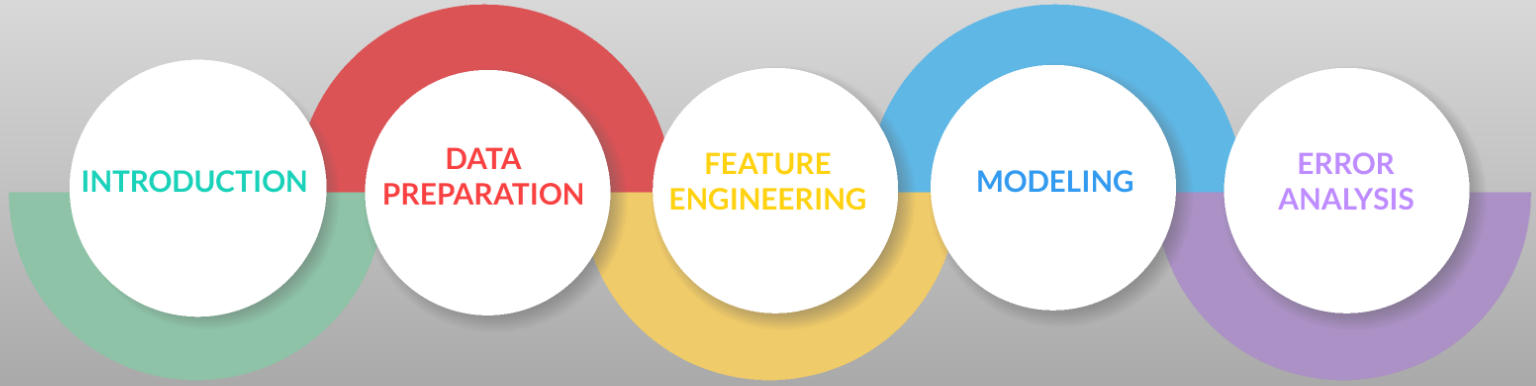
	Multinomial Naive Bayes	K-Nearest Neighbors (K-NN)	Support Vector Machine (SVM)	Decision Tree	Random Forest
n = 10	50.46%	18.07%	44.82%	38.36%	45.61%
n = 20	73.61%	20.00%	60.75%	50.32%	57.82%
n = 30	84.36%	23.00%	75.71%	58.86%	67.68%
n = 40	88.93%	27.25%	83.18%	68.00%	78.36%
n = 50	93.68%	29.29%	88.39%	69.14%	82.82%
n = 60	97.54%	34.21%	92.54%	80.00%	88.43%
n = 70	98.43%	41.61%	94.04%	78.00%	91.00%
n = 80	99.07%	45.29%	96.54%	84.18%	93.57%
n = 90	99.71%	45.50%	97.64%	82.89%	95.39%
n = 100	99.82%	50.64%	97.71%	83.36%	96.61%
n = 110	99.68%	65.36%	99.14%	88.29%	98.11%
n = 120	99.71%	60.36%	99.11%	86.29%	97.82%
n = 130	100.00%	64.04%	98.82%	88.32%	98.71%
n = 140	99.89%	71.82%	99.64%	86.71%	99.14%
n = 150	100.00%	70.86%	99.61%	90.25%	99.11%
n = 160	100.00%	66.46%	99.36%	90.11%	99.21%
n = 170	100.00%	71.64%	99.64%	92.25%	99.54%
n = 180	100.00%	72.75%	99.68%	90.57%	99.14%
n = 190	100.00%	76.86%	99.54%	91.86%	99.71%
Average	93.94%	50.26%	90.83%	78.83%	88.83%

Table 4-2 TF-IDF

	Multinomial Naive Bayes	K-Nearest Neighbors (K-NN)	Support Vector Machine (SVM)	Decision Tree	Random Forest
n = 10	49.54%	15.93%	48.11%	35.54%	42.93%
n = 20	70.64%	57.36%	66.43%	51.89%	58.61%
n = 30	80.86%	69.11%	78.46%	60.14%	66.86%
n = 40	89.71%	78.68%	86.71%	68.11%	76.36%
n = 50	92.86%	83.54%	92.43%	68.11%	82.86%
n = 60	95.11%	86.82%	94.50%	73.50%	86.11%
n = 70	98.32%	89.39%	97.00%	77.57%	90.64%
n = 80	98.04%	91.82%	98.00%	80.46%	92.96%
n = 90	98.50%	95.07%	98.75%	80.50%	93.43%
n = 100	98.61%	94.89%	99.18%	85.43%	95.93%
n = 110	99.57%	96.00%	99.50%	83.18%	97.14%
n = 120	99.43%	96.96%	99.21%	85.18%	97.82%
n = 130	99.79%	98.43%	99.39%	86.04%	97.82%
n = 140	99.89%	98.18%	99.57%	88.57%	98.39%
n = 150	99.89%	98.61%	99.86%	89.64%	98.93%
n = 160	99.75%	99.00%	99.75%	89.32%	99.14%
n = 170	100.00%	98.75%	100.00%	89.75%	99.57%
n = 180	100.00%	99.57%	99.89%	90.04%	99.46%
n = 190	100.00%	99.75%	99.93%	91.21%	99.50%
Average	93.18%	86.73%	92.46%	77.59%	88.13%

	Multinomial Naive Bayes	K-Nearest Neighbors (K-NN)	Support Vector Machine (SVM)	Decision Tree	Random Forest
n = 10	42.89%	25.25%	35.04%	28.64%	41.14%
n = 20	58.36%	30.21%	49.93%	35.89%	51.93%
n = 30	69.71%	37.50%	58.79%	46.18%	59.32%
n = 40	79.04%	44.36%	69.61%	50.64%	69.29%
n = 50	83.46%	47.71%	73.71%	52.04%	72.79%
n = 60	87.36%	53.11%	78.00%	56.21%	76.89%
n = 70	91.14%	59.46%	83.36%	63.39%	83.04%
n = 80	93.29%	62.18%	84.79%	63.18%	84.46%
n = 90	94.93%	64.32%	86.89%	64.46%	87.39%
n = 100	95.46%	70.07%	89.96%	72.39%	89.21%
n = 110	96.79%	69.32%	91.64%	70.96%	89.50%
n = 120	97.57%	77.64%	92.96%	71.18%	93.00%
n = 130	98.43%	79.21%	93.36%	75.32%	94.43%
n = 140	99.04%	82.36%	93.54%	73.61%	94.21%
n = 150	99.00%	81.71%	95.71%	76.57%	95.71%
n = 160	98.71%	85.04%	96.32%	78.32%	96.00%
n = 170	99.21%	84.50%	95.64%	81.71%	96.36%
n = 180	99.57%	85.79%	96.57%	77.50%	95.25%
n = 190	99.36%	88.75%	98.36%	80.29%	97.61%
Average	88.60%	64.66%	82.33%	64.13%	82.50%

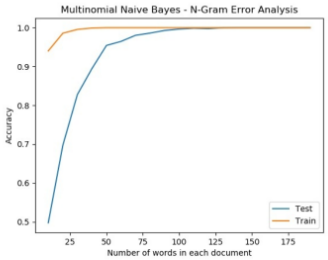
Text Classification:

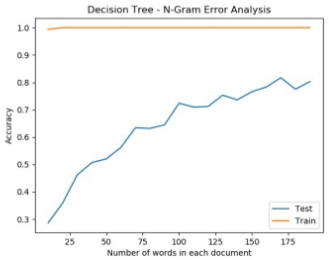


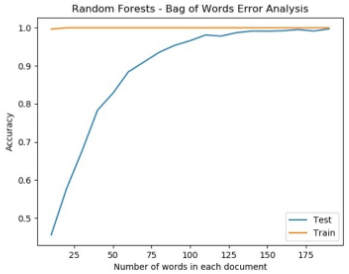
Presented by Sahand Malek
Soroush Salehi
Hossein Zinaghaji

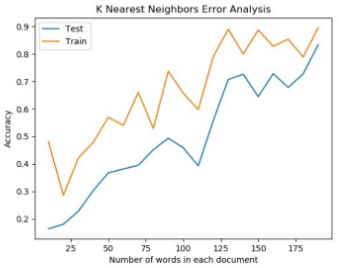

ERROR ANALYSIS
Accuracy
Confusion Matrix
Recall
Precision
F1-Score
CONCLUSION

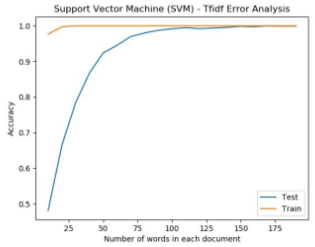
$$\begin{pmatrix} 33 & 5 & 7 & 0 & 0 & 4 & 1 \\ 2 & 37 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 31 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 39 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 47 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 32 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 37 \end{pmatrix}$$

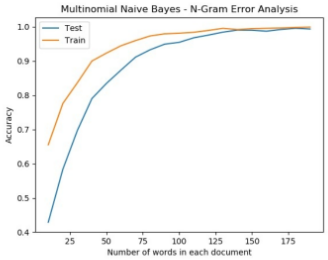

Multinomial Naive Bayes - N-Gram Error Analysis
Accuracy
Number of words in each document
Test
Train

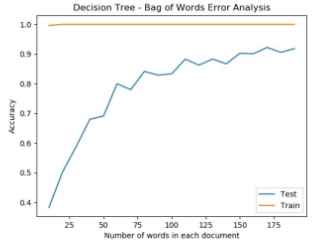

Decision Tree - N-Gram Error Analysis
Accuracy
Number of words in each document
Test
Train

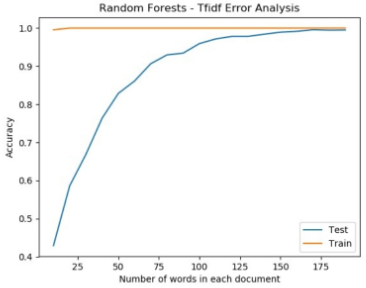

Random Forests - Bag of Words Error Analysis
Accuracy
Number of words in each document
Test
Train

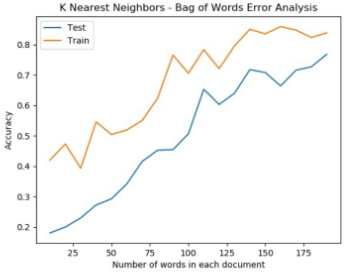

K Nearest Neighbors Error Analysis
Accuracy
Number of words in each document
Test
Train

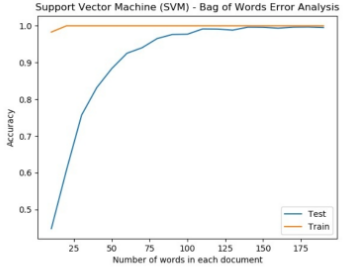

Support Vector Machine (SVM) - Tfidf Error Analysis
Accuracy
Number of words in each document
Test
Train

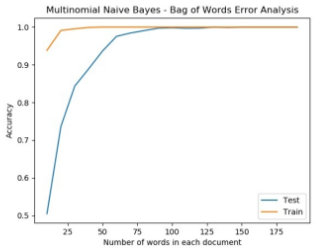

Multinomial Naive Bayes - N-Gram Error Analysis
Accuracy
Number of words in each document
Test
Train

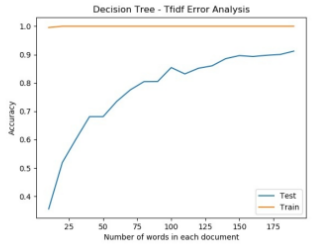

Decision Tree - Bag of Words Error Analysis
Accuracy
Number of words in each document
Test
Train

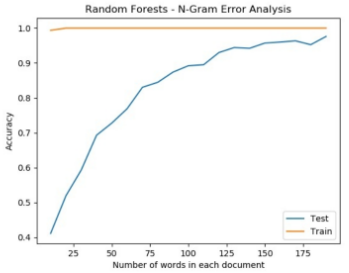

Random Forests - Tfidf Error Analysis
Accuracy
Number of words in each document
Test
Train

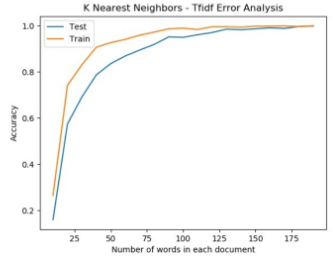

K Nearest Neighbors - Bag of Words Error Analysis
Accuracy
Number of words in each document
Test
Train


Support Vector Machine (SVM) - Bag of Words Error Analysis
Accuracy
Number of words in each document
Test
Train


Multinomial Naive Bayes - Bag of Words Error Analysis
Accuracy
Number of words in each document
Test
Train


Decision Tree - Tfidf Error Analysis
Accuracy
Number of words in each document
Test
Train


Random Forests - N-Gram Error Analysis
Accuracy
Number of words in each document
Test
Train


K Nearest Neighbors - Tfidf Error Analysis
Accuracy
Number of words in each document
Test
Train

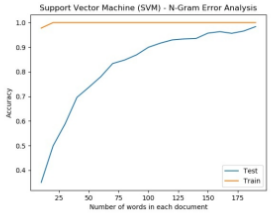

Support Vector Machine (SVM) - N-Gram Error Analysis
Accuracy
Number of words in each document
Test
Train

Table 5-1 An example of Classification Report (BoW Random Forest)

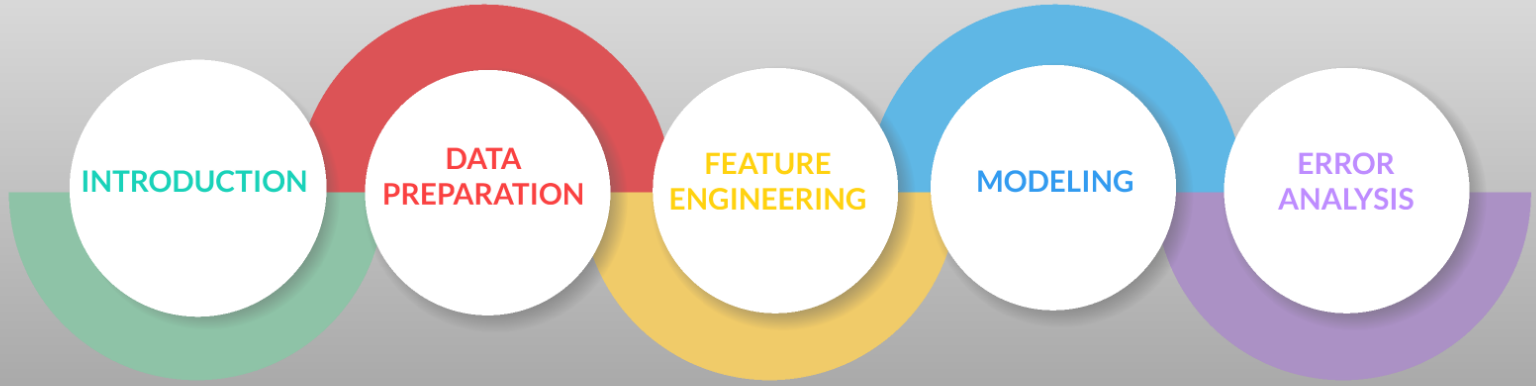
	Precision	Recall	F1-Score	Support
0	0.66	0.74	0.7	50
1	0.84	0.95	0.89	39
2	0.82	0.85	0.84	33
3	0.83	0.97	0.89	39
4	0.95	0.89	0.92	47
5	1	0.74	0.85	34
6	1	0.82	0.9	38
micro avg	0.85	0.85	0.85	280
macro avg	0.87	0.85	0.86	280
weighted avg	0.86	0.85	0.85	280

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Text Classification:



Presented by Sahand Malek
Soroush Salehi
Hossein Zinaghaji