Agenda

To solve a challenging Application security automation problem.

CLOUDSEK

Making Machines Think about Security

Core Team

Mohanlal Menon [CEO and Founder]

- 25 years of experience as an Entrepreneur, Angel Investor, Executive/Board level Coach & Mentor.
- Formerly: Dell, Eicher, Dupont and APC

Rahul Sasi [CTO and Founder]

- Security Researcher.
- Invited speaker on information security issues in 28 global forums at 17 different countries.
- Formerly: Citrix and iSight .

Most seen security issue on Facebook

insecure direct object

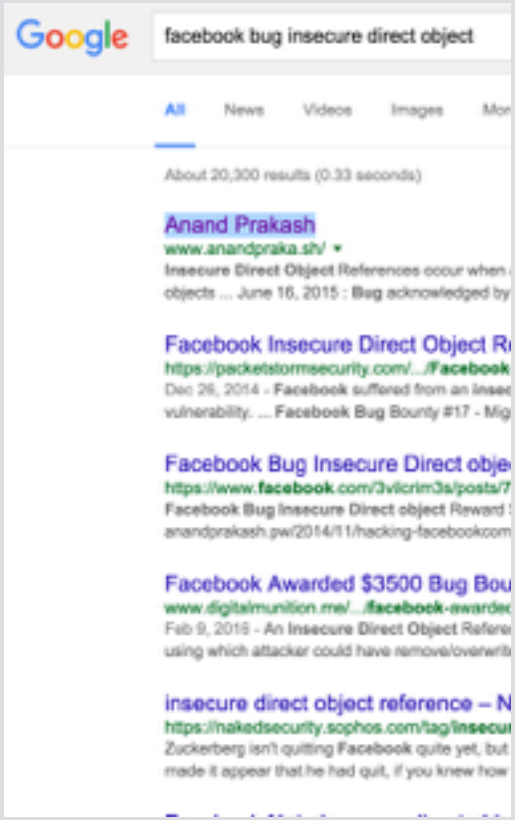
Attacker user-id = 1234

Victim user-id = 6666

Attacker replaces attacker user-id with victims and performs an action on his behalf.

Hard to automate

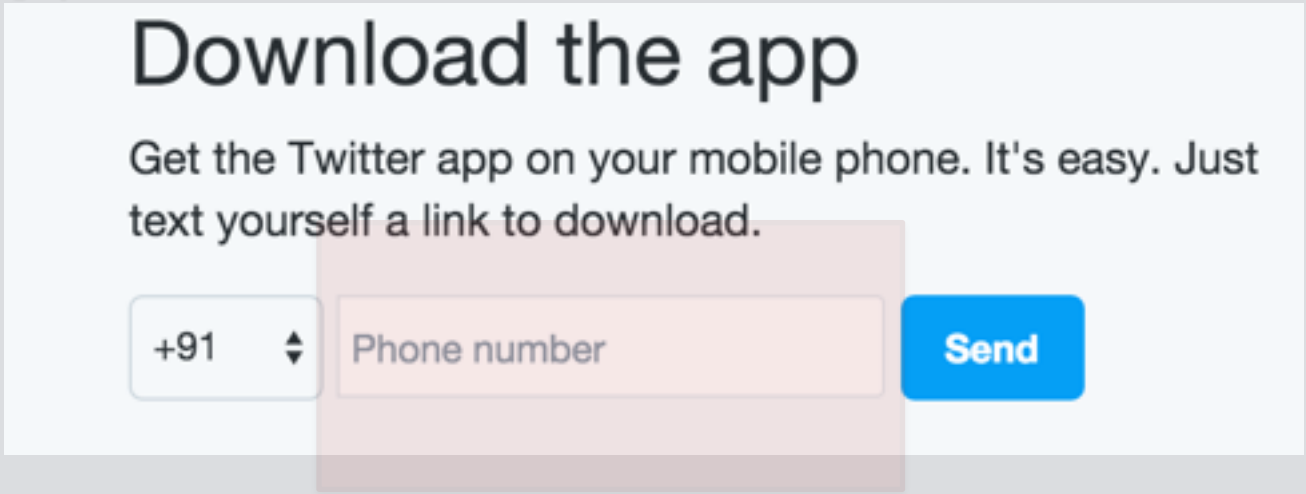
- Such bugs often get left out by automated tools.
- Manually testing for such bugs is time consuming.



The Problem.

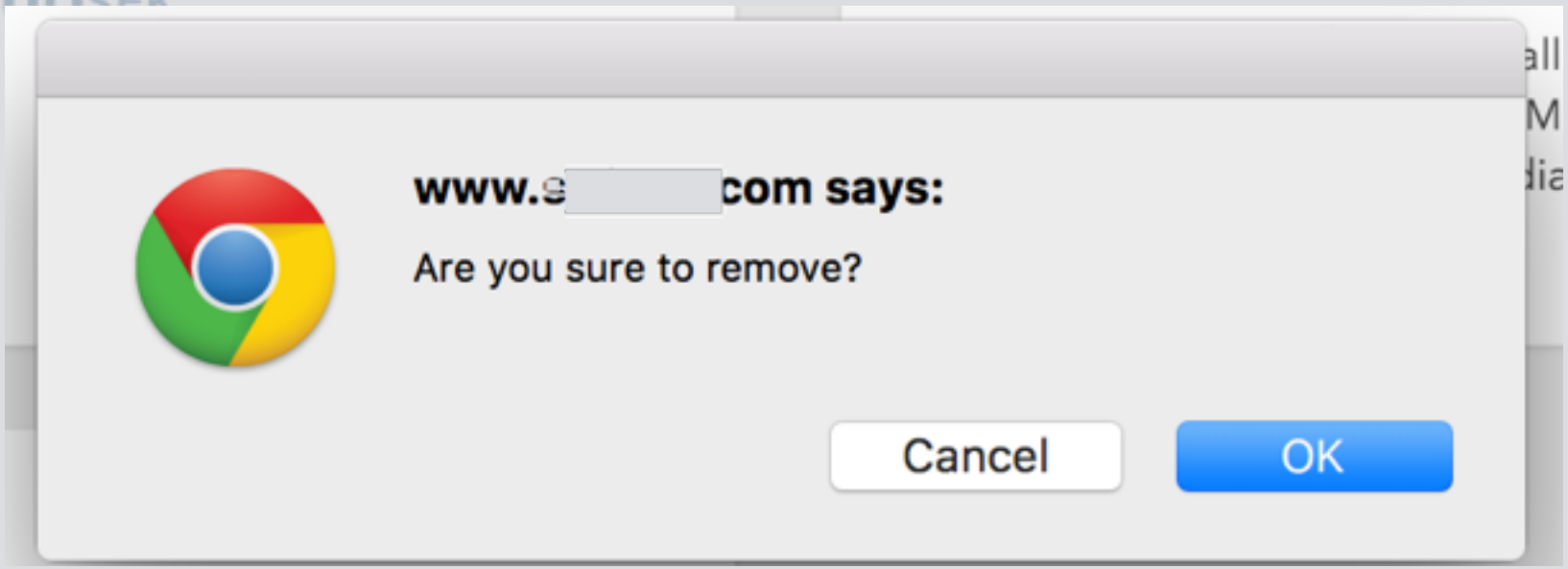
- Application scanners are blind and are not intelligent.
- Scanner do not understand the application they are scanning.
- Scanners treat every webpage as an html form and nothing more.
- Scanners currently understand very little of the meaning of human language.
- Modern web applications run on Javascript.

The Problem Example:



- POST /update_phone.php {phone:"9999313373" }
  GET/update_phone_sucess.php?user_name=cloudsek

The Problem Example:



- GET /app/delete_data.php?uid=1337

Disclaimer

- We are not re-writing application security automation methods.
- We are trying to improve it.
- We are simply trying to make security automation more human.

Solution

- Build programs that are capable of understanding and using Web Applications like a normal human being .
- Using [ML] Machine Learning and [NLP] Natural Language processing to build this highly ambitious project.

AI = systems that can do intelligent things

NLP = systems that can understand language

ML = systems that can learn from experience

NLP ∩ ML = systems that can learn how to understand language.

Supervised ML= systems that can predict from known input

UnSupervised = systems that can predict from unknown input

Building Next generation automation tool

We wanted our overly ambitious tool to perform the following.

- Understand and differentiate between two different functionalities.

ex: differentiate a login page and forgot password page.

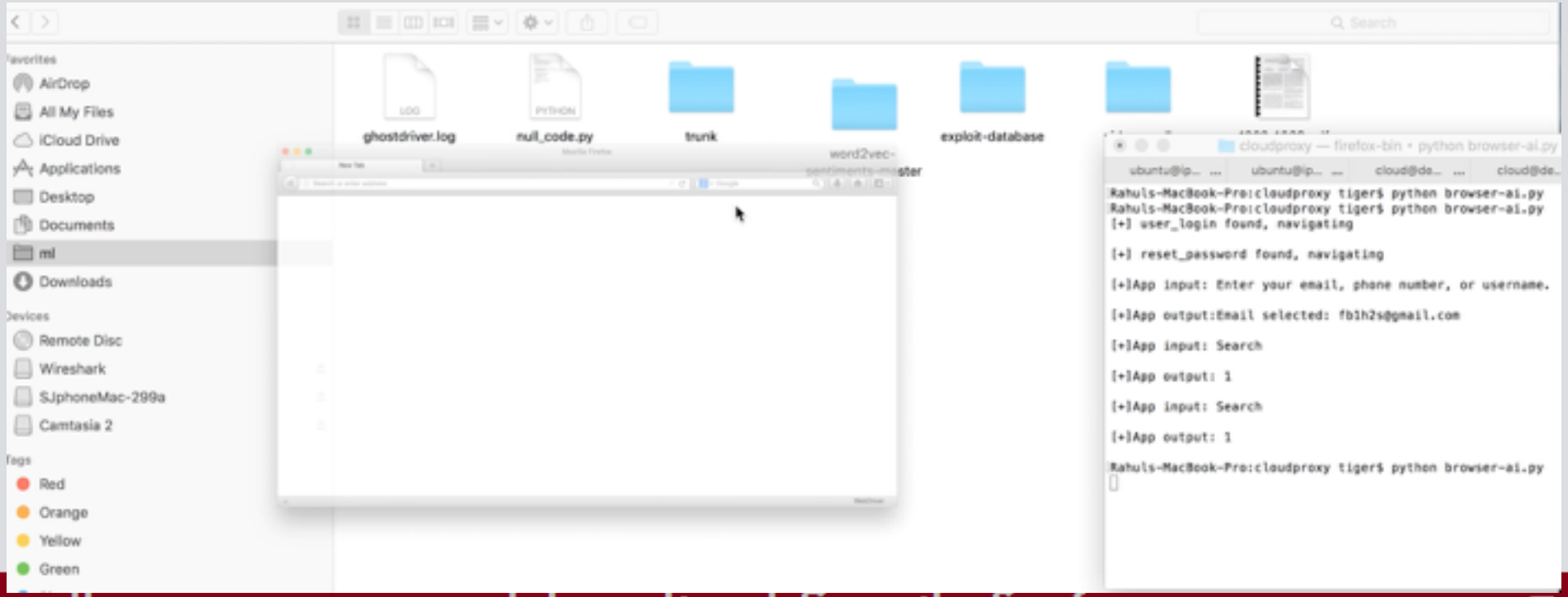
- Crawl [Navigate] a web application the same way a normal human being uses it

ex: first register a new user and then proceed to login

or traverse multiple pages to reach final end point .

Browser AI Functioning

- Can identify natural language and identify instructions.
- Can pass in the right input based on application request.
- Browser AI on twitter navigating forgot password option.



Implementation Details

- NLP: Identifying functionalities of an application.

d1: Sign Up

d2: Create new user.

d3: Register now!

Register_User()

d1: Forgot password

d2: Reset password

d3: Forgot your password

d4: Reset Now!

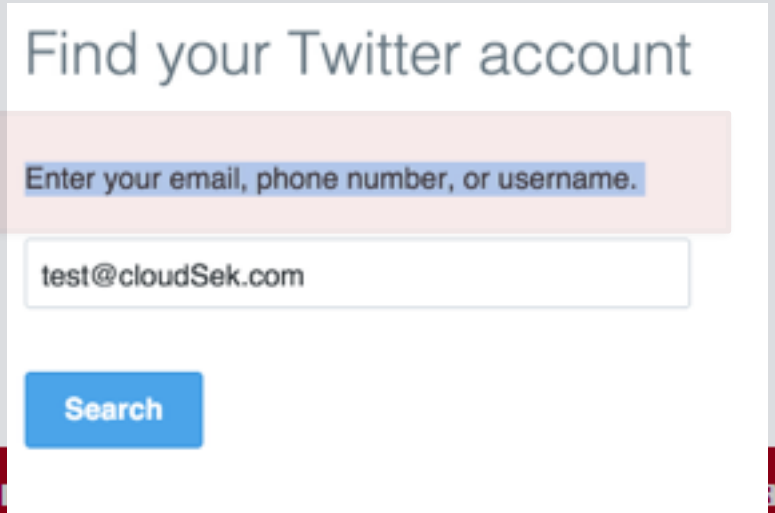
Forgot_password()

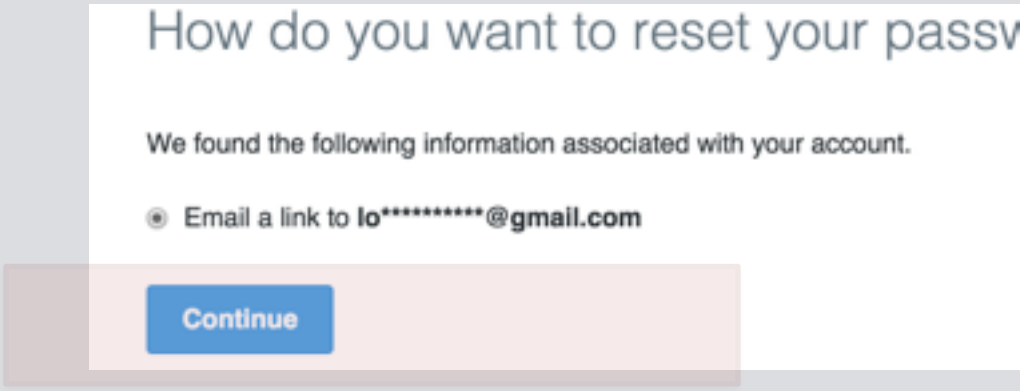
Implementation Details

- NLP: Making tools understand the scanned application .

1) Passing correct information, based on what application is expecting.

2) Identifying and responding to decision making modules.





Building The Intelligent system

Preprocessing

- Identifying and extracting instructions.
- text to lowercase,
- token extraction,
- filter stop words,

d1: sign up

d2: create new user.

d3: register now!

Register_User()

d1: forgot password

d2: reset password

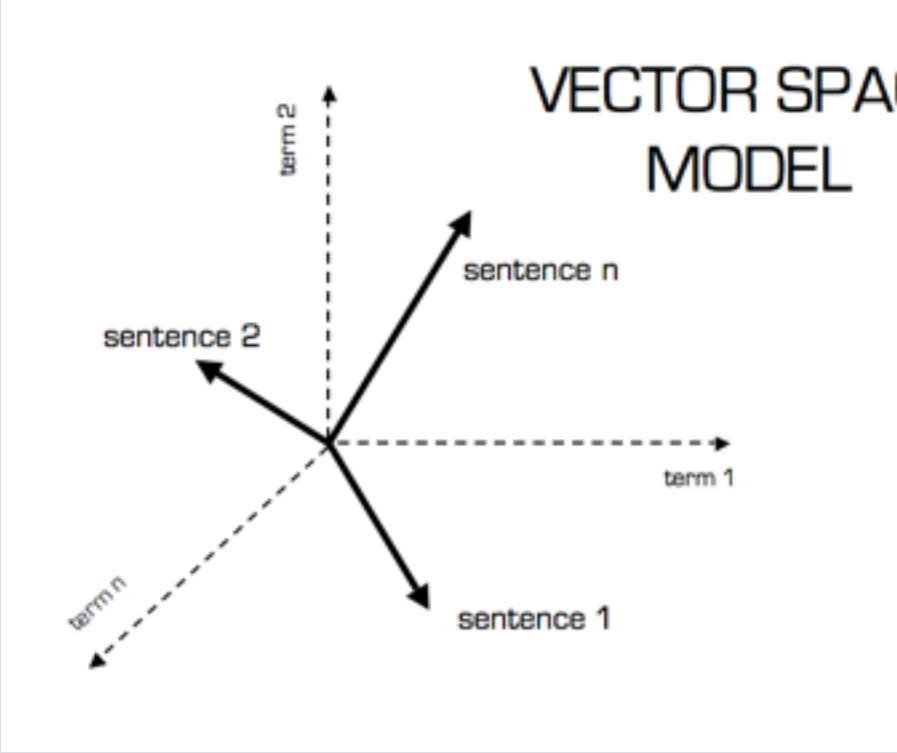
d3: forgot your password

d4: reset now!

Forgot_password()

NLP/ML Basics

- Convert the sentences to some kind of numeric representation for machine learning
- VSM, is a space where text is represented as a vector of numbers instead of its original string.
- Building a supervised models.
- Building training data.
- Running our ML models on our trained data.



Vectorisation

A vector could represent,

- the importance of a term (tf–idf) in a document.
- the absence or presence (Bag of Words) of a word in a document
- hashing vectoriser .

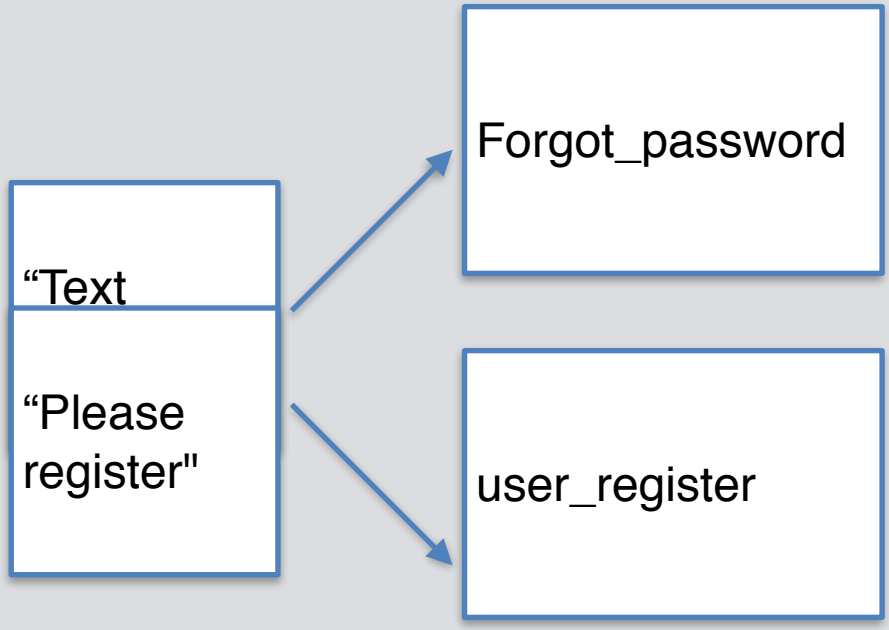
TF-IDF vectorisation

- TF: Term Frequency, which measures how frequently a term occurs in a document.
  - TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).
  - Ex: Document = “How is null conference better than other conferences”
  - TF(Conference) = 2 / 8 = X
- IDF: Inverse Document Frequency, which measures how important a term is.
  - IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

$$tfidf = TermFrequency x log\left(\frac{TotalDocuments}{DocumentFrequency}\right)$$

Classification.

What a Classifier does

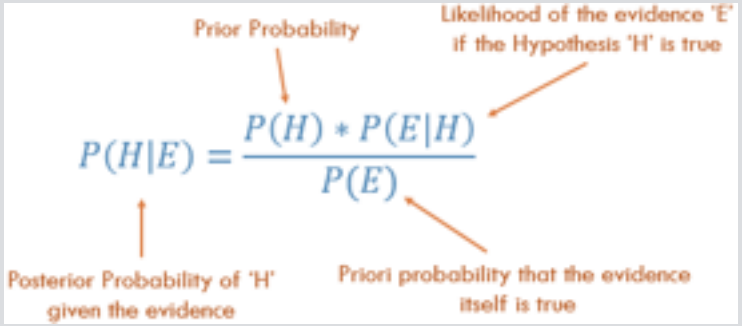


How a Classifier works

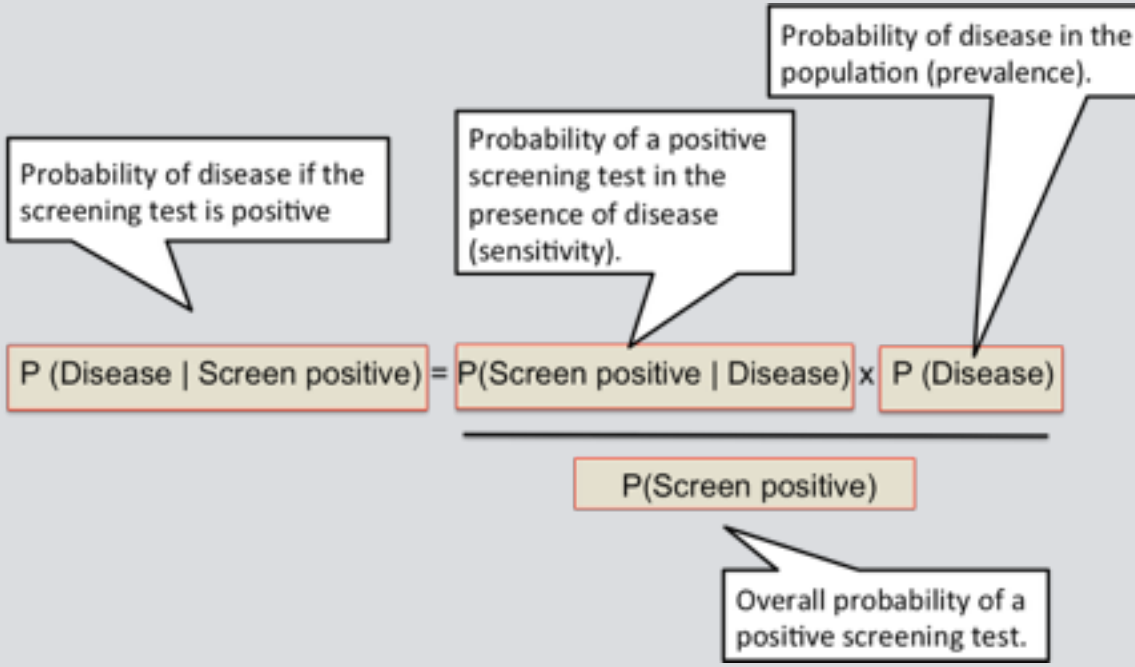
Temp (Degrees C)	Rain	Play?
15	No	Yes
23	Yes	Yes
-6	Yes	No
-6	No	Yes

- Given documents d1,d2,d3 predict a tag.
- Different ML classifiers work for this problem.
  - Naive Bayes (NB)
  - Support Vector Machines (SVM)
  - Maximum Entropy
- In our testing we had better results with Naive Bayes

(Reverent Thomas Bayes 1702-1761)



Bayes Theorem Explained



Building Training data

- Initially we extracted and labeled data manually

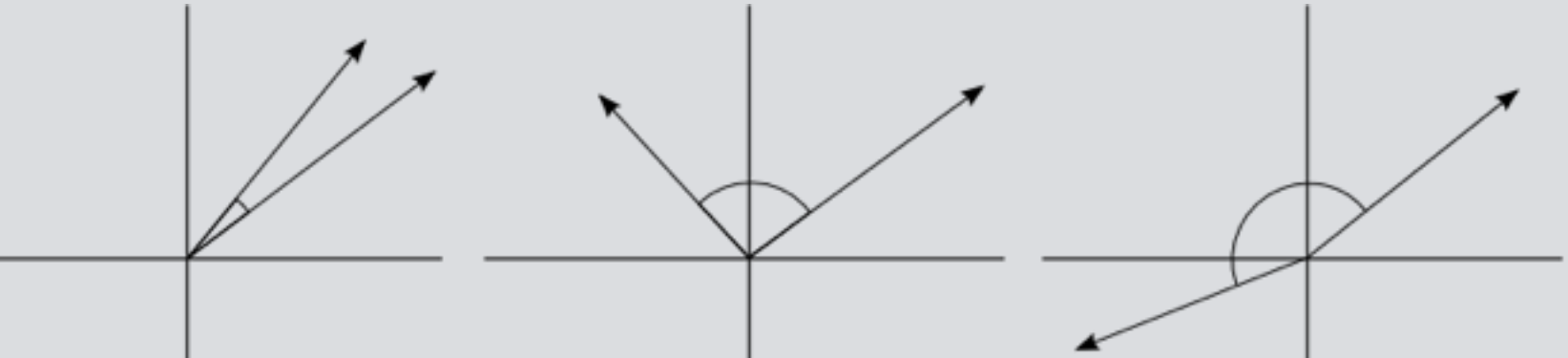
Training Data	Label
forgot password	[password-reset]
Create new user.	[register_user]
Log in here	[Login_user]
reset password	[password-reset]
Sign Up	[register_user]
Forgot your password	[password-reset]
Log in to your store	[Login_user]
Reset Now!	[password-reset]
Forgot your username or password?	[password-reset]
Log me in	[Login_user]
Request Password reset	[password-reset]
forgot credentials	[password-reset]
Register now!	[register_user]

The Cosine Similarity

- calculation of the angle between two vectors.
- a metric that says how related are two documents by looking at the angle.
- documents with the closest cosine similarity would be similar documents

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$



Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%

Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%

Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

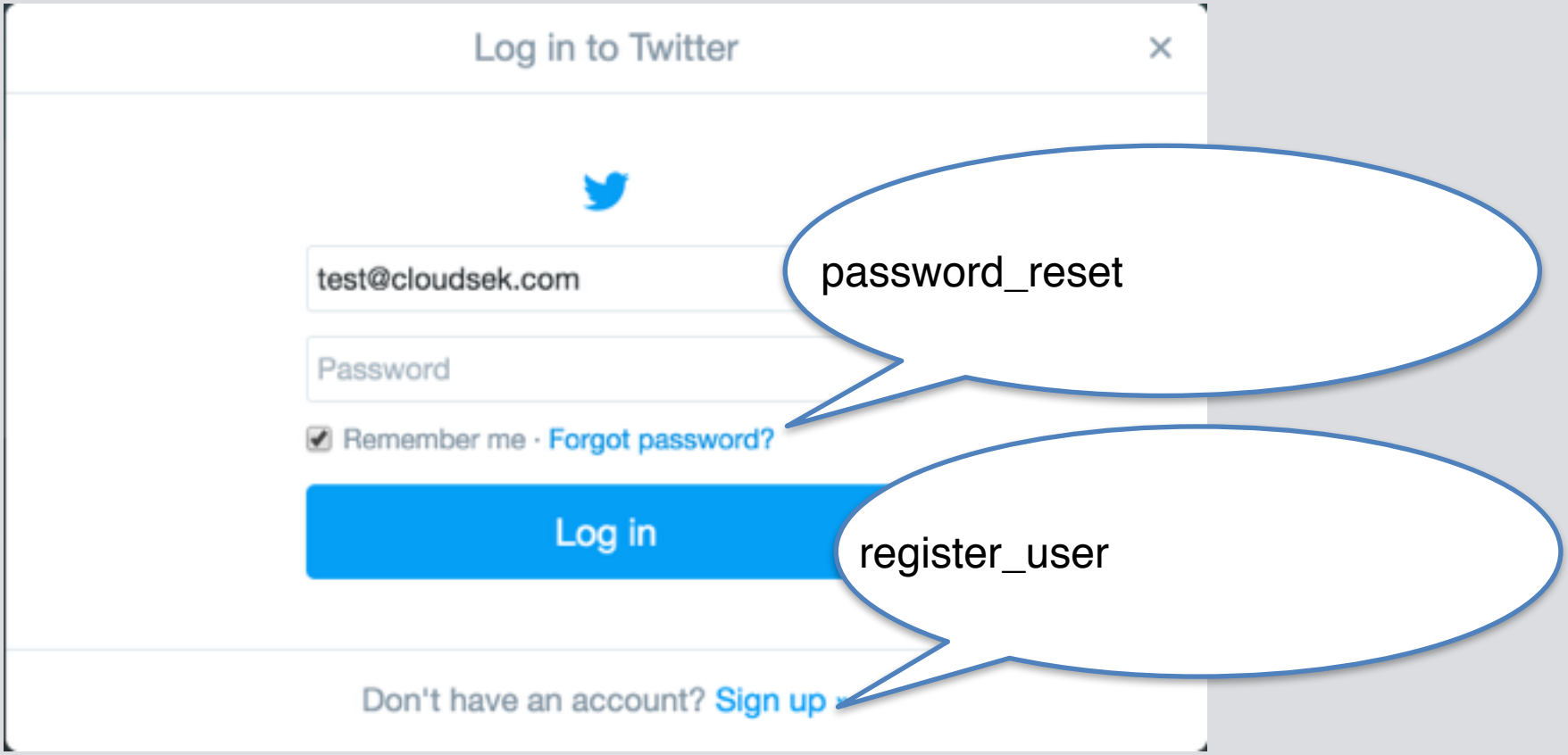
Building Training data

- We used Cosine similarity to improve the training data automatically.

Training Data	Label	Cosine Similarity
forgot password	[password-reset]	
Forgot your password	[password-reset]	
Reset Now!	[password-reset]	“Forgot user password”
Forgot your username or password?	[password-reset]	“Change user password”
Request Password reset	[password-reset]	
forgot credentials	[password-reset]	

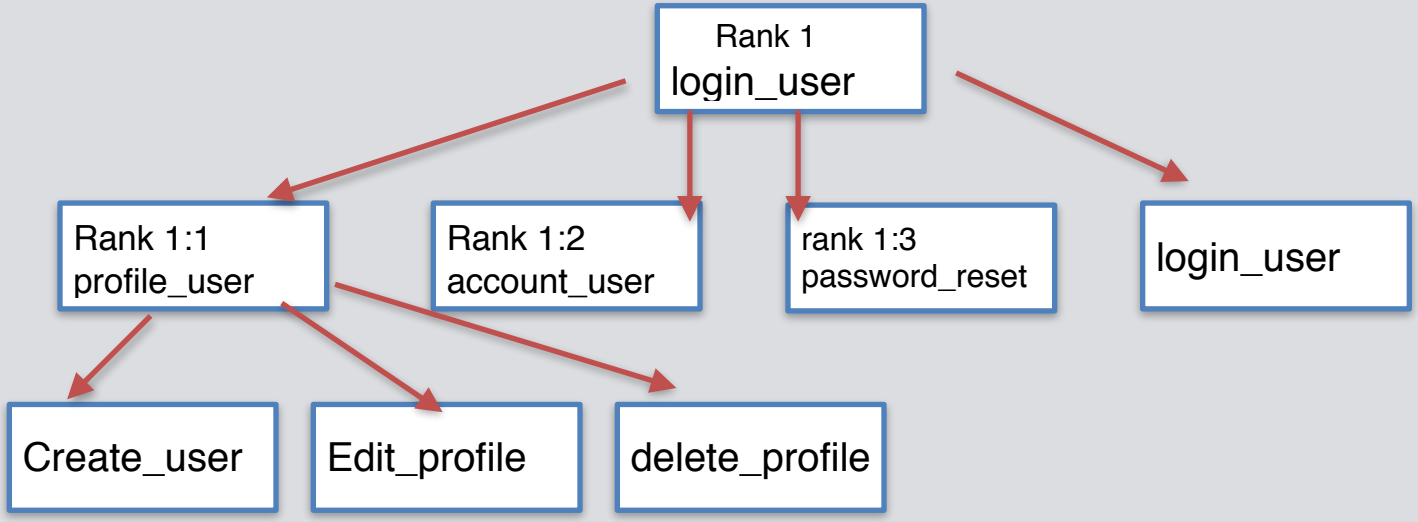
Now system can Identify labels

Label identification



Decision Tree

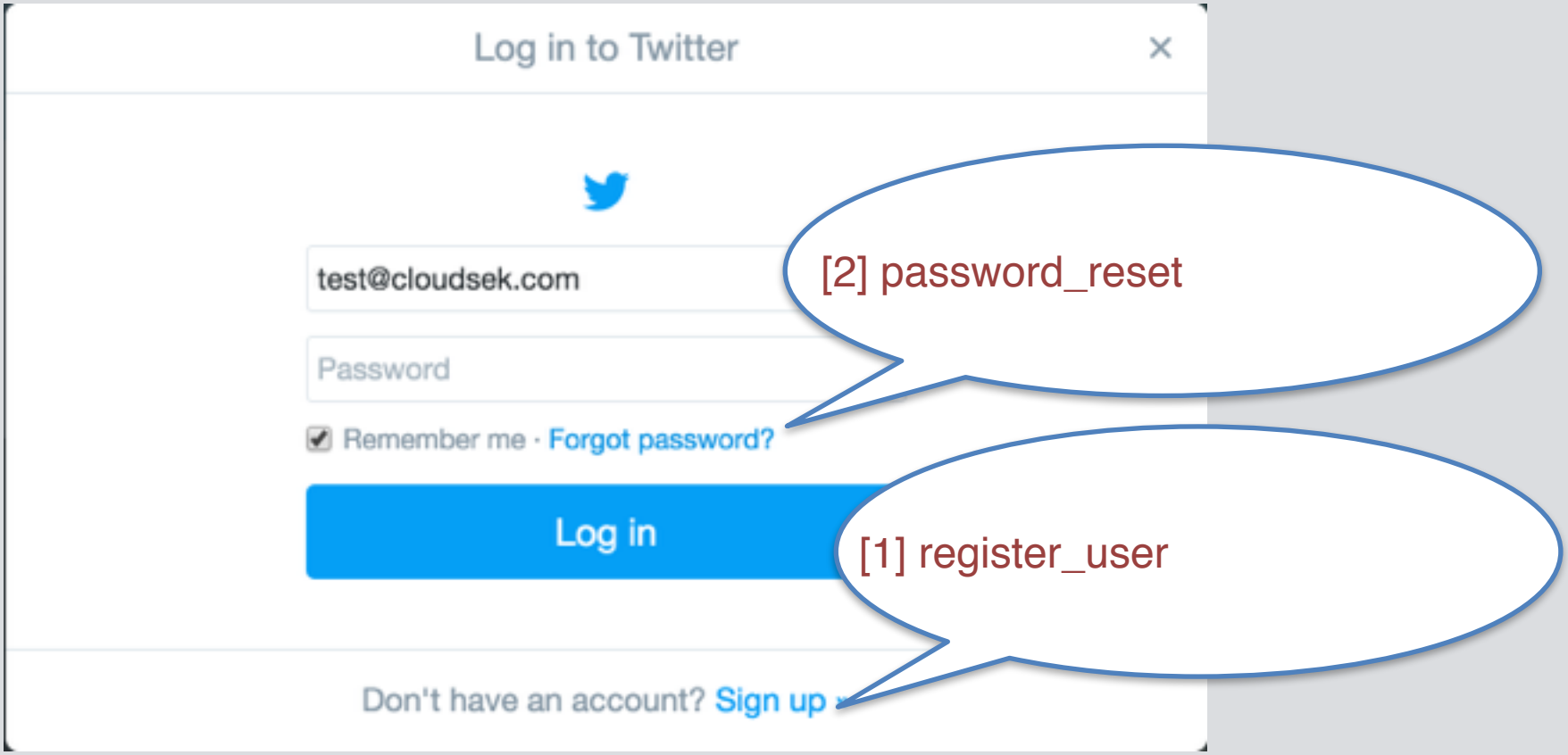
- Navigation algorithm based on a label ranking.
- We marked labels forward [1] and backward [0] .
- And ran our navigation algorithm based on this data.



Forward {Labels}	Next	Yes	Continue	Search	Repeat	Follow	Order	Save
Backward {labels}	Back	No	Abrot	Cancel	Exit	unfollow		

Building the navigation system.

- Now system can Identify labels.
- It can also identify the steps to reach an end result.

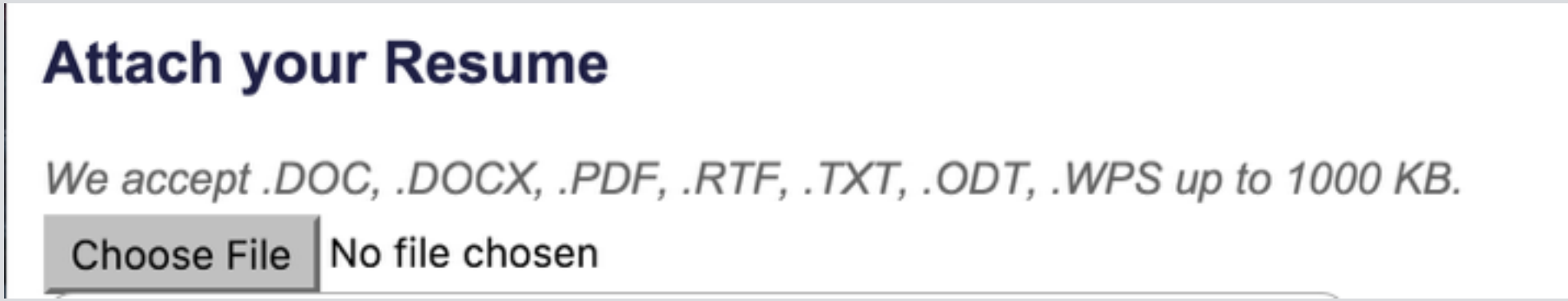


Demo

Use cases and results

1) Use cases and results

- File upload bugs.
- Can detect the allowed file types based on labels.



- Reported RCE bug to a large event booking company.

Bypass PHP GD Process To RCE

2) Use cases and results

- Insecure direct object on food delivery app.
- Repeat an order on behalf of another user.

attacker user-id modified to victim user-id , and you get free pizza .

Better results with API

Questions?

- Thanks to Nullcon /Null.
- Thanks to Garage4Hackers
- Special thanks to :
  - Finny Abraham and Rahul Babu of CloudSek team.
  - LavaKumar [IronWasp]

Whitfield, Bangalore
Phone: 09880669337
Web: https://www.CloudSek.com/
Email: theoracle@cloudsek.com

CloudSek an InfoSec Risk Assessment company

- With the rise in number of cloud applications, Enterprises are often affected with
  A. Vulnerable web applications that leaks sensitive data.
  B. Hacktivism, targeted attacks and other online leak of critical information.
- Security is an ongoing process and needs a more comprehensive strategy, that include tools that monitor internet resources 24/7 for potential security risks.

CloudSek Monitors,

- track clients' various internet resources 24/7 for potential security risks.
- provides actionable intelligence needed to tackle internet threats.
- CloudSek provides a unique AI based social-media/web threat monitor and a Cloud security monitor.