جامعة كارنيجي ميلون في قطر
Carnegie Mellon Qatar

مؤسسة قطر
Qatar Foundation

Summer 2010 Research Project

Spam Filtering by Text Classification

Manoj Reddy

Advisor: Dr. Behrang Mohit

NLP

- Field of Computer Science and Linguistics concerned with interactions between computers and languages

- Part of other areas of Computer Science such as Artificial Intelligence, Machine Learning & Computational Linguistics

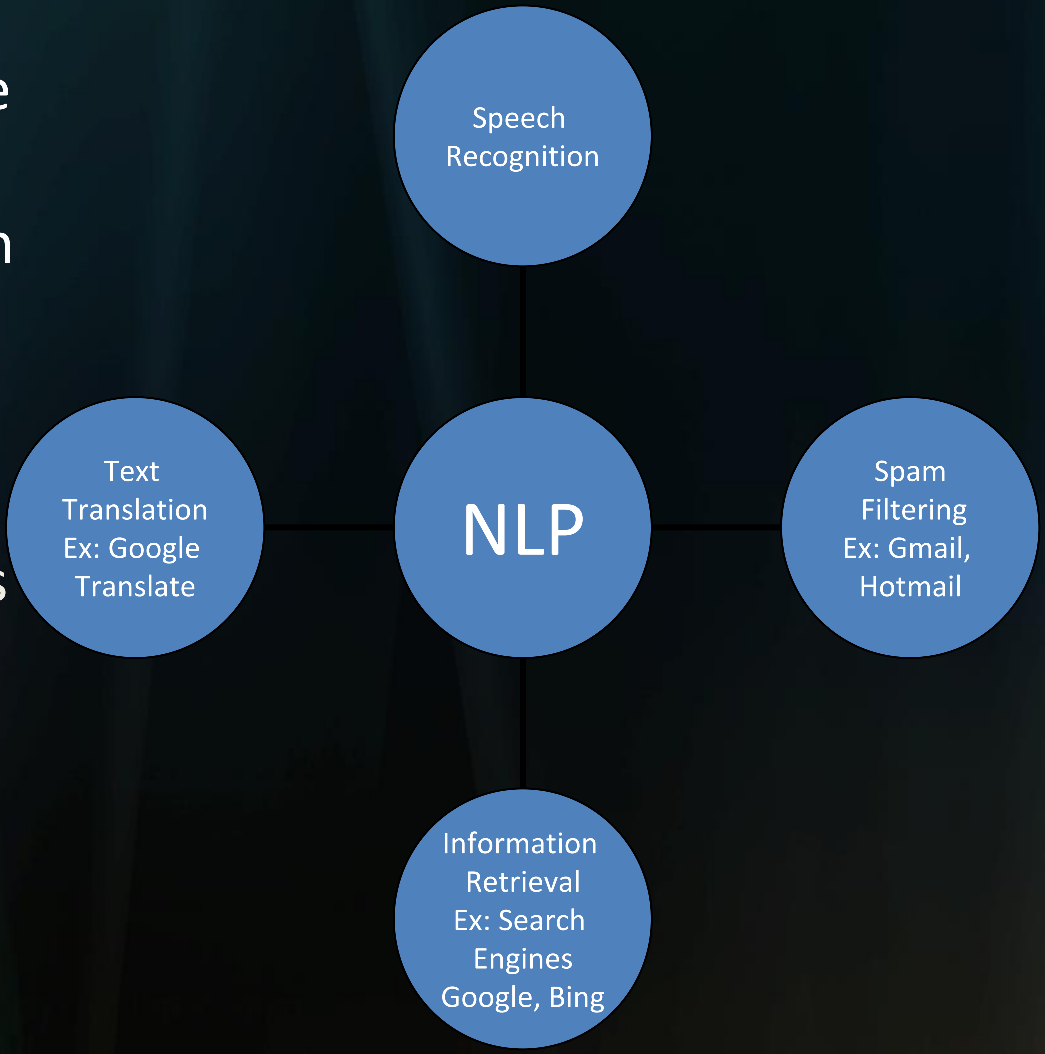
- Applications include -

NLP

- Field of Computer Science and Linguistics concerned with interactions between computers and languages

- Part of other areas of Computer Science such as Artificial Intelligence, Machine Learning & Computational Linguistics

Applications Include:



Speech Recognition
Text Translation Ex: Google Translate
NLP
Spam Filtering Ex: Gmail, Hotmail
Information Retrieval Ex: Search Engines Google, Bing

My Main Focus – Spam Filtering

- ## Spam

Display images below   MOTTO: FIGHTING POVERTY ROUND THE WORLD.
UK NATIONAL LOTTERY HEADQUARTERS:
Watford , Herts. WD18 9RN
UNITED KINGDOM (Customer Services)

(Customer Services)
FINAL NOTIFICATION

ATTENTION: WINNER

This is to inform you that you have been selected for a cash prize of (£450,000.00) held 18th August 2010 in London. Selection process was carried out through random selection in our computerized selection of e-mail from a database of more than 598,000,000 e-mails around the world from which you have been selected. UK Online Lottery International is approved by the British Gambling Board.

To begin processing your price, You must contact our fiduciary claims for more information as regards Procedures to claim your prize. However, it is not mandatory that you must claim your winning hence your choice to withdraw your claim. Congratulations once again on your wining.

REF NO: UK 200-26937
LOT NO: 2007MJL-01

Send the following details:

1.NAMES:
2.ADDRESS:
3.SEX:
4.AGE:
5.MARITAL STATUS:
6.OCCUPATION:
7.TELEPHONE NUMBER:
8.COUNTRY:

You may contact the Claims Agent:

Agent: John Rudolph
Tel: +447024030510
Email:claimsdirectorate2010@gmail.com

- ## Legitimate Email

Dear John ,

All the best to you for the New Year! How are things going in the Land of the Rising Sun? I must say, I really envy you getting that Tokyo gig with the company. Somehow they overlooked me on that one and I am forced to slug it out here through another frigid and snowy Montreal winter. Brrrr!

Hope everything is going great with you. I really cherish the time we had in London this August.

Remember the offer of winning a Porsche through the Lottery at the mall. We should have taken part in it because many people have won big prizes.

Did you hear about Margie Bronson suddenly leaving the company just before year-end? It was a bit of a shock to say the least. She gave one week's notice and was gone. Nobody knows for sure what's up with her but rumors have been flying fast and furious that she went through a bit of a personal meltdown and has now gone underground to lick her wounds for a while. There could be some truth to that since her long time relationship ended recently and three months ago she was passed over for that director position that was up for grabs. I'll keep you posted when we find out more.

As for me, I am quite busy these days on the Branscombe Systems Project. We are entering Phase Two now, and that is expected to run for three years, at least. Frank Schindler is Senior Project Manager and I am Team Leader of the Embedded Systems Group. I am enjoying it so far. Whether I'll feel the same way in three years, I'm not sure. By then I might be ready to join you in Japan.

I'm still kicking butt in the squash court and am managing to get in two or three matches per week. What about squash in Japan? Have you been able to play any over there? Are there even any squash courts? I suppose since you are in Tokyo there must be some. Let me know.

Nice talking to you.

Sincerely,
- David

My Main Focus – Spam Filtering

Spam

- Spam

Display images below MOTTO: FIGHTING POVERTY ROUND THE WO
UK NATIONAL LOTTERY HEADQUARTERS:
Watford , Herts. WD18 9RN
UNITED KINGDOM (Customer Services)

(Customer Services)
FINAL NOTIFICATION

ATTENTION: WINNER

This is to inform you that you have been selected for a c
(£450,000.00) held 18th August 2010 in London. Sele
carried out through random selection in our compute
from a database of more than 598,000,000 e-mails
have been selected. UK Online Lottery Int
British Gambling Board.

To begin processing your price, You must co
information as regards Procedures t
mandatory that you must claim yo
your claim. Congratulations once

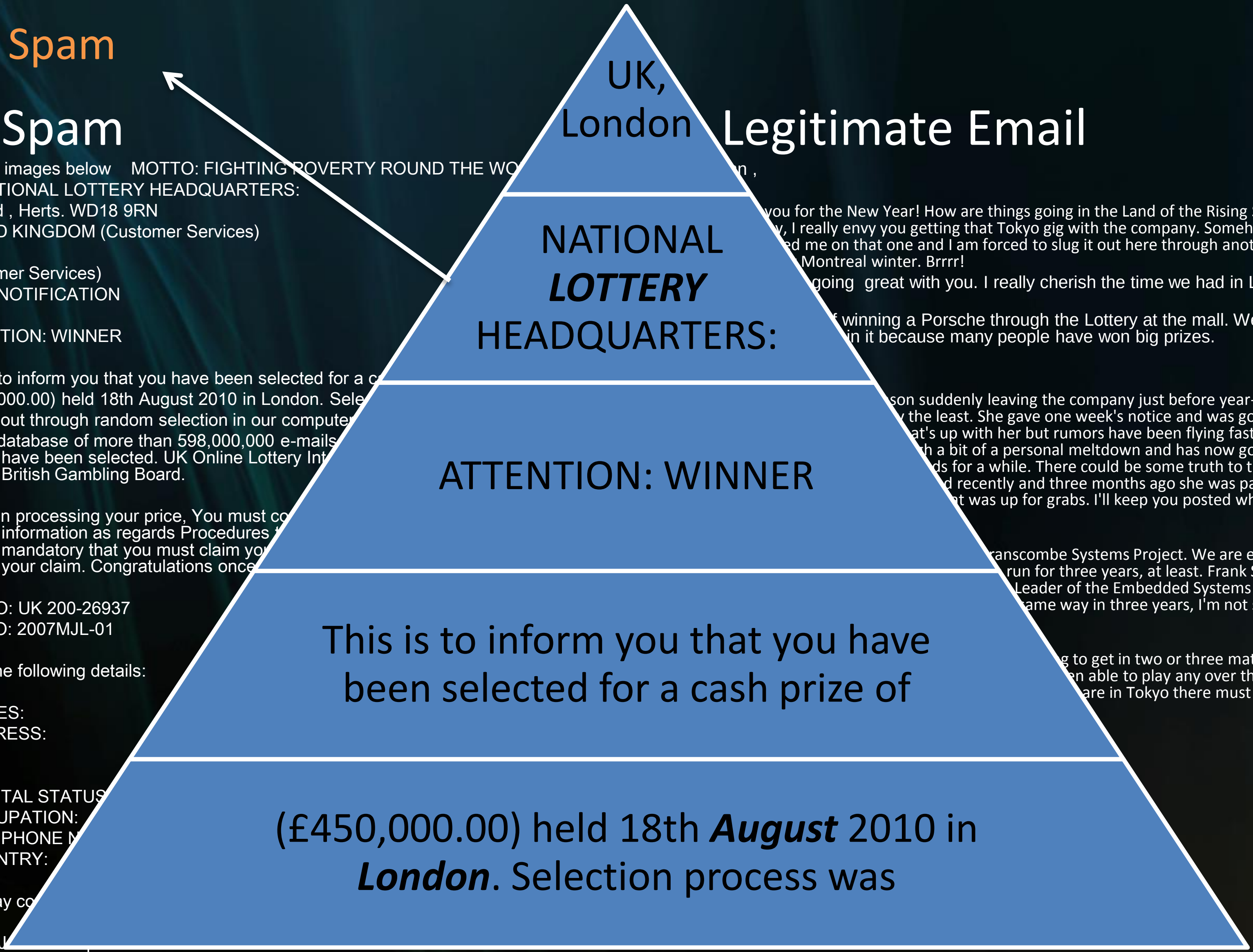
REF NO: UK 200-26937
LOT NO: 2007MJL-01

Send the following details:

1.NAMES:
2.ADDRESS:
3.SEX:
4.AGE:
5.MARITAL STATUS
6.OCCUPATION:
7.TELEPHONE N
8.COUNTRY:

You may co

Agent: J
Tel: +447024030510



UK, London

NATIONAL ***LOTTERY*** HEADQUARTERS:

ATTENTION: WINNER

This is to inform you that you have been selected for a cash prize of

(£450,000.00) held 18th ***August*** 2010 in ***London***. Selection process was

Legitimate Email

n ,

you for the New Year! How are things going in the Land of the Rising Sun? I
y, I really envy you getting that Tokyo gig with the company. Somehow they
ed me on that one and I am forced to slug it out here through another frigid
Montreal winter. Brrrr!
going great with you. I really cherish the time we had in London

f winning a Porsche through the Lottery at the mall. We should
in it because many people have won big prizes.

son suddenly leaving the company just before year-end? It
the least. She gave one week's notice and was gone.
at's up with her but rumors have been flying fast and
h a bit of a personal meltdown and has now gone
ds for a while. There could be some truth to that since
d recently and three months ago she was passed
t was up for grabs. I'll keep you posted when we

ranscombe Systems Project. We are entering
run for three years, at least. Frank Schindler
Leader of the Embedded Systems Group. I
ame way in three years, I'm not sure. By

g to get in two or three matches per
n able to play any over there? Are
are in Tokyo there must be some.

Legitimate
Email

My Main Focus – Spam Filtering

●

Display
UK NAT
Watford
UNITED

(Customer Services)
FINAL NOTIFICATION

ATTENTION: WINNER

This is t
(£450,0
carried
from a c

To begi

REF NO: UK 200-26937
LOT NO: 2007MJL-01

Send the following details:

1.NAME
2.ADDR
3.SEX:
4.AGE:
5.MARI
6.OCCU
7.TELE
8.COUN

You ma

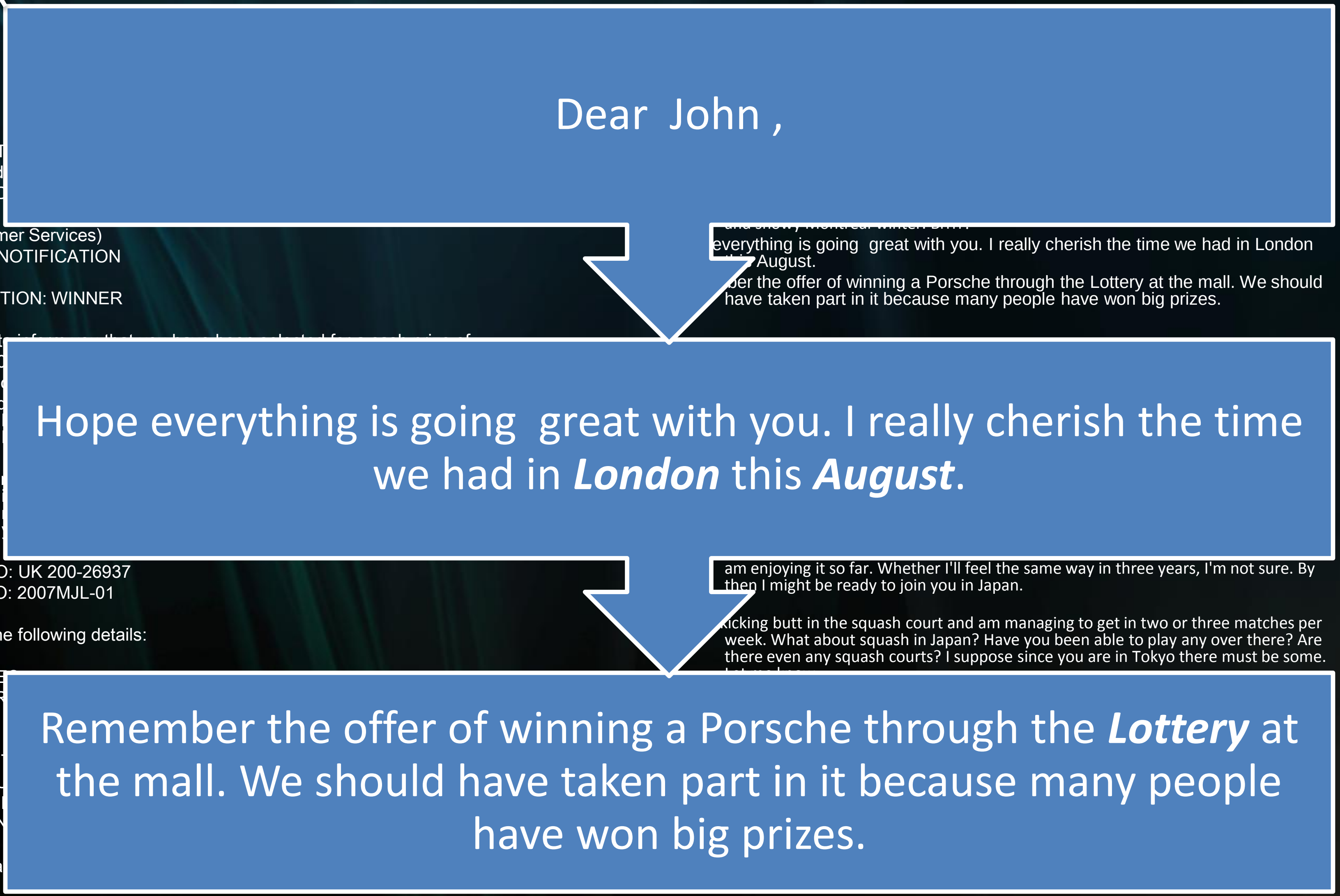
Agent: John Rudolph
Tel: +447024030510

everything is going great with you. I really cherish the time we had in London August.

er the offer of winning a Porsche through the Lottery at the mall. We should have taken part in it because many people have won big prizes.

am enjoying it so far. Whether I'll feel the same way in three years, I'm not sure. By then I might be ready to join you in Japan.

icking butt in the squash court and am managing to get in two or three matches per week. What about squash in Japan? Have you been able to play any over there? Are there even any squash courts? I suppose since you are in Tokyo there must be some.



Dear John ,

Hope everything is going great with you. I really cherish the time we had in ***London*** this ***August***.

Remember the offer of winning a Porsche through the ***Lottery*** at the mall. We should have taken part in it because many people have won big prizes.

My Main Focus – Spam Filtering

- Spam

Display images below MO ING POVERTY ROUND THE WORLD.
UK NATIONAL LOTTERY H ERS:
Watford , Herts. WD18 9R
UNITED KINGDOM (Cus

(Customer Services)
FINAL NOTIFICATIO

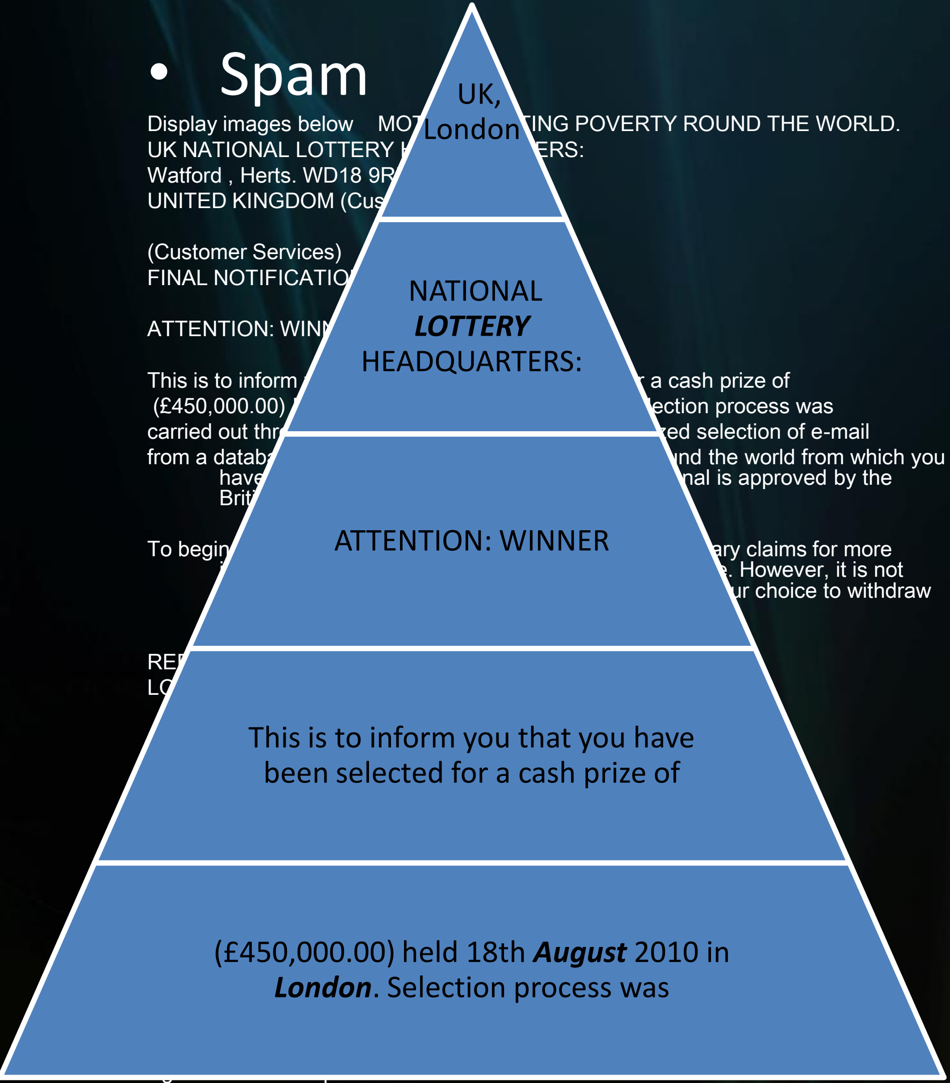
ATTENTION: WIN

This is to inform a cash prize of
(£450,000.00) ection process was
carried out thr zed selection of e-mail
from a datab und the world from which you
have nal is approved by the
Brit

To begin ary claims for more
. However, it is not
ur choice to withdraw

RE
LO

Tel: +447024030510


UK, London

NATIONAL *LOTTERY* HEADQUARTERS:

ATTENTION: WINNER

This is to inform you that you have been selected for a cash prize of

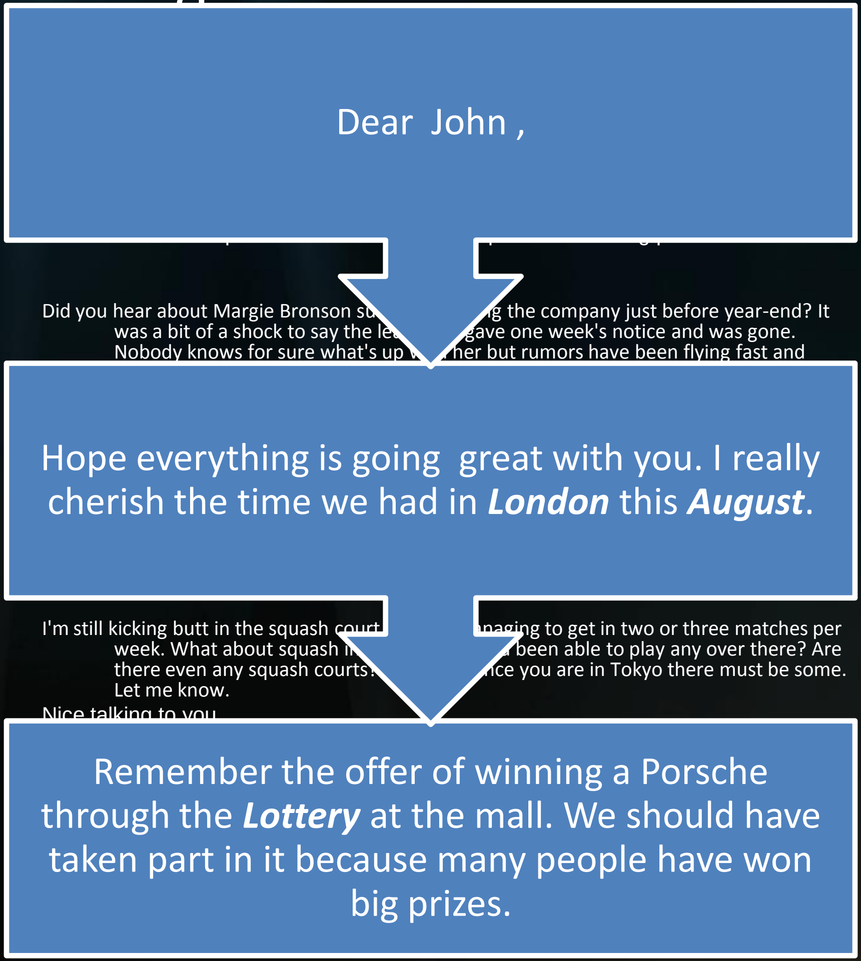
(£450,000.00) held 18th ***August*** 2010 in ***London***. Selection process was

- Legitimate Email

Did you hear about Margie Bronson su g the company just before year-end? It was a bit of a shock to say the le gave one week's notice and was gone. Nobody knows for sure what's up w her but rumors have been flying fast and

I'm still kicking butt in the squash court naging to get in two or three matches per week. What about squash i been able to play any over there? Are there even any squash courts ice you are in Tokyo there must be some. Let me know.

Nice talking to you.


Dear John ,

Hope everything is going great with you. I really cherish the time we had in ***London*** this ***August***.

Remember the offer of winning a Porsche through the ***Lottery*** at the mall. We should have taken part in it because many people have won big prizes.

Why is this important?

" Like almost everyone who uses e-mail, I receive a ton of spam every day. Much of it offers to help me get out of debt or get rich quick. It would be funny if it weren't so irritating. "

- **Bill Gates, 2003**

Why is this important?

- Businesses all around the world rely on email communication
- Along with this bliss there is a growing increase in unsolicited mail
- Advertising companies use it since it is very very cheap (0.01 c per email)
- Businesses spend millions of dollars on technologies that solve this problem

Why is this important?

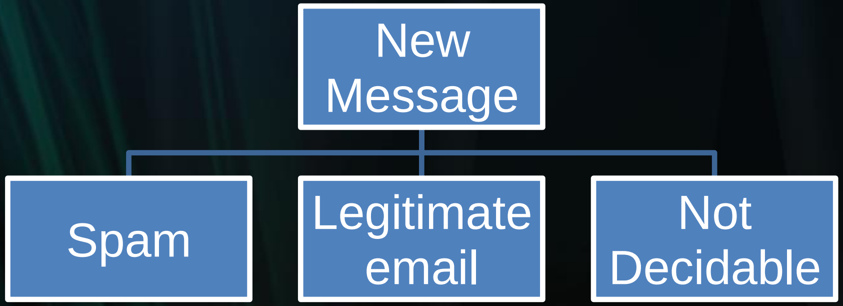
- Businesses all around the world rely on email communication
- Along with this bliss there is a growing increase in unsolicited mail
- Advertising companies use it since it is very very cheap (0.01 c per email)
- Businesses spend millions of dollars on technologies that solve this problem

Main Objective

- Creating a filter that looks at a mail and decides whether it is Spam or Legitimate or Undecidable

Main Objective

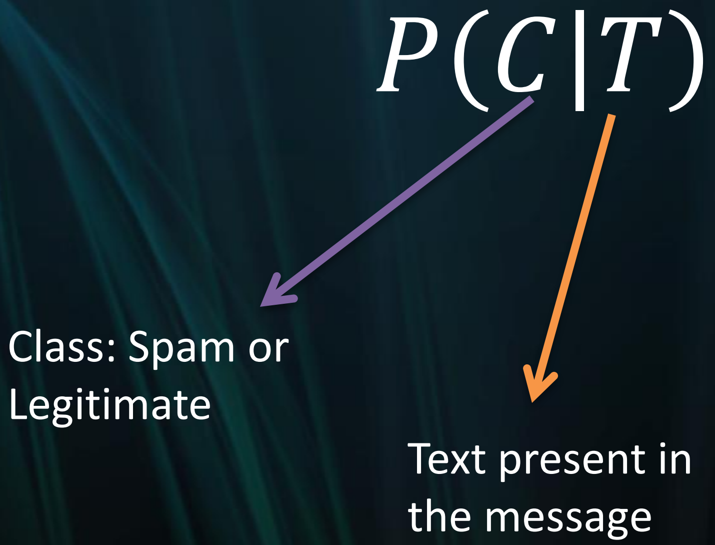
- Creating a filter that looks at a mail and decides whether it is Spam or legitimate or Undecidable



Solution: Statistical Classification

- Build a Classifier software
- **Idea:** Use previously seen data (emails) to train a probabilistic *classifier* for spam.
    - **Training:** Learn what are the important features of spam emails.
        - Use probabilities
    - **Test:** Use what you learned from the data to classify a new email.

Class Probability



$$P(C|T)$$

Class: Spam or Legitimate

Text present in the message

Naïve Bayes Theorem

- Basically, derived from the theorem in Probability called Bayes Theorem.

$$P(C|T) = \frac{P(T|C)P(C)}{P(T)}$$

Calculating the probability a message being spam or legitimate.

Naïve Bayes Classifier

- C – Class { Legitimate or Spam}
- T –Text { Text in the Message}

$$P(C|t_1, t_2, t_3 \ldots t_n = \frac{P(C)P(t_1, t_2 \ldots . t_n|C)}{P(t_1, t_2, t_3, \ldots . t_n)}$$

- Further breakdown of the formula

$$= P(C)P(t_1|C)P(t_2|C)P(t_3|C) \ldots . P(t_n|C)$$

Naïve Bayes Classifier

- C – Class { Legitimate or Spam}
- T –Text {Text in the Message}

$$P(C|t_1, t_2, t_3 \ldots t_n = \frac{P(C)P(t_1, t_2 \ldots . t_n|C)}{P(t_1, t_2, t_3, \ldots . t_n)}$$

Further breakdown of the formula

$$= P(C)P(t_1|C)P(t_2|C)P(t_3|C) \ldots . P(t_n|C)$$

Implementation

- Training
  - Data Set: TREC-05, contains Enron emails
  - Involves creating a table of probabilities of each word with respect to legitimate/Spam
  - Indirectly requires calculating the frequency distribution of each word in the data set.
  - Code was written entirely in Python, since it was best suited for NLP

Implementation

- Training
  - Data Set: TREC-05, contains Enron emails
  - Involves creating a table of probabilities of each word with respect to legitimate/Spam
  - Indirectly requires calculating the frequency distribution of each word in the data set.
  - Code was written entirely in Python, since it was best suited for NLP

Implementation

- Training
  - Data Set: TREC-05, contains Enron emails
  - Involves creating a table of probabilities of each word with respect to legitimate/Spam
  - Indirectly requires calculating the frequency distribution of each word in the data set.
  - Code was written entirely in Python, since it was best suited for NLP

Implementation

- Training
  - Data Set: TREC-05, contains Enron emails
  - Involves creating a table of probabilities of each word with respect to legitimate/Spam
  - Indirectly requires calculating the frequency distribution of each word in the data set.
  - Code was written entirely in Python, since it was best suited for NLP

Implementation

- Testing
  - Involves reading 50 **UNSEEN** messages and determine whether they are legitimate/Spam, so the baseline is 50
  - Assigns a probability for each message being spam/legitimate and then the greatest of the two is considered.
  - The results of the test are observed and an evaluation is made based on the results.

Implementation

- Testing
  - Involves reading 50 **UNSEEN** messages and determine whether it is legitimate/Spam, so the baseline is 50
  - Assigns a probability for each message being spam/legitimate and then the greatest of the two is considered.
  - The results of the test are observed and an evaluation is made based on the results.

Implementation

- Testing
  - Involves reading 50 **UNSEEN** messages and determine whether it is legitimate/Spam, so the baseline is 50
  - Assigns a probability for each message being spam/legitimate and then the greatest of the two is considered.
  - The results of the test are observed and an evaluation is made based on the results.

Features

- Naïve Bayes is just an algorithm & there are many modules that can be attached to the system to reach our end goal.
- We also introduced many novel features:
  - Observation:
    - Spam has a lot of misspelled words
  - Solution:
    - Maintain an English Dictionary to highlight whenever there is a misspelled word
    - Increase the probability of a message being a spam if majority of the words are non-ASCII or garbage

Features

- Naïve Bayes is just an algorithm & there are many modules that can be attached to the system to reach our end goal.
- We also introduced many novel features:
  - Observation:
    - Spam has a lot of misspelled words
  - Solution:
    - Maintain an English Dictionary to highlight whenever there is a misspelled word
    - Increase the probability of a message being a spam if majority of the words are non-ASCII or garbage

Other notable features

- Stop words
  - Observation:
    - Words such as “the”, “is” are trivial
  - Solution:
    - Removing them help us focus on the essential text

- Addresses
  - Insight:
    - Keeping track of the “To:” and “From:” fields help determine what type of email is each person sending.

Other notable features

- Stop words
  - Observation:
    - Words such as “the”, “is” are trivial
  - Solution:
    - Removing them help us focus on the essential text
- Addresses
  - Insight:
    - Keeping track of the “To:” and “From:” fields help determine what type of email is each person sending.

Evaluation

- Accuracy
  - We need to know how well is the classifier working
  - 4 possibilities

Legitimate email classified as a **L**egitimate	**L**egitimate classified as a **S**pam
Spam classified as a **L**egitimate	**S**pam email classified as a **S**pam

Accuracy

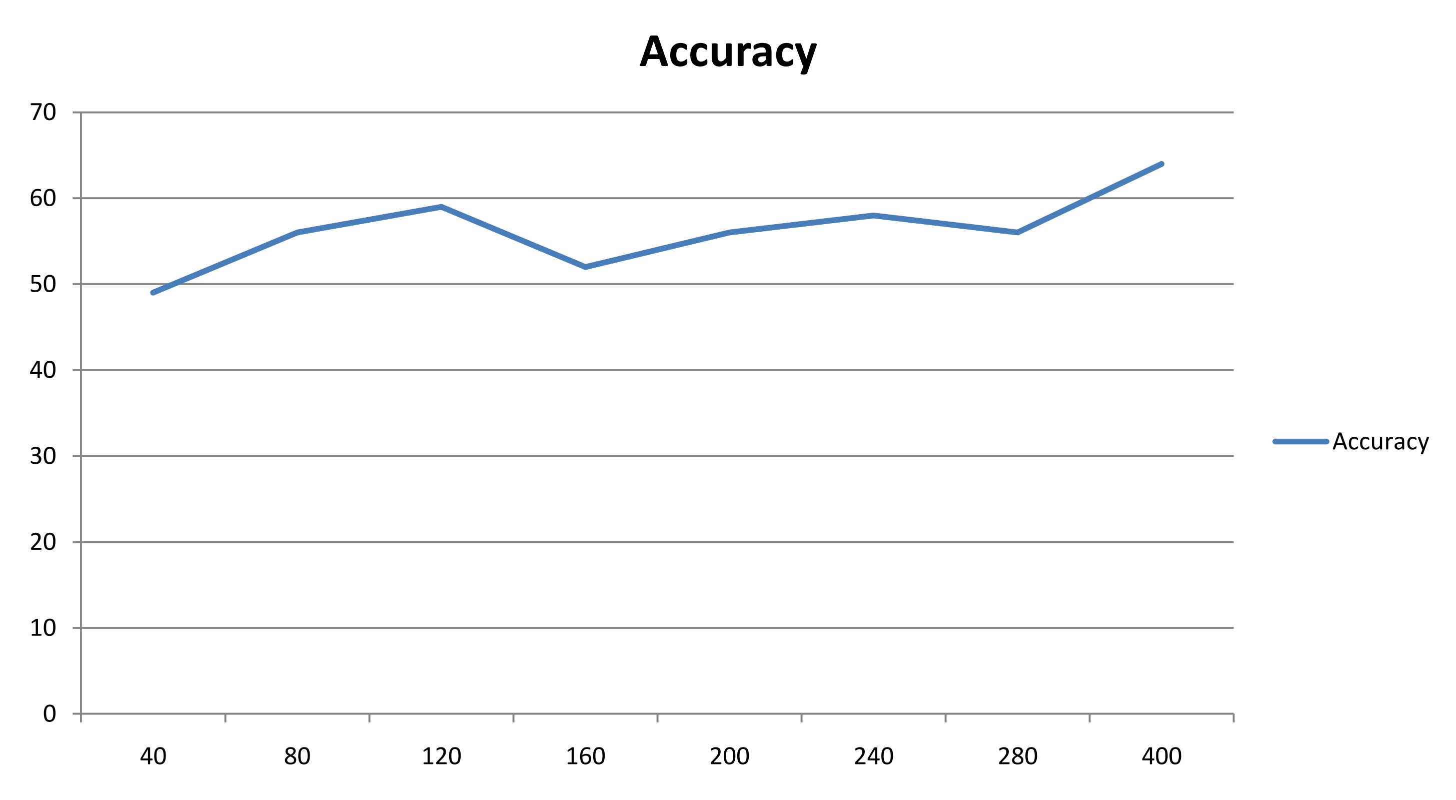
- So the formula will be

$$\text{Accuracy} = \frac{LL + SS}{SS + LL + SL + LS}$$

Legitimate email classified as a **L**egitimate	**L**egitimate classified as a **S**pam
Spam classified as a **L**egitimate	**S**pam email classified as a **S**pam

Evaluation

- Upon trying different sizes of data sets there is a change in accuracy:



Future Work

- This is just a Naïve Bayes classifier
  - It considers the probability of each word independently of the other.
  - This might not be always true
  - Consider the words:
    - Ice cream

After the word ICE there is a high probability that the next word is CREAM.

Future Work

- This is just a Naïve Bayes classifier
  - It considers the probability of each word independently of the other.
  - This might not be always true
  - Consider the words:
    - Ice cream

After the word ICE there is a high probability that the next word is CREAM.

So an efficient classifier should take into account the relation between text

Future Work

- This is just a Naïve Bayes classifier
  - It considers the probability of each word independently of the other.
  - This might not be always true
  - Consider the words:
    - Ice cream

After the word ICE there is a high probability that the next word is CREAM.

So an efficient classifier should take into account the relation between text

Future Work

- IP Address Management
  - Track of IP Addresses and locate the sender of the email
- Image Filtering
  - Involves categorizing images as Spam/legitimate
  - This takes it to a whole new level
- Multi – class Text Classification
  - Imagine documents automatically being sorted out into different label simply by looking at the text in it.
  - Cool, isn't it ?
- References & Books : Report.

Future Work

- IP Address Management
  - Track of IP Addresses and locate the sender of the email
- Image Filtering
  - Involves categorizing images as Spam/legitimate
  - This takes it to a whole new level
- Multi – class Text Classification
  - Imagine documents automatically being sorted out into different label simply by looking at the text in it.
  - Cool, isn't it ?
- References & Books : Report.

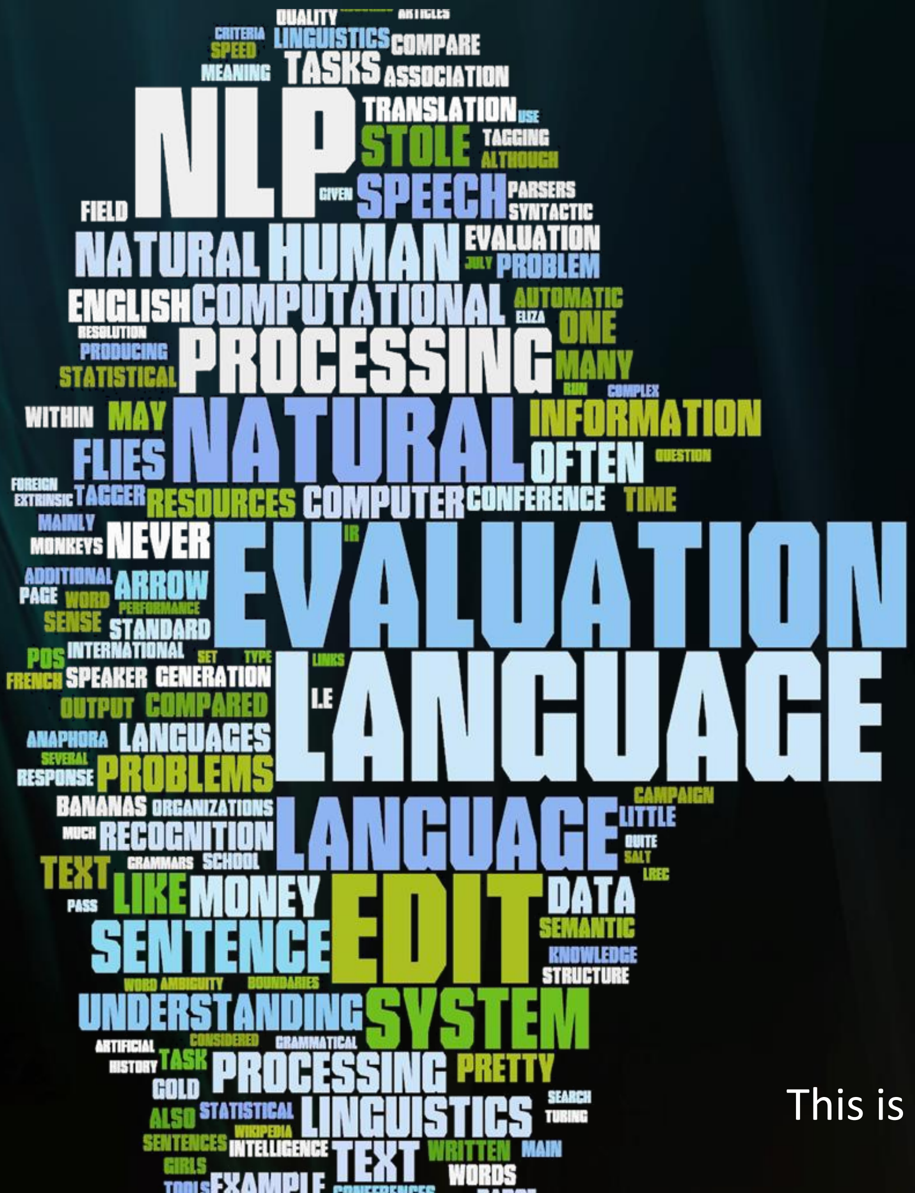
Future Work

- IP Address Management
  - Track of IP Addresses and locate the sender of the email
- Image Filtering
  - Involves categorizing images as Spam/Legitimate
  - This takes it to a whole new level
- Multi – class Text Classification
  - Imagine documents automatically being sorted out into different label simply by looking at the text in it.
  - Cool, isn’t it ?

Thank You !!!

Questions / Comments

NLP (Natural Language Processing)



This is from wordle.net