



SPAM or ham?

Which should we chew on first?

BY: CLARKE WILLIAMS

ENTITY

Bio

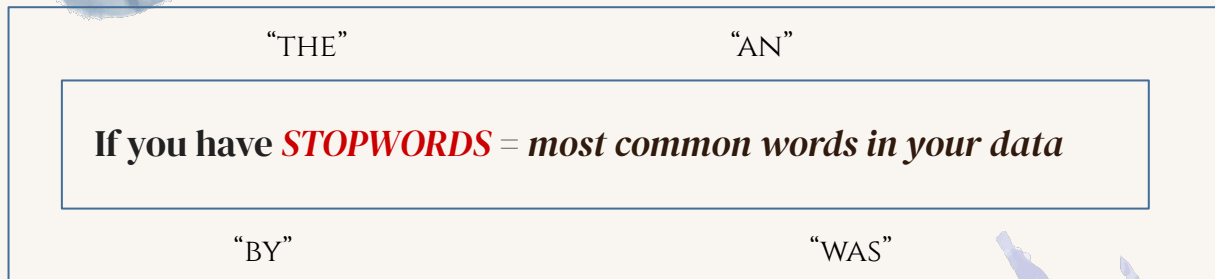
- ★ Undergrad education from UT Arlington as a Business major and City College of San Francisco
 - Certificate of Completions from PT Global for Personal Training, and from MSI for Business Office Manager
- ★ Employment has been geared around customer service, working in teams, and problem-solving for 10+ years in many different forms: from IHOP server, to licensed cosmetologist, to call center representative, to now personal trainer (and hopefully soon-to-be data scientist!)
- ★ Currently I'm employed as a Certified Personal Trainer at Eden Fitness Studio in San Francisco, CA



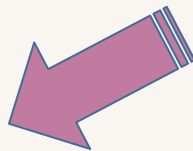
Why THIS project?

1. What is actually considered a “Ham” message?
2. How do we catch the “Spam” messages to put it in the appropriate folder?

(text, emails, and/or social media)



IT'S PROBABLY TRASH
(literally)



KNOWN & UNKNOWN SENDERS

The Method to the Madness



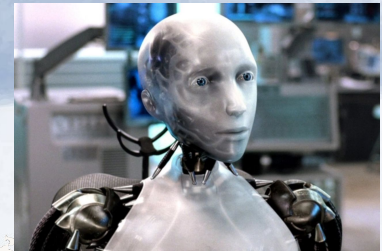
Languages: Python

- using Pandas, NLTK, & Sklearn packages

★ Using Natural Language Processing (NLP) based on machine learning to analyze text through many processes:

- **List comprehension** = a list created while filtering through an existing dataset
- **Stemming** = reducing a word to its root form by removing the suffix

MACHINE LEARNING (ML) → the use of computer systems to allow certain softwares to predict more accurate outcomes, without *actually* telling the computer to do so




- **TF-IDF** = a type of vectorizing/scoring that tells you the occurrence of a word and how important it is in a document



TF
(term frequency)



IDF
(inverse document frequency)



	good	movie	not	a	did	like
good movie	1	1	0	0	0	0
not a good movie	1	1	1	1	0	0
did not like	0	0	1	0	1	1

- **Naive Bayes** = an algorithm (based on Bayes Theorem) that classifies each predictor as independent, given the type of class
- **Random Forest** = an algorithm model randomly trying to accurately predict a categorical variable with as **LESS** bias as possible (used for *classification & regression* problems)

Results

Importing data:

	label	message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...



After the wrangling process:

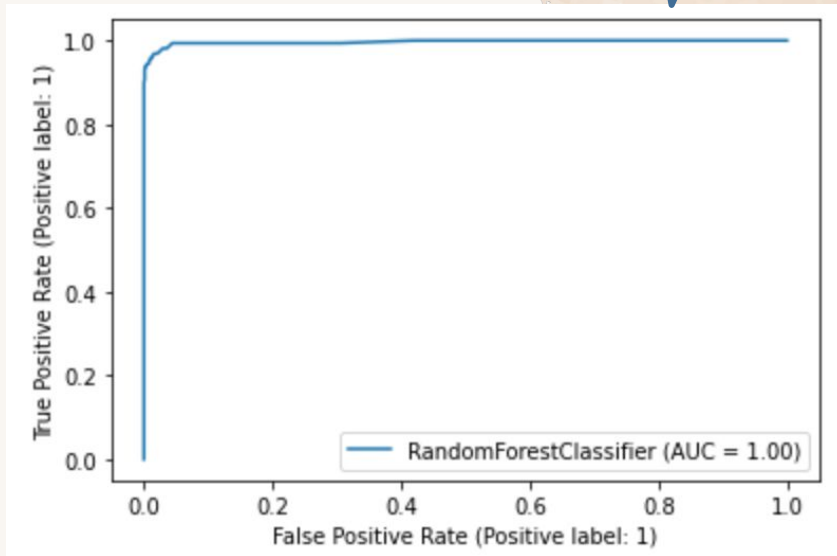
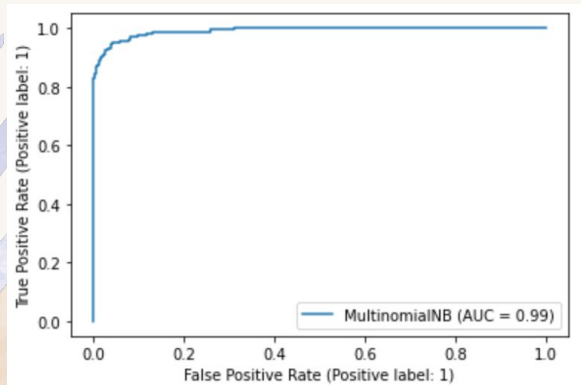
```
['go jurong point crazi avail bugi n great world la e buffet cine got amor wat',  
'ok lar joke wif u oni',  
'free entri wkli comp win fa cup final tkt st may text fa receiv entri question std txt rate c appli',  
'u dun say earli hor u c already say',  
'nah think goe usf live around though',  
'freemsg hey darl week word back like fun still tb ok xxx std chg send rcv',  
'even brother like speak treat like aid patent',
```

Random Forest = WINNER!

About
98% accurate!



(Naive Bayes)



Precision : 1.0
Recall : 0.788
fscore : 0.881

Precision : 1.0
Recall : 0.894
fscore : 0.944

What's the takeaway?

SORT YOUR DATA!

how it can help:



- ★ Helps prevent possible phishing/malware content on software, while operating a business
 - (causing system failure, stealing personal information, etc)
- ★ Creates folders & sorts through content for easier management and navigation

QUESTIONS?

