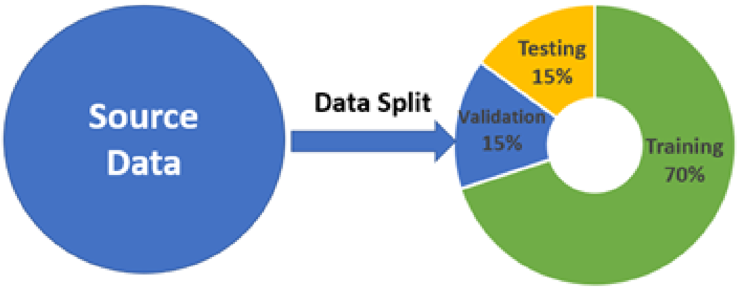
IT 331

Adversarial Attacks - Malicious URLs

Terms:

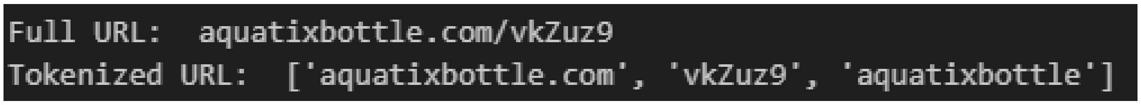
The separation of training and testing data:



- **The training set**
  - Used to train and make the model learn hidden features and patterns in the data
- **The validation set (in theory)**
  - Used to validate the model's performance during training
- **The test set**
  - Used to test the model after training

Tokenization:

- The process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security.



```
Full URL:  aquatixbottle.com/vkZuz9
Tokenized URL:  ['aquatixbottle.com', 'vkZuz9', 'aquatixbottle']
```

It is tokenized based on special characters like "/" and "." (as seen in the example above)

Vectorization:

- A linear transformation which converts the matrix into a vector.

Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']

↓

	The	quick	brown	fox	jumps	over	lazy	dog
Data	2	1	1	1	1	1	1	1

TF-IDF: Term-frequency times inverse document-frequency .

Main:

- **TF(t,d)**: the count of how many times the term appears in the document.
- **IDF** (**I**nverse **d**ocument-**f**requency): helps identify rare words that may be more significant.
- It reduces the importance of common words and applies Euclidean Normalization.
- A matrix out of the token count

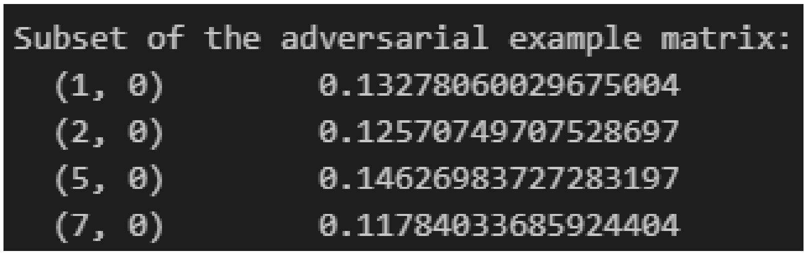
$$\mathrm{TF}(t,d) = \frac{\text{number of times } t \text{ appears in } d}{\text{total number of terms in } d}$$

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

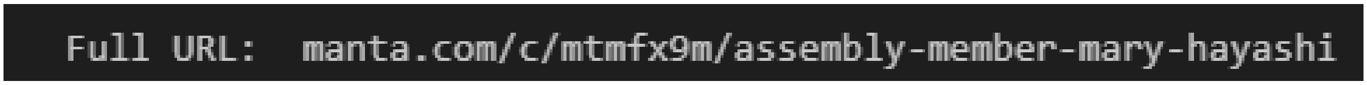
$$\mathrm{IDF}(t) = \log\frac{1+n}{1+\mathrm{DF}(t)} + 1$$

This is the formula used in TF-IDF

A statistical measure that evaluates how relevant a word is to a document.

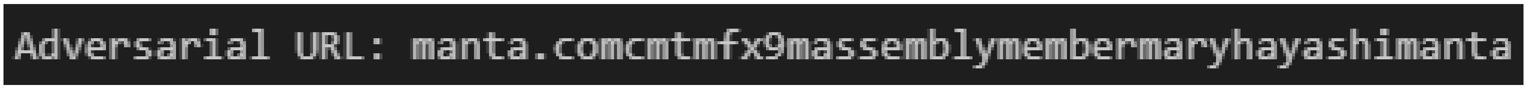


```
Subset of the adversarial example matrix:
  (1, 0)        0.13278060029675004
  (2, 0)        0.12570749707528697
  (5, 0)        0.14626983727283197
  (7, 0)        0.11784033685924404
```



```
Full URL:  manta.com/c/mtmfx9m/assembly-member-mary-hayashi
```

manta.com/c/mtmfx9m/assembly-member-mary-hayashi (Normal URL)



```
Adversarial URL: manta.comcmtmfx9massemblymembermaryhayashimanta
```

manta.com/c/mtmfx9m/assemblymembermaryhayashimanta (Adversarial URL)

IT 331

Minutes from our meetings:

The code was run on the team leader's computer and uploaded on her github.

Mar 4, 2024 11:00 AM

First meeting: discussed the scope of the projects and did general research and formulated questions where we had concerns.

1. What exactly is expected of us
    - Are we detecting the malicious urls or creating them?
    - Creating a model to detect good/bad urls?
    - or creating a model that creates bad urls that can go undetected?
    - Are we corrupting a created model that can detect good/bad URLs
2. When is it due?
3. What is being graded? What do we hand it?

Mar 6, 2024 11:00 AM

Met with the instructor to clarify concerns and fully inquire on the scope of the project.
Notes from the meeting

- we are supposed to create the malicious URLs
- Go to the malware detection code found on the project list
- take the code and generate an adversarial example (?)
- The main concept of the malicious url example posted by him
    - splitting the dataset into training and testing
    - splitting the url "www" "google" ".com" into strings
    - vectorizing the URLs (converting the strings into numbers so the computer can understand it)
        - numbers are assigned based on the closeness of the word "googel" vs "google" would have similar numbers
- he is expecting us to convert a vectorized url back into the string url
- maybe research a bit about machine learning;
    - train class and test class ..
    - how tokenization works ,
    - token to vector …
    - vector to string
    - how to train a model …
    - background knowledge
- Aim for end of the month…best

Mar 15, 2024 11:00 AM

Met to divide workload.

lmutisya@student.concordia.ab.ca - was to run the code and debug

kabdi@student.concordia.ab.ca - was to do research on the terminologies and come up with the master documents template.

Mercyline Lelei - was to do research, write minutes and organize meetings.

Mar 18, 2024 11:00 AM

Met to check progress and further understand the different segments of the code such as the tokenization process and vectorization process.

Apr 13, 2024 12:00 PM

Met to check code functionality and to debug, made the master document and prepared for the presentation.

Additional meetings were done to check on the code and discuss further on the topic.

Apr 17, 2024 11:45 AM

Project evaluation and feedback.

1. Showing the subset of the adversarial example matrix.
2. Showing the adversarial URL in plain text.
3. Adding comments to the code.

Apr 17, 2024 12:00 AM

We worked on responding to the feedback by adding the required sudsets, url and comments. We then posted it on Github.