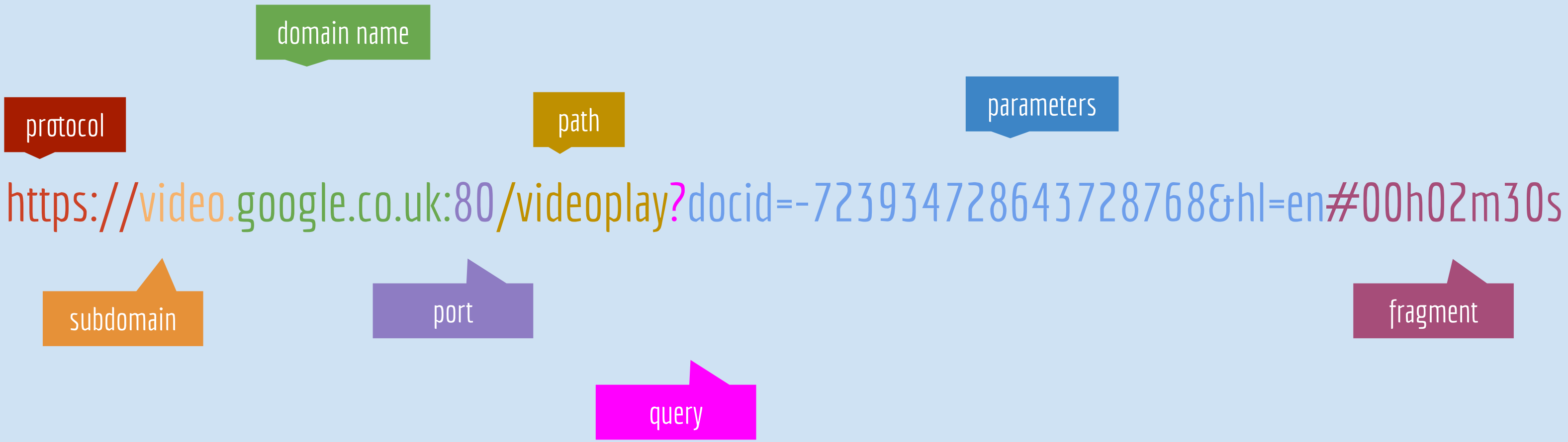
Spam Page Detection

Project Presentation

M151: Web Information Retrieval

George Panagiotopoulos   Maria Despoina Siampou

Introduction: URL



domain name

protocol

path

parameters

https://video.google.co.uk:80/videoplay?docid=-723934728643728768&hl=en#00h02m30s

subdomain

port

fragment

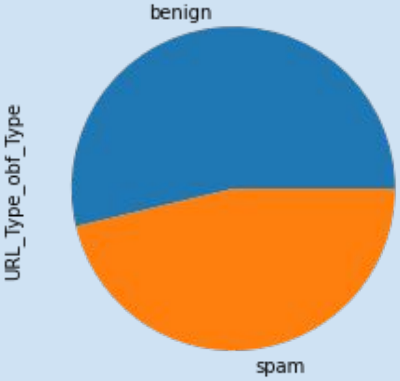
query

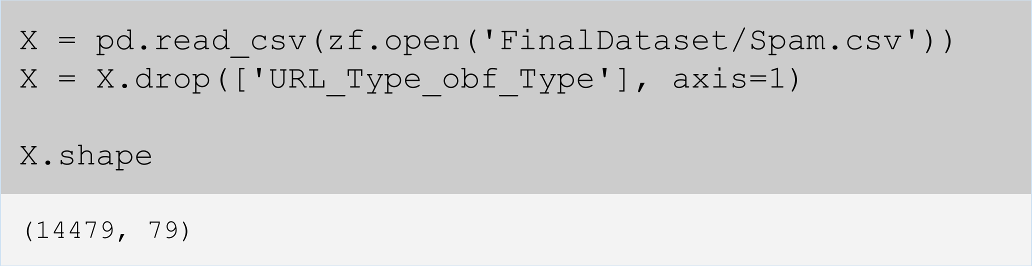
Introduction: URL

SPAM URL !

http://paypal.com.webscr.cmd.login.submit.dispatch.5885d80a13c0db1f8e263663d3faee8db2b24f7b84f1819343fd6c338b1d9d.
222studio.com/UK/

ISCXURL2016 Dataset: Overview

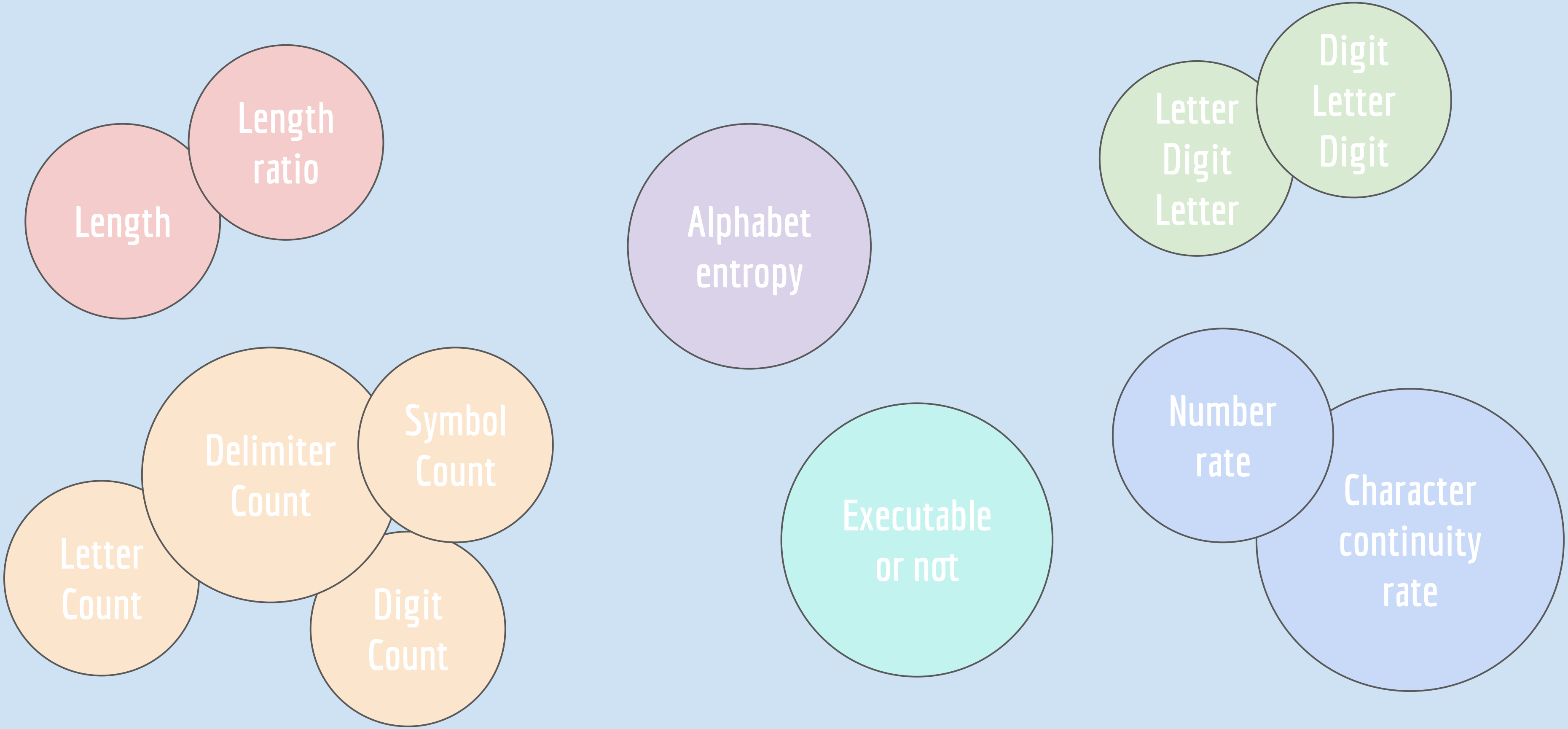




```
X = pd.read_csv(zf.open('FinalDataset/Spam.csv'))
X = X.drop(['URL_Type_obf_Type'], axis=1)

X.shape
```

```
(14479, 79)
```

ISCXURL2016 Dataset: Features



Length

Length ratio

Alphabet entropy

Letter Digit Letter

Digit Letter Digit

Letter Count

Delimiter Count

Symbol Count

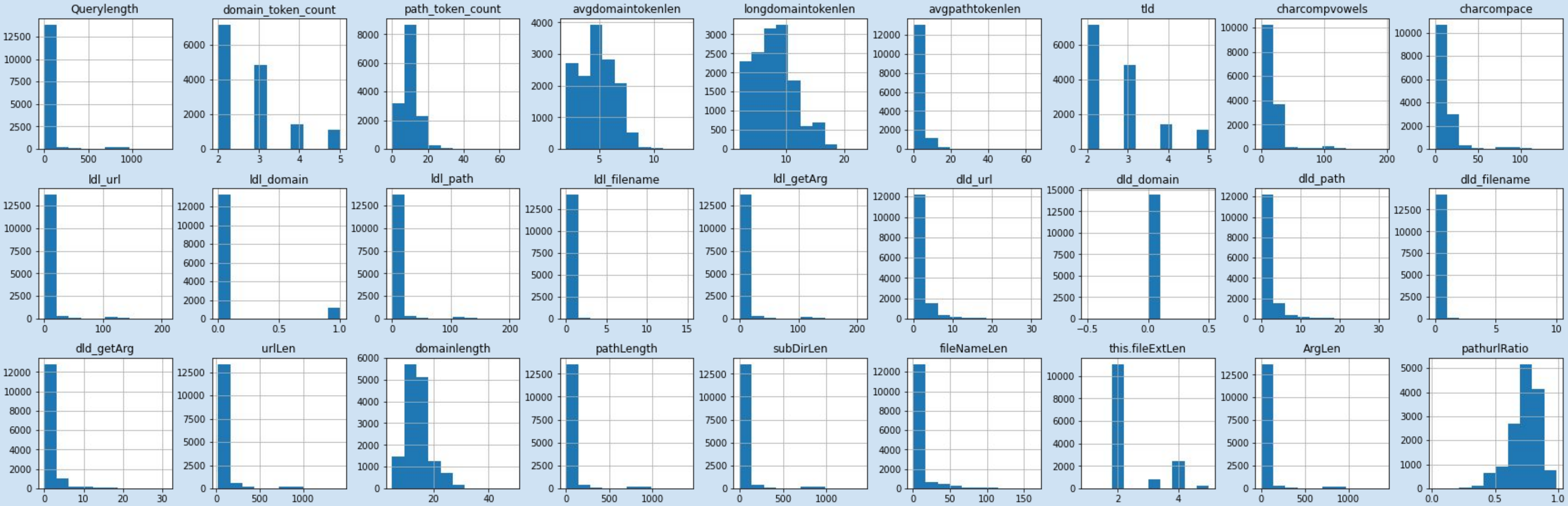
Digit Count

Executable or not

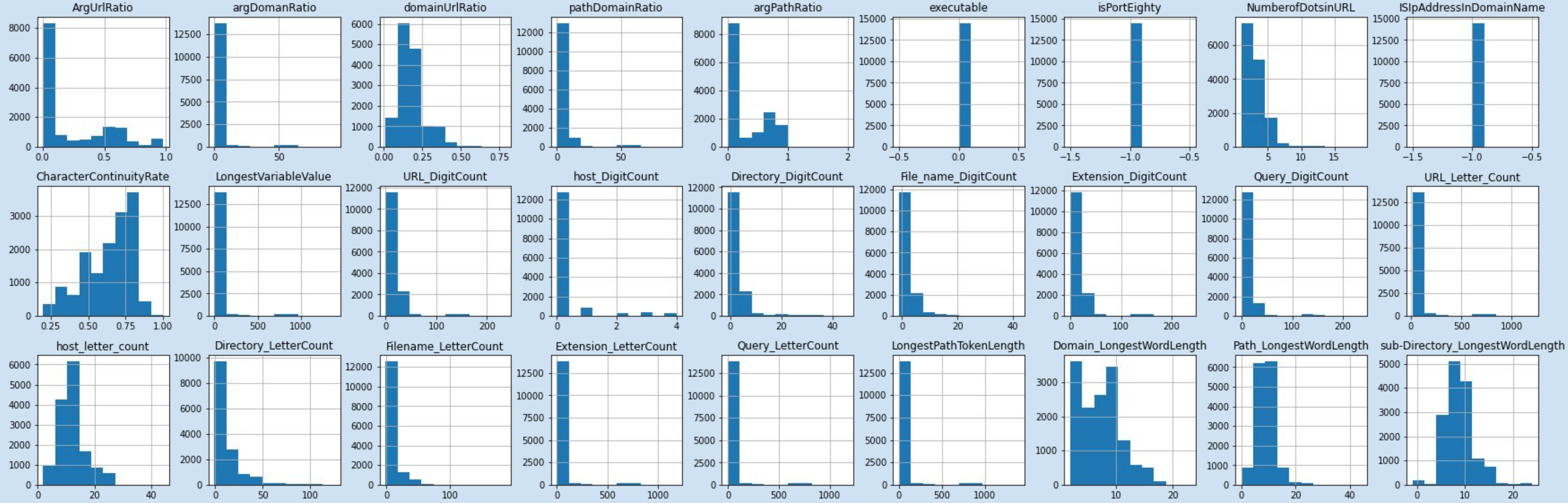
Number rate

Character continuity rate

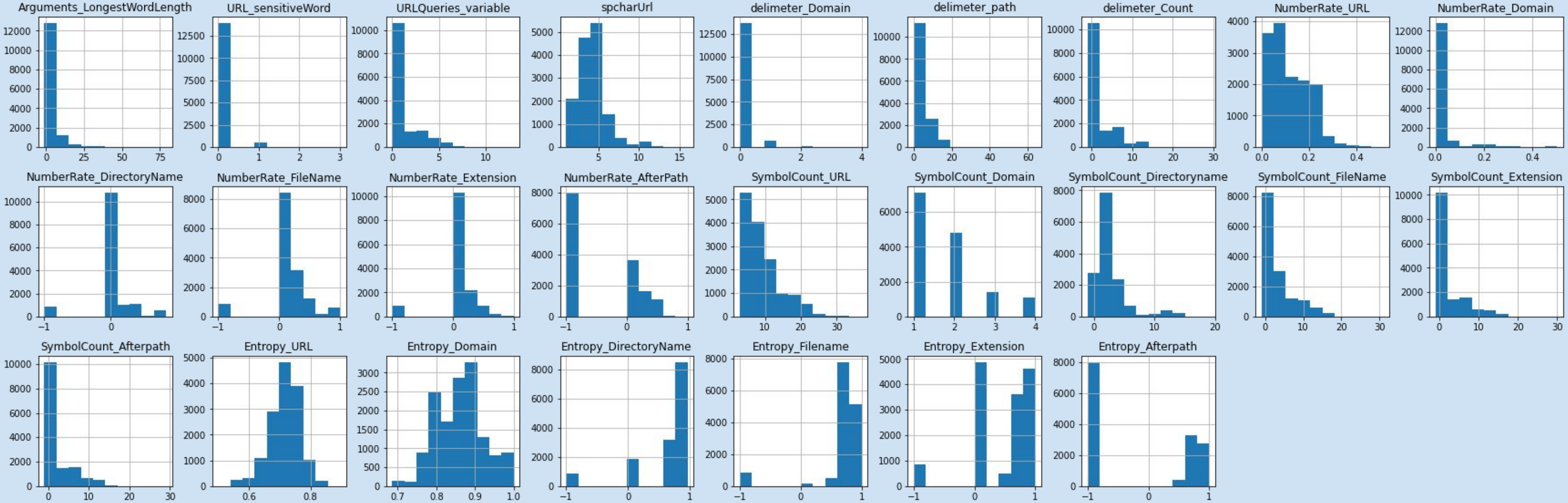
ISCXURL2016 Dataset: Feature exploration (1/3)



ISCXURL2016 Dataset: Feature exploration (2/3)



ISCXURL2016 Dataset: Feature exploration (3/3)



ISCXURL2016 Dataset: Features

Having a big number of features may:

- Lead to overfitting
- Suffer from curse of dimensionality
- Consume time to compute and process unnecessary features

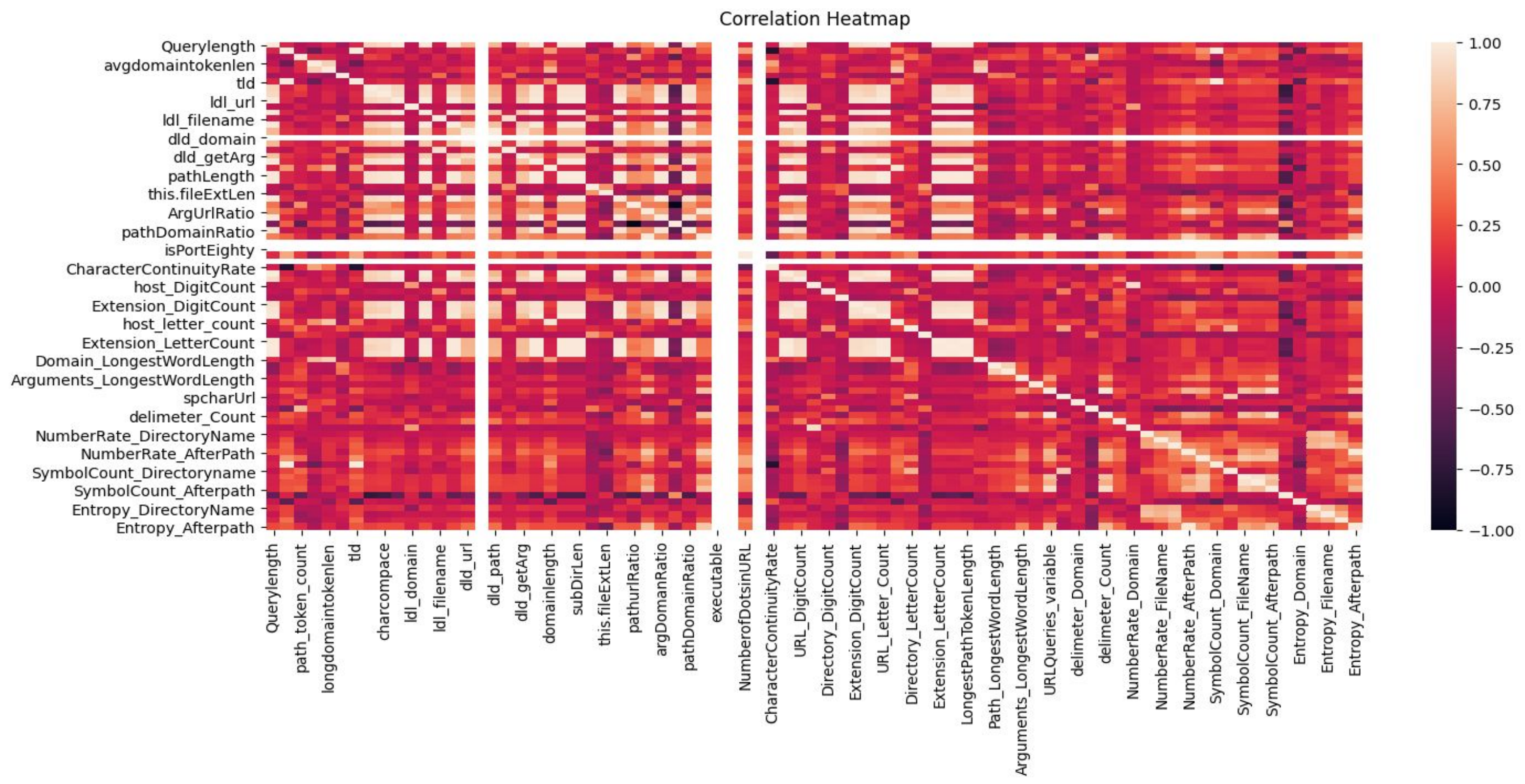
ISCXURL2016 Dataset: Features

Having a big number of features may:

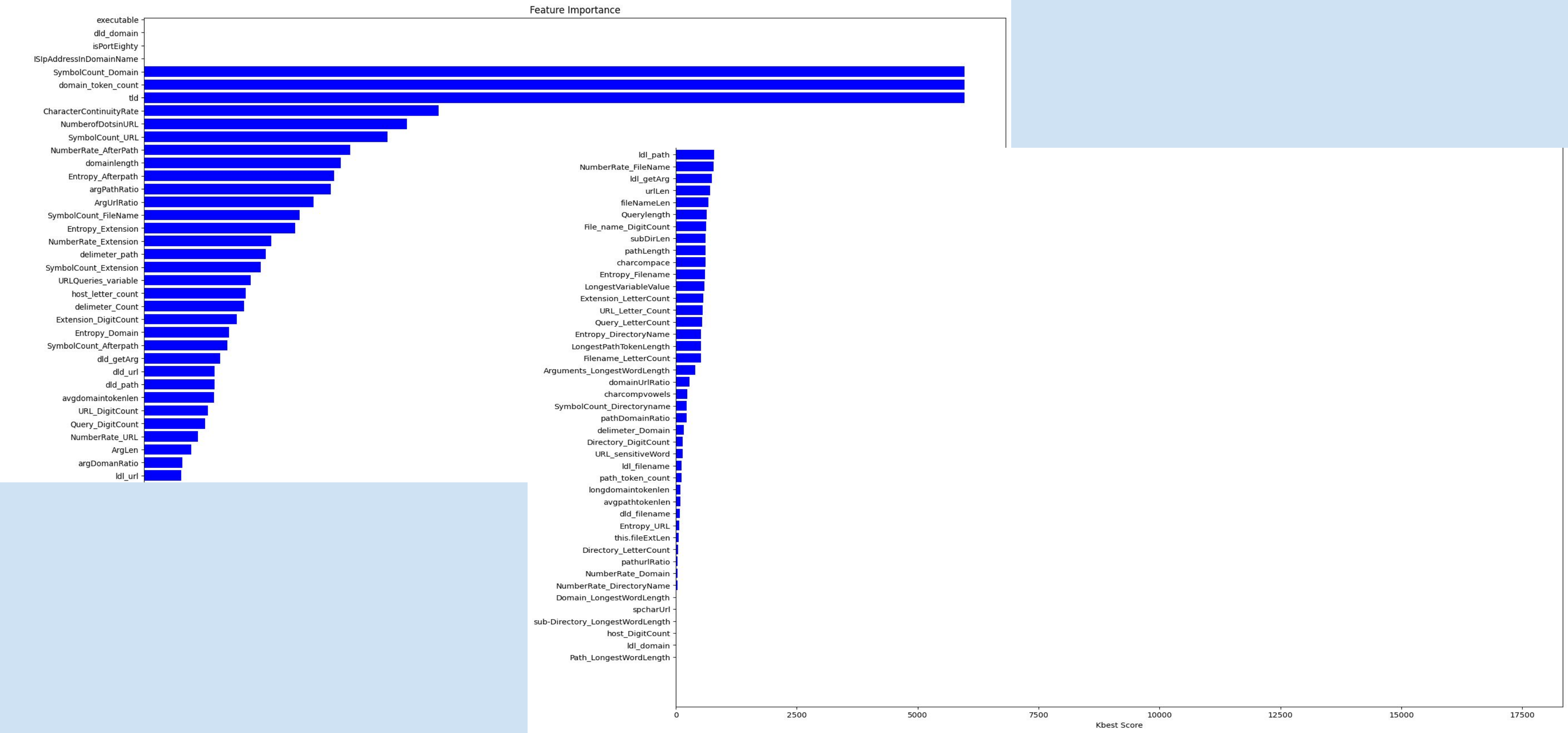
- Lead to overfitting
- Suffer from curse of dimensionality
- Consume time to compute and process unnecessary features

→ Keep only meaningful features

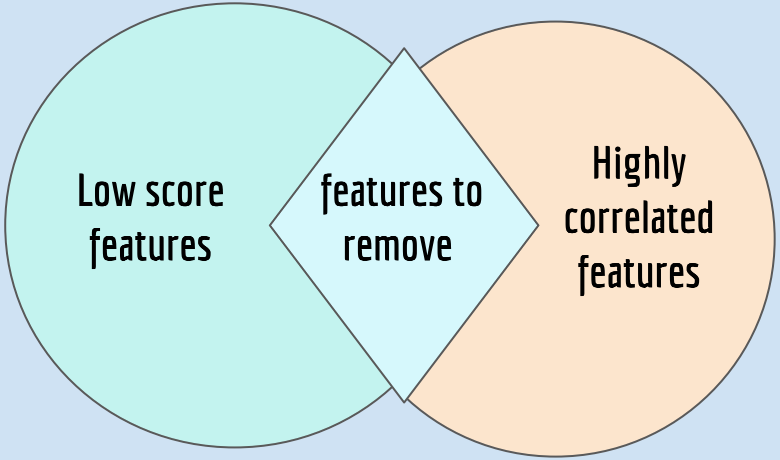
Feature Selection: Feature Correlation

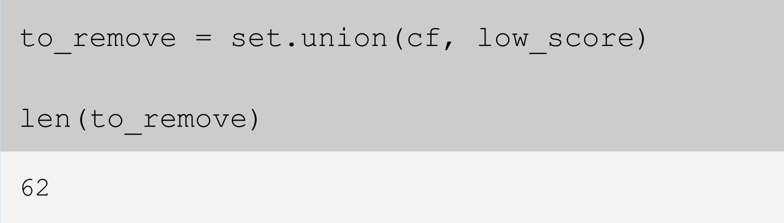


Feature Selection: K- best Features



Feature Selection: Removed Features





```
to_remove = set.union(cf, low_score)

len(to_remove)
```

```
62
```

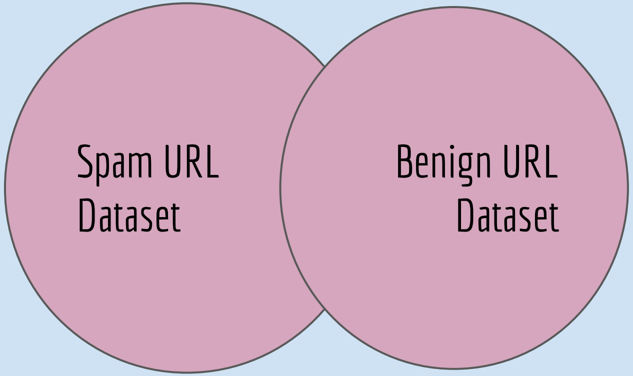
Feature Selection: Results

1. domain token count
2. average domain token length
3. digit letter digit pattern count url
4. domain length
5. argument - url ratio
6. number of dots in url
7. character continuity rate
8. url queries variable
9. delimiter count path
10. number rate url
11. number rate filename
12. number rate extension
13. number rate afterpath
14. symbol count url
15. symbol count filename
16. entropy domain
17. entropy extension

Classifier Selection: 10-Fold Cross Validation

Classifier	Mean F1-Score
bagging-dtree	0.997
decision-tree	0.995
knn	0.994
logisticreg	0.986
naive-bayes	0.941
random-forest	0.998
linear-svc	0.988
voting-classifier	0.988

Experiments: Creating a new dataset



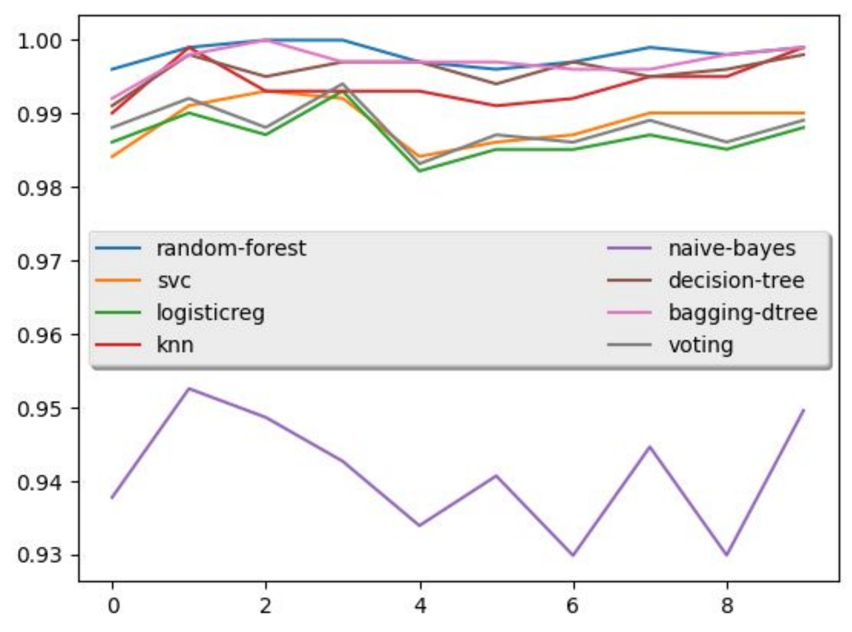
```
X_ = Benign.append(Spam)

X_.shape
```

```
(47378, 2)
```

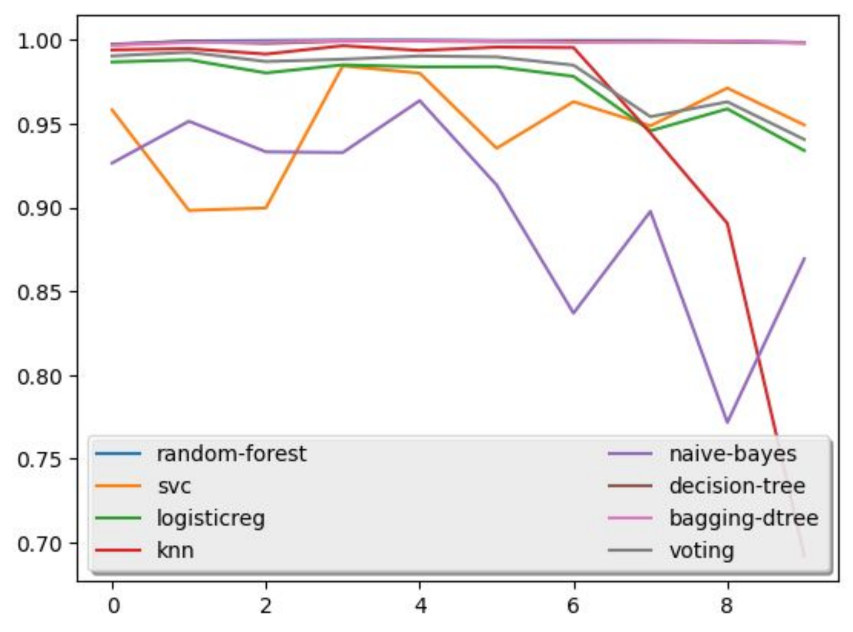
```
X_new = feature_extraction(X_)

X_new.shape
```

```
(47378, 17)
```

Experiments: Datasets' Comparison

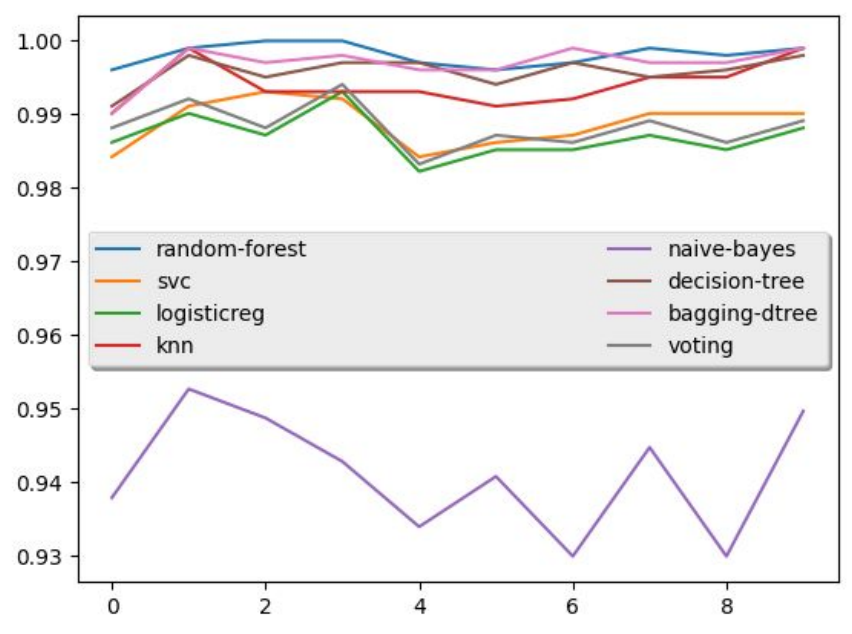


F1-Scores of ISCXURL2016 dataset

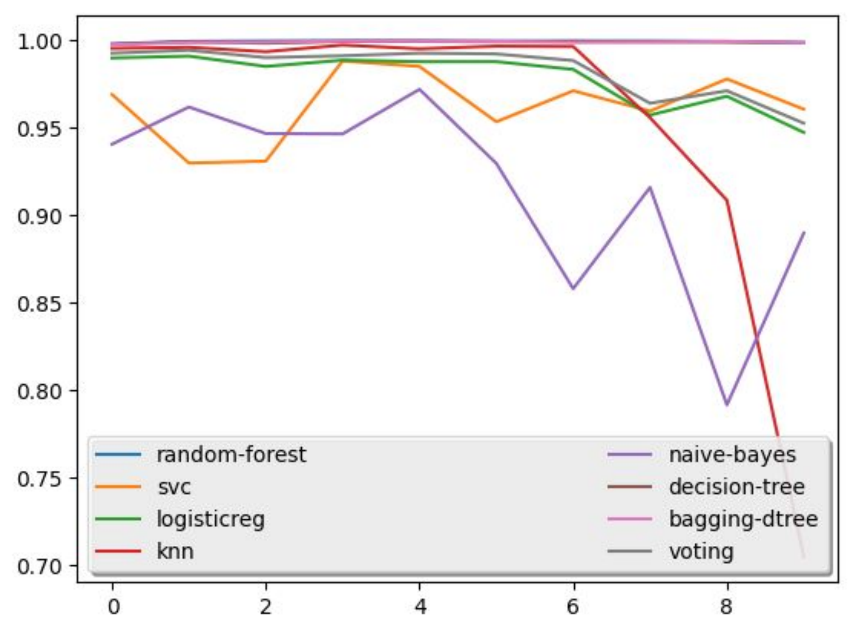


F1-Scores of Spam+Benign dataset

Experiments: Datasets' Comparison



Accuracy of ISCXURL2016 dataset



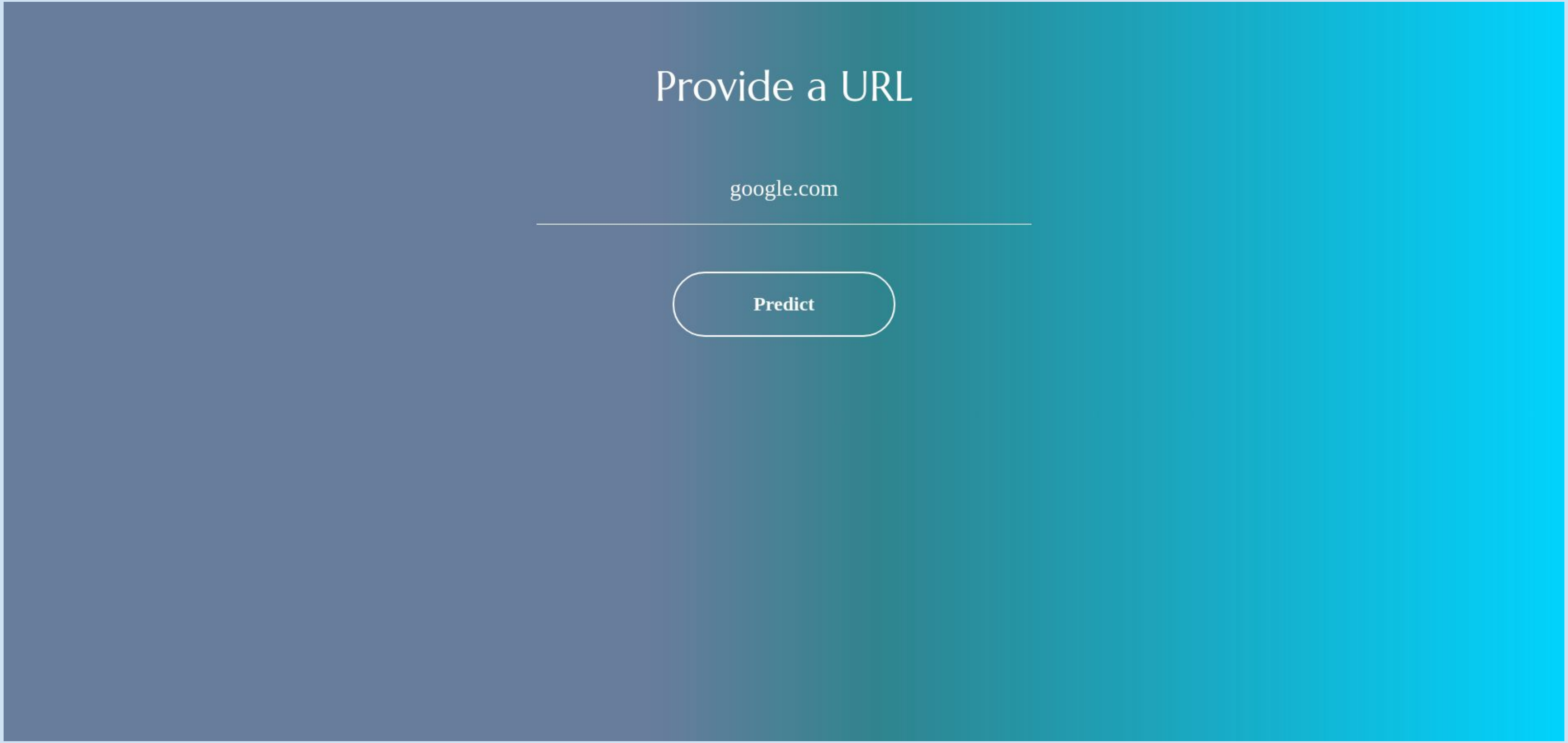
Accuracy of Spam+Benign dataset

Experiments: Results

Parameters	Values
Classifier	Random Forest
Number of features	17
Accuracy	0.996
Precision	0.999
Recall	0.997
F1 - Score	0.998

17 features are enough !

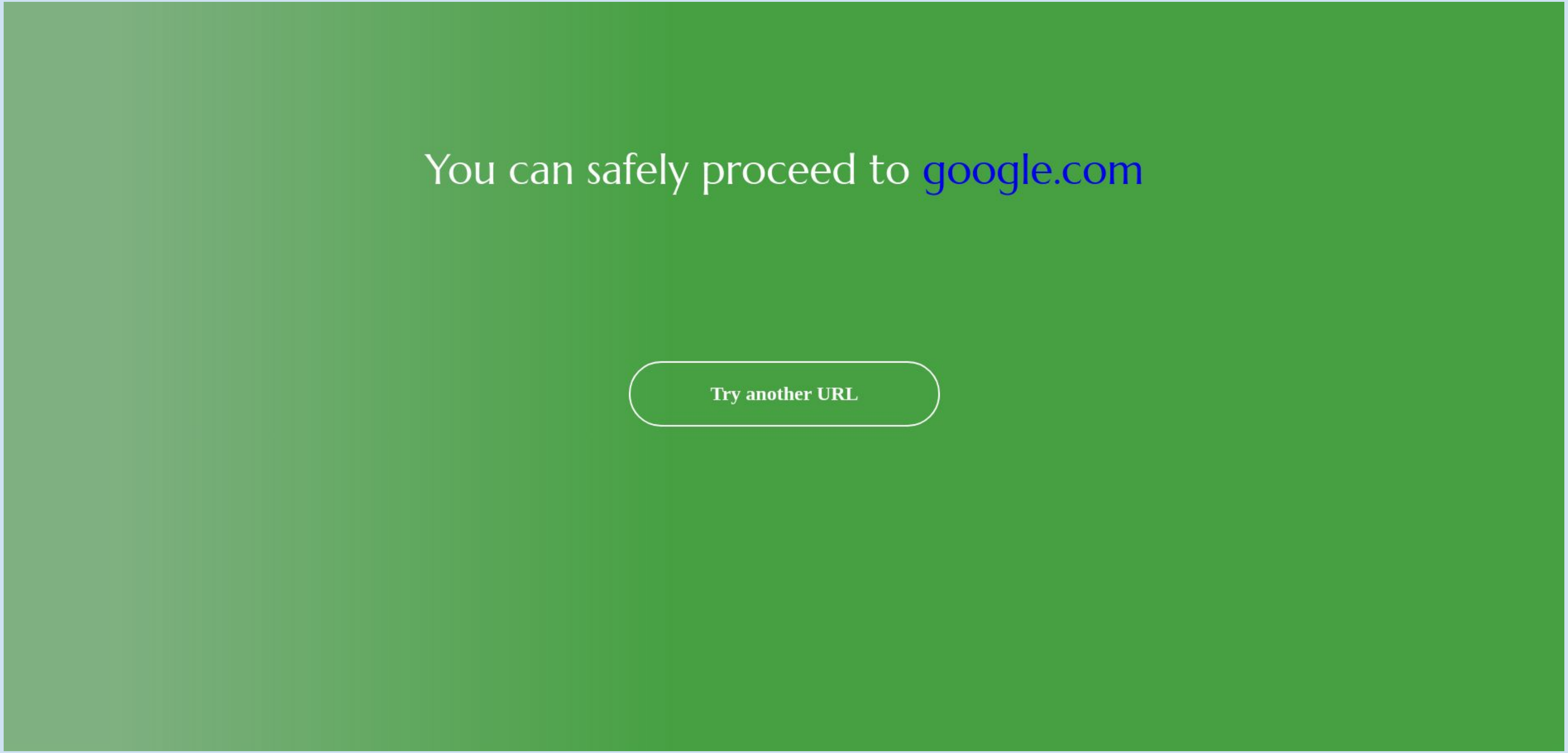
UI Demonstration



Provide a URL

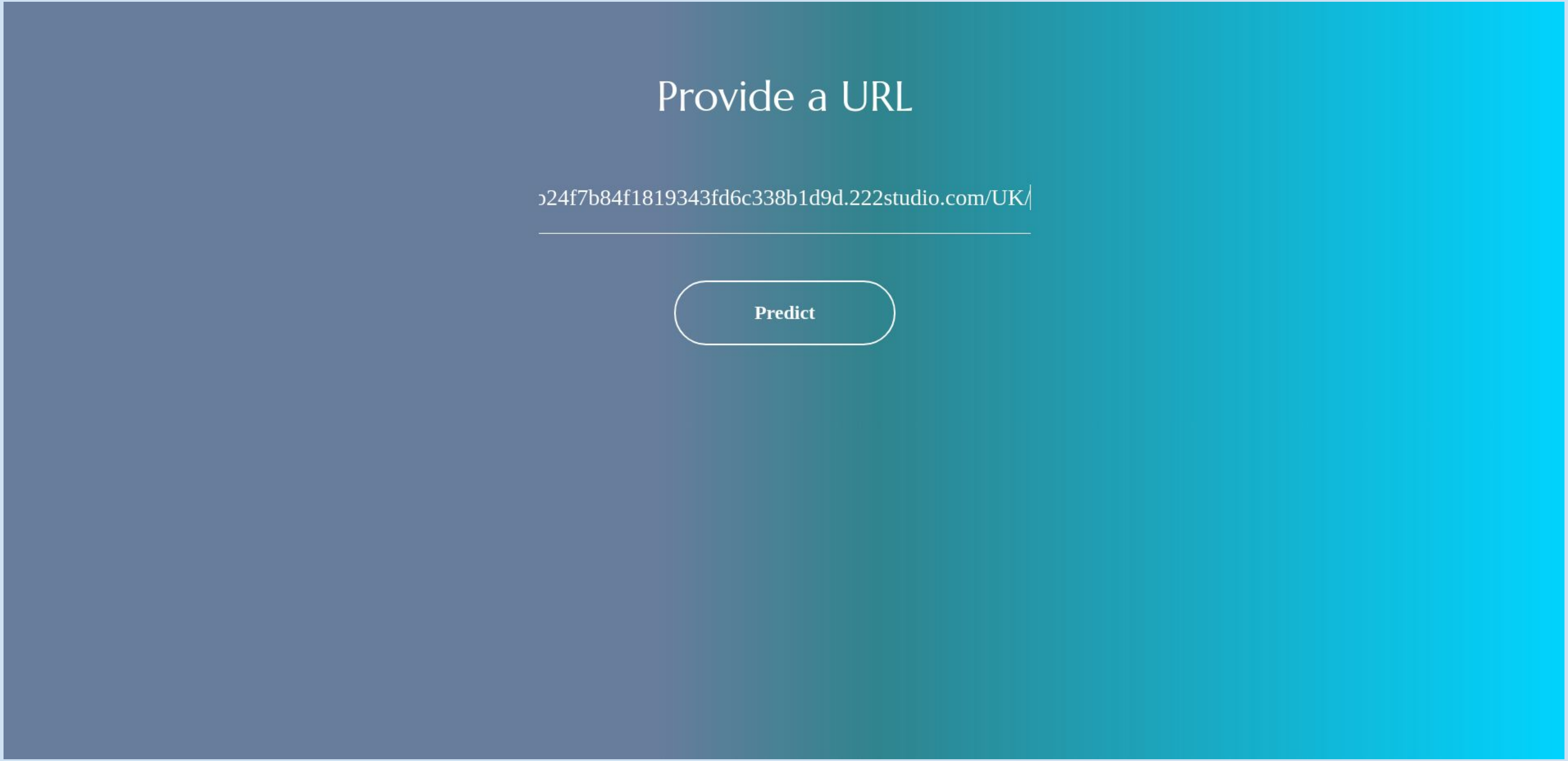
google.com

UI Demonstration



You can safely proceed to google.com

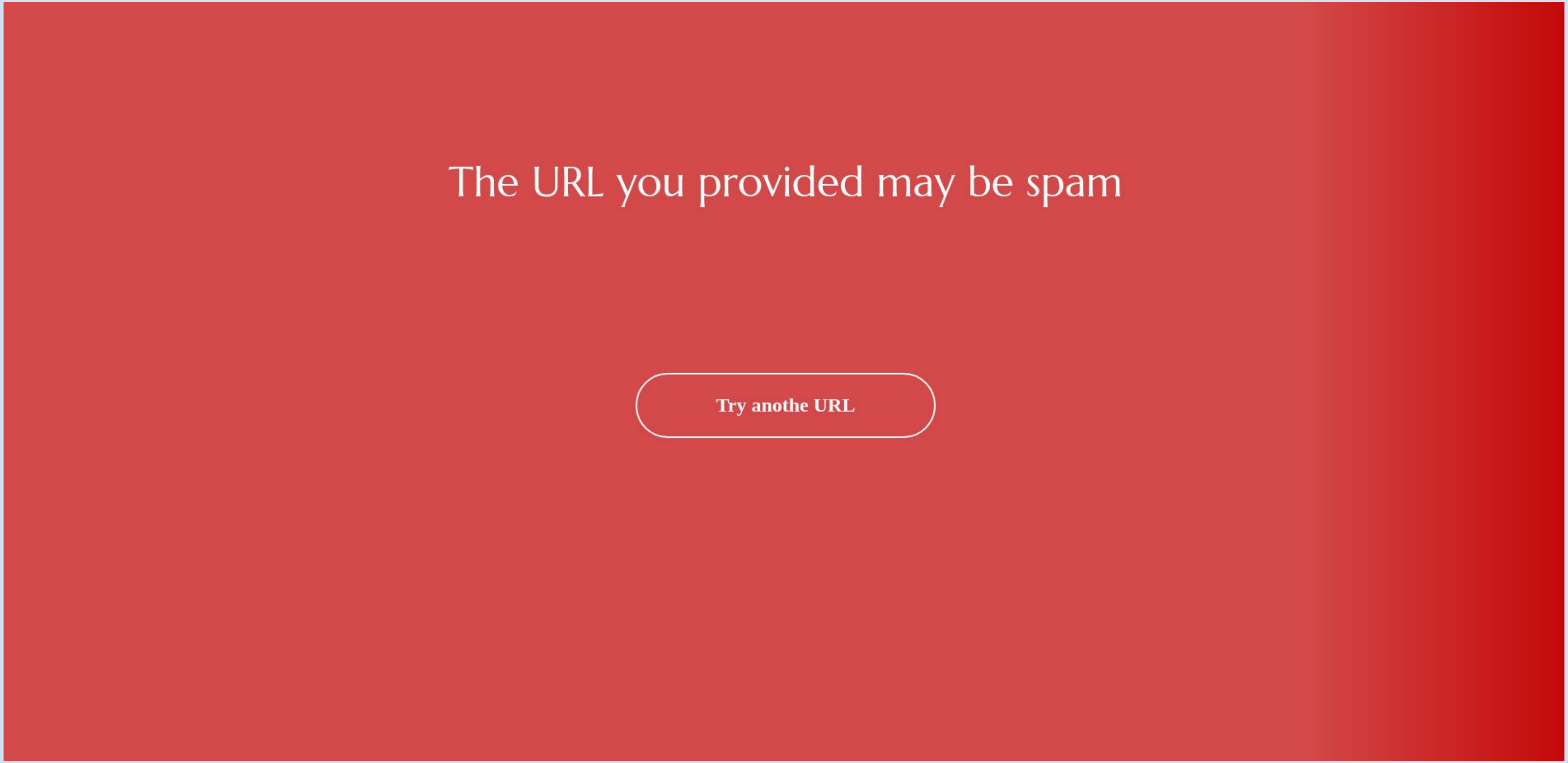
UI Demonstration



Provide a URL

24f7b84f1819343fd6c338b1d9d.222studio.com/UK/

UI Demonstration



The URL you provided may be spam