SD-TSIA214 Machine Learning For Text Mining

June 11, 2018

Author: ZHANG Bolong, ZHU Fangda

```
In [186]: # import numpy as np
          import difflib
```

1 Task : automatic segmentation of mails, problem statement

This Lab aims to build an email segmentation tool, dedicated to separate the email header from its body. It is proposed to perform this task by learning a HMM (A; B;) with two states, one (state 1) for the header, the other (state 2) for the body. In this model, it is assumed that each mail actually contains a header : the decoding necessarily begins in the state 1.

1.1 Q1 : Give the value of the π vector of the initial probabilities

According to the Task, the decoding necessarily begins in the states1, sp our HMM necessarily has an initial state of 1.So the π vector of the initial probabilities should be:

$$\pi = (1, 0)^T$$

□.

Knowing that each mail contains exactly one header and one body, each mail follows once the transition from 1 to 2. The transition matrix $(A(i; j) = P(j|i))$ estimated on a labeled small corpus has thus the following form :

$$A = \begin{pmatrix} 0.9992180780358 12 & 0.0007819219641879 74 \\ 0 & 1 \end{pmatrix}$$

1.2 Q2. What is the probability to move from state 1 to state 2 ? What is the probability to remain in state 2 ? What is the lower/higher probability ? Try to explain wh

We can find the probabilities of movements between states in the transition matrx A given above. The row index determines the starting state, while the column index determines the arriving state. Thus: - Probability from state 1 to state 2:

$$A(1, 2) = 0.0007819219641879 74$$

- Probability to remain in state 2:

$$A(2, 2) = 1.0$$

For now, the probability to remain state 2 is the highest, this value is relative intuitive: once our email in the state of body, we can not get back to the header state when we continue to read following charaters.□

A mail is represented by a sequence of characters. Let N be the number of different characters. Each part of the mail is characterized by a discrete probability distribution on the characters $P(c|s)$, with s = 1 or s = 2.

1.3 To implement

```
In [187]: def viterbi(obs, states, start_prob, trans, emission_prob, log=False):
              """
                  Viterbi Algorithm Implementation

                  Keyword arguments:
                      - obs: sequence of observation
                      - states:list of states
                      - start_prob:vector of the initial probabilities
                      - trans: transition matrix
                      - emission_prob: emission probability matrix
                  Returns:
                      - seq: sequence of state
              """
              start_prob = np.log(start_prob)
              trans = np.log(trans)
              emission_prob = np.log(emission_prob)
              T1 = np.zeros((obs.size, states.size))
              T2 = np.zeros((obs.size, states.size), dtype = np.uint8)
              for state in states:
                  T1[0,state] = start_prob[state] +  emission_prob[obs[0]][state]
                  T2[0,state] = 0

              for index in range(1,obs.size):
                  for state in states:
                      liste = [T1[index-1,start_state] + trans[start_state, state] for start_st
                      T1[index, state] = np.max(liste) + emission_prob[obs[index]][state]
                      T2[index, state] = np.argmax(liste)
                      if(log == False): continue

              path = np.zeros(len(obs), dtype= np.uint8)

              path[-1] = np.argmax(T1[-1])
              for i in range(len(obs)-2,-1,-1):
                  path[i] = T2[i, path[i+1]]
              return path
```

```
In [188]: # Test
          A = np.array([[0.7,0.3], [0.4, 0.6]])
          start_prob = np.array([0.6, 0.4])
```

```
emis = np.array([[0.1,0.6], [0.4,0.3], [0.5,0.1]])
obs = np.array([0,1,2])
states = np.array([0,1])

x = viterbi(obs, states, start_prob, A, emis)
print(x)
```

```
[0 1 0]
```

```
In [189]: # Test
A = np.array([[0.5,0.2, 0.3], [0.3, 0.5, 0.2], [0.2,0.3,0.5]])
start_prob = np.array([0.2, 0.4, 0.4])
emis = np.array([[0.5,0.5], [0.4,0.6], [0.7,0.3]])
obs = np.array([0,1,0])
states = np.array([0,1,2])

x = viterbi(obs, states, start_prob, A, emis.T)
print(x)
```

```
[0 2 2]
```

1.4 Question 4 Print the track and present and discuss the results obtained on mail11.txt to mail30.txt

```
In [191]: # Get emission matrix
emission = np.loadtxt('P.text')

# Get Transition matrix
trans = np.array([[0.999218078035812,0.000781921964187974],[1e-100,1-1e-100]])
# Set start probalility
start_prob = np.array([1-1e-100,1e-100])

# Import data file
with open('dat/mail.lst', 'r') as file:
    file_list = file.read().splitlines()
datas = [np.loadtxt('dat/'+ x, dtype = int) for x in file_list]
states = np.array([0,1])
```

```
In [236]: # Text for mail1 - mail10
def spliceText(path, text):
    '''splice the text according to the states path'''
    vals, index = np.unique(path, return_index=True)
    index = index[1]
    return text[0:index], text[index:]

def verify_res(number):
        path = viterbi(np.loadtxt('dat/mail'+ str(number) + '.dat', dtype = int), sta
```

```
                val, index = np.unique(path, return_index=True)
                index = index[1]
                print('+---------------------------------------- mail %d ----------------
                print('Test result:')
                print("state 1: 0 ~ " + str(index-1))
                print('state 2:' + str(index) + ' ~ ' + str(len(path)))
                with open('dat/mail' + str(number) + 'h.txt') as file:
                    text_h = file.read()
                    nb_h = len(text_h)
                with open('dat/mail' + str(number) + 'c.txt') as file:
                    text_c = file.read()
                    nb_c = len(text_c)
                with open('dat/mail'+ str(number) + '.txt') as file:
                    text = file.read()
                    header,body = spliceText(path,text)
                print('Real result:')
                print("state 1: 0 ~ " + str(nb_h))
                print('state 2:' + str(nb_h+1) + ' ~ ' + str(nb_h + nb_c))

                print("<------------------------------ Diff ---------------------------------
                diff = difflib.context_diff(text_c.splitlines(), body.splitlines())
                print('\n'.join(list(diff)))


        # def print_data():
        #     for i,data in enumerate(['mail' + str(i) + '.dat' for i in range(1,11)]):
```

```
In [237]: for i in range(1,11):
              verify_res(i)
```

```
+---------------------------------------- mail 1 -----------------------------------+
Test result:
state 1: 0 ~ 3796
state 2:3797 ~ 5216
Real result:
state 1: 0 ~ 3611
state 2:3612 ~ 5216
<------------------------------ Diff ------------------------------------------------------>
***

---

***************

*** 1,7 ****
```

```
-
-     Date:        Wed, 21 Aug 2002 10:54:46 -0500
-     From:        Chris Garrigues <cwg-dated-1030377287.06fa6d@DeepEddy.Com>
-     Message-ID:  <1029945287.4797.TMDA@deepeddy.vircio.com>


    | I can't reproduce this error.
--- 1,3 ----

+---------------------------------------- mail 2 -----------------------------------+
Test result:
state 1: 0 ~ 2445
state 2:2446 ~ 3376
Real result:
state 1: 0 ~ 2476
state 2:2477 ~ 3376
<----------------------------- Diff ------------------------------------------------>
***

---

***************

*** 1,3 ****

--- 1,4 ----

+ ntent-Transfer-Encoding: 7bit

  Martin A posted:
  Tassos Papadopoulos, the Greek sculptor behind the plan, judged that the
+---------------------------------------- mail 3 -----------------------------------+
Test result:
state 1: 0 ~ 2263
state 2:2264 ~ 3934
Real result:
state 1: 0 ~ 2182
state 2:2183 ~ 3934
<----------------------------- Diff ------------------------------------------------>
***

---

***************

*** 1,8 ****

!
```

```
! Man Threatens Explosion In Moscow
!
! Thursday August 22, 2002 1:40 PM
! MOSCOW (AP) - Security officers on Thursday seized an unidentified man who
  said he was armed with explosives and threatened to blow up his truck in
  front of Russia's Federal Security Services headquarters in Moscow, NTV
  television reported.
--- 1,4 ----

! - Security officers on Thursday seized an unidentified man who
  said he was armed with explosives and threatened to blow up his truck in
  front of Russia's Federal Security Services headquarters in Moscow, NTV
  television reported.
+---------------------------------------- mail 4 -------------------------------------+
Test result:
state 1: 0 ~ 2304
state 2:2305 ~ 3424
Real result:
state 1: 0 ~ 2296
state 2:2297 ~ 3424
<------------------------------ Diff ------------------------------------------------->
***

---

***************

*** 1,5 ****

!
! Klez: The Virus That Won't Die

  Already the most prolific virus ever, Klez continues to wreak havoc.

--- 1,4 ----

! e Virus That Won't Die

  Already the most prolific virus ever, Klez continues to wreak havoc.

+---------------------------------------- mail 5 -------------------------------------+
Test result:
state 1: 0 ~ 2302
state 2:2303 ~ 3386
Real result:
state 1: 0 ~ 2333
state 2:2334 ~ 2989
<------------------------------ Diff ------------------------------------------------->
```

```
***

---

***************

*** 1,3 ****

--- 1,4 ----

+ ntent-Transfer-Encoding: 7bit

  >  in adding cream to spaghetti carbonara, which has the same effect on pasta as
  >  making a pizza a deep-pie;
***************

*** 14,16 ****

--- 15,33 ----

  Stewart Smith
  Scottish Microelectronics Centre, University of Edinburgh.
  http://www.ee.ed.ac.uk/~sxs/
+
+
+ ------------------------ Yahoo! Groups Sponsor ----------------------~-->
+ 4 DVDs Free +s&p Join Now
+ http://us.click.yahoo.com/pt6YBB/NXiEAA/mG3HAA/7gSolB/TM
+ ---------------------------------------------------------------------~->
+
+ To unsubscribe from this group, send an email to:
+ forteana-unsubscribe@egroups.com
+
+
+
+ Your use of Yahoo! Groups is subject to http://docs.yahoo.com/info/terms/
+
+
+
+---------------------------------------- mail 6 -----------------------------------+
Test result:
state 1: 0 ~ 2436
state 2:2437 ~ 3211
Real result:
state 1: 0 ~ 2467
state 2:2468 ~ 3211
<----------------------------- Diff ------------------------------------------------->
***
```

```
---

***************

*** 1,3 ****

--- 1,4 ----

+ ntent-Transfer-Encoding: 7bit


  > I just had to jump in here as Carbonara is one of my favourites to make and
+---------------------------------------- mail 7 -----------------------------------+
Test result:
state 1: 0 ~ 2555
state 2:2556 ~ 3848
Real result:
state 1: 0 ~ 2521
state 2:2522 ~ 3848
<------------------------------ Diff ------------------------------------------------------>
***

---

***************

*** 1,7 ****

!
! The Scotsman - 22 August 2002
!
!  Playboy wants to go out with a bang


   AN AGEING Berlin playboy has come up with an unusual offer to lure women into
--- 1,4 ----

! ayboy wants to go out with a bang


   AN AGEING Berlin playboy has come up with an unusual offer to lure women into
+---------------------------------------- mail 8 -----------------------------------+
Test result:
state 1: 0 ~ 2305
state 2:2306 ~ 3558
Real result:
state 1: 0 ~ 2336
```

```
state 2:2337 ~ 3558
<------------------------------ Diff ------------------------------------------------------>
***

---

***************

*** 1,3 ****

--- 1,4 ----

+ ntent-Transfer-Encoding: 7bit

  Martin Adamson wrote:
  >
+---------------------------------------- mail 9 ----------------------------------+
Test result:
state 1: 0 ~ 2540
state 2:2541 ~ 8625
Real result:
state 1: 0 ~ 2509
state 2:2510 ~ 8625
<------------------------------ Diff ------------------------------------------------------>
***

---

***************

*** 1,7 ****


- The Scotsman
-
-  Thu 22 Aug 2002

   Meaningful sentences

--- 1,4 ----

+---------------------------------------- mail 10 ---------------------------------+
Test result:
state 1: 0 ~ 2846
state 2:2847 ~ 3715
Real result:
state 1: 0 ~ 2847
state 2:2848 ~ 3715
```

<---------------------------- Diff -->

The results for the 10 first mail seem consistent.

```
In [253]: # Text for mail1 - mail10
          def spliceText(path, text):
              '''splice the text according to the states path'''
              vals, index = np.unique(path, return_index=True)
              index = index[1]
              return text[0:index], text[index:]

          def verify_res_last(number):
                  path = viterbi(np.loadtxt('dat/mail'+ str(number) + '.dat', dtype = int), sta
                  val, index = np.unique(path, return_index=True)
                  index = index[1]
                  print('+------------------------------------ mail %d ---------------------
                  print('Test result:')
                  print("state 1: 0 ~ " + str(index-1))
                  print('state 2:' + str(index) + ' ~ ' + str(len(path)))
                  with open('dat/mail'+ str(number) + '.txt') as file:
                      text = file.read()
                      header,body = spliceText(path,text)
                  print('<------------------------------ header ------------------------------
                  print('\n'.join(header.splitlines()[0:6]))
                  print('..............')
                  print('\n'.join(header.splitlines()[-6:]))
                  print('<------------------------------ body --------------------------------
                  print('\n'.join(body.splitlines()[0:9]))
                  print('..............\n')
                  print('<------------------------------ End ---------------------------------
          for i in range(10,31):
              verify_res_last(i)
```

```
+------------------------------------ mail 10 ----------------------------+
Test result:
state 1: 0 ~ 2846
state 2:2847 ~ 3715
<------------------------------ header ------------------------------------------------->
From spamassassin-talk-admin@lists.sourceforge.net  Thu Aug 22 15:25:29 2002
Return-Path: <spamassassin-talk-admin@example.sourceforge.net>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id B48D543F99
        for <zzzz@localhost>; Thu, 22 Aug 2002 10:25:28 -0400 (EDT)
...
```

```
List-Id: Talk about SpamAssassin <spamassassin-talk.example.sourceforge.net>
List-Unsubscribe: <https://example.sourceforge.net/lists/listinfo/spamassassin-talk>,
    <mailto:spamassassin-talk-request@lists.sourceforge.net?subject=unsubscribe>
List-Archive: <http://www.geocrawler.com/redir-sf.php3?list=spamassassin-talk>
X-Original-Date: Thu, 22 Aug 2002 10:16:36 -0400
Date: Thu, 22 Aug 2002 10:16:36 -0400
<------------------------------ body ---------------------------------------------------->

I have been trying to research via SA mirrors and search engines if a canned
script exists giving clients access to their user_prefs options via a
web-based CGI interface. Numerous ISPs provide this feature to clients, but
so far I can find nothing. Our configuration uses Amavis-Postfix and ClamAV
for virus filtering and Procmail with SpamAssassin for spam filtering. I
would prefer not to have to write a script myself, but will appreciate any
suggestions.

...

<------------------------------ End ----------------------------------------------------->
+------------------------------------ mail 11 ---------------------------+
Test result:
state 1: 0 ~ 2851
state 2:2852 ~ 3475
<------------------------------ header ------------------------------------------------->
From spamassassin-devel-admin@lists.sourceforge.net  Thu Aug 22 15:25:29 2002
Return-Path: <spamassassin-devel-admin@example.sourceforge.net>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id AE2D043F9B
        for <zzzz@localhost>; Thu, 22 Aug 2002 10:25:29 -0400 (EDT)
...
List-Id: SpamAssassin Developers <spamassassin-devel.example.sourceforge.net>
List-Unsubscribe: <https://example.sourceforge.net/lists/listinfo/spamassassin-devel>,
    <mailto:spamassassin-devel-request@lists.sourceforge.net?subject=unsubscribe>
List-Archive: <http://www.geocrawler.com/redir-sf.php3?list=spamassassin-devel>
X-Original-Date: Thu, 22 Aug 2002 16:19:48 +0200
Date: Thu, 22 Aug 2002 16:19:48 +0200
<------------------------------ body ---------------------------------------------------->

Hello, have you seen and discussed this article and his approach?

Thank you

http://www.paulgraham.com/spam.html
-- "Hell, there are no rules here-- we're trying to accomplish something."
-- Thomas Alva Edison

...
```

```
<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 12 ---------------------------+
Test result:
state 1: 0 ~ 2938
state 2:2939 ~ 3993
<------------------------------ header ----------------------------------------------->
From spamassassin-devel-admin@lists.sourceforge.net  Thu Aug 22 16:27:25 2002
Return-Path: <spamassassin-devel-admin@example.sourceforge.net>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id DF2DD43F9B
        for <zzzz@localhost>; Thu, 22 Aug 2002 11:27:24 -0400 (EDT)
...
List-Id: SpamAssassin Developers <spamassassin-devel.example.sourceforge.net>
List-Unsubscribe: <https://example.sourceforge.net/lists/listinfo/spamassassin-devel>,
    <mailto:spamassassin-devel-request@lists.sourceforge.net?subject=unsubscribe>
List-Archive: <http://www.geocrawler.com/redir-sf.php3?list=spamassassin-devel>
X-Original-Date: Thu, 22 Aug 2002 08:14:12 -0700
Date: Thu, 22 Aug 2002 08:14:12 -0700
<------------------------------ body ------------------------------------------------>

Yes - great minds think alike. But even withput eval rules it would be very
useful. It would allow us to respond quickly to spammer's tricks.

Theo Van Dinter wrote:
> On Thu, Aug 22, 2002 at 07:27:52AM -0700, Marc Perkel wrote:
>
>>Has anyone though of the idea of live updates of rules after release? The
>>idea being that the user can run a cron job once a week or so and get the
...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 13 ---------------------------+
Test result:
state 1: 0 ~ 2304
state 2:2305 ~ 3328
<------------------------------ header ----------------------------------------------->
From ilug-admin@linux.ie  Thu Aug 22 16:27:21 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 7A28A43F99
        for <zzzz@localhost>; Thu, 22 Aug 2002 11:27:21 -0400 (EDT)
...
X-Mailman-Version: 1.1
Precedence: bulk
List-Id: Irish Linux Users' Group <ilug.linux.ie>
```

```
X-Beenthere: ilug@linux.ie

On Mon, Aug 19, 2002 at 03:08:16PM +0100,
<------------------------------ body ---------------------------------------------------->
 John P. Looney mentioned:
>  This is likely because to get it to boot, like the cobalt, I'm actually
> passing root=/dev/hda5 to the kernel, not /dev/md0.

 Just to solve this...the reason I was booting the box with
root=/dev/hda5, not /dev/md0 was because /dev/md0 wasn't booting - it
would barf with 'can't find init'.

 It turns out that this is because I was populating md0 with tar. Which
...

<------------------------------ End ----------------------------------------------------->
+------------------------------------ mail 14 ---------------------------+
Test result:
state 1: 0 ~ 4813
state 2:4814 ~ 6576
<------------------------------ header ------------------------------------------------->
From exmh-workers-admin@redhat.com  Thu Aug 22 16:37:36 2002
Return-Path: <exmh-workers-admin@spamassassin.taint.org>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 50AF343F9B
        for <zzzz@localhost>; Thu, 22 Aug 2002 11:37:35 -0400 (EDT)
...
> > Ouch...I'll get right on it.
> >
> > > From:  Robert Elz <kre@munnari.OZ.AU>
> > > Date:  Wed, 21 Aug 2002 19:30:01 +0700
> > >
> > > An
<------------------------------ body ---------------------------------------------------->
y chance of having that lengthen instead?   I like all my exmh stuff
> > > in nice columns (fits the display better).   That is, I use the detache
> d
> > > folder list, one column.   The main exmh window takes up full screen,
> > > top to bottom, but less than half the width, etc...
>
> I thought about that.  The first order approximation would be to just add
> using pack ... -side top instead of pack ... -side left, however, since their
> each a different width, it would look funny.
...

<------------------------------ End ----------------------------------------------------->
+------------------------------------ mail 15 ---------------------------+
```

```
Test result:
state 1: 0 ~ 2183
state 2:2184 ~ 6808
<------------------------------ header ------------------------------------------------>
From fork-admin@xent.com  Thu Aug 22 16:37:41 2002
Return-Path: <fork-admin@xent.com>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 5DB9843F99
        for <zzzz@localhost>; Thu, 22 Aug 2002 11:37:40 -0400 (EDT)
...
List-Unsubscribe: <http://xent.com/mailman/listinfo/fork>,
    <mailto:fork-request@xent.com?subject=unsubscribe>
List-Archive: <http://xent.com/pipermail/fork/>
Date: Thu, 22 Aug 2002 12:39:47 -0300

SpamAs
<------------------------------ body ------------------------------------------------>
sassin is hurting democracy!
Owen
-------------------------------------------------------------------------------------------

http://www.bayarea.com/mld/mercurynews/news/opinion/3900215.htm

Internet can level the political playing field
By Mike McCurry and Larry Purpuro

...

<------------------------------ End ------------------------------------------------->
+------------------------------------ mail 16 ----------------------------+
Test result:
state 1: 0 ~ 1971
state 2:1972 ~ 2627
<------------------------------ header ------------------------------------------------>
From iiu-admin@taint.org  Thu Aug 22 17:08:44 2002
Return-Path: <iiu-admin@taint.org>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id B989743F9B
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:08:43 -0400 (EDT)
...
List-Post: <mailto:iiu@iiu.taint.org>
List-Help: <mailto:iiu-request@iiu.taint.org?subject=help>
List-Subscribe: <http://iiu.taint.org/mailman/listinfo/iiu>,
    <mailto:iiu-request@iiu.taint.org?subject=subscribe>
List-Archive: <http://iiu.taint.org/pipermail/iiu/>
Date: Thu, 22 Aug 2002 16:58:37 +0100
```

```
<------------------------------ body ------------------------------------------------>

Hi all,

apologies for the possible silly question (i don't think it is, but),
but is Eircom's aDSL service NAT'ed?

and what implications would that have for VoIP? I know there are
difficulties with VoIP or connecting to clients connected to a NAT'ed
network from the internet wild (i.e. machines with static, real IPs)
...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 17 ---------------------------+
Test result:
state 1: 0 ~ 2282
state 2:2283 ~ 3425
<------------------------------ header ----------------------------------------------->
From robert.chambers@baesystems.com  Thu Aug 22 17:19:36 2002
Return-Path: <robert.chambers@baesystems.com>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id F2AD843F99
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:19:25 -0400 (EDT)
...
Reply-To: zzzzteana@yahoogroups.com
Content-Type: text/plain; charset=US-ASCII
Content-Transfer-Encoding: 7bit

--- In forteana@y..., "D.McMann" <dmcmann@b...> wrote:
> Robert Moaby, 33,
<------------------------------ body ------------------------------------------------>
 who sent death threats to staff, was also jailed
> for hoarding indecent pictures of children on his home computer.
> =========
>
> Hmm, if I didn't trust our government and secret police, I could
look at
> this another way...

There is a bit of circumstantial evidence - apparently some MT
...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 18 ---------------------------+
Test result:
state 1: 0 ~ 2367
state 2:2368 ~ 3077
```

```
<------------------------------ header ------------------------------------------------>
From ilug-admin@linux.ie  Thu Aug 22 17:19:31 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 6622747C69
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:19:23 -0400 (EDT)
...
> Sent: 22 August 2002 17:23
> To: ILUG
> Subject: [ILUG] Sun Solaris..
>
>
> Ca
<------------------------------ body ------------------------------------------------>
n someone explain what type of operating system Solaris
> is... as ive never seen or used it i dont know wheather to
> get a server from Sun or from DELL i would prefer a linux
> based server and Sun seems to be the one for that but im not
> sure if Solaris is a distro of linux or a completely
> different operating system? can someone explain...
>
> Kiall Mac Innes
>
...

<------------------------------ End ------------------------------------------------->
+------------------------------------ mail 19 ---------------------------+
Test result:
state 1: 0 ~ 2101
state 2:2102 ~ 2620
<------------------------------ header ------------------------------------------------>
From robert.chambers@baesystems.com  Thu Aug 22 17:19:36 2002
Return-Path: <robert.chambers@baesystems.com>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id C1E1343F9B
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:19:26 -0400 (EDT)
...
List-Unsubscribe: <mailto:zzzzteana-unsubscribe@yahoogroups.com>
Date: Thu, 22 Aug 2002 16:17:39 -0000
Subject: [zzzzteana] Which Muppet Are You?
Reply-To: zzzzteana@yahoogroups.com
Content-Type: text/plain; charset=US-ASCII
Co
<------------------------------ body ------------------------------------------------>
ntent-Transfer-Encoding: 7bit
```

```
Apols if this has been posted before:

http://www.pinkpaperclips.net/subs/quiz2.html

Rob


...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 20 --------------------------+
Test result:
state 1: 0 ~ 1840
state 2:1841 ~ 2434
<------------------------------ header ----------------------------------------------->
From ilug-admin@linux.ie  Thu Aug 22 17:19:25 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id CD34B47C67
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:19:21 -0400 (EDT)
...
Subject: [ILUG] Sun Solaris..
Sender: ilug-admin@linux.ie
Errors-To: ilug-admin@linux.ie
X-Mailman-Version: 1.1
Precedence: bulk
Li
<------------------------------ body ------------------------------------------------>
st-Id: Irish Linux Users' Group <ilug.linux.ie>
X-Beenthere: ilug@linux.ie

Can someone explain what type of operating system Solaris is... as ive never
seen or used it i dont know wheather to get a server from Sun or from DELL i
would prefer a linux based server and Sun seems to be the one for that but
im not sure if Solaris is a distro of linux or a completely different
operating system? can someone explain...

...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 21 --------------------------+
Test result:
state 1: 0 ~ 2103
state 2:2104 ~ 2664
<------------------------------ header ----------------------------------------------->
From timc@2ubh.com  Thu Aug 22 17:31:00 2002
Return-Path: <timc@2ubh.com>
```

```
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id A97BE43F99
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:30:58 -0400 (EDT)
...
List-Unsubscribe: <mailto:zzzzteana-unsubscribe@yahoogroups.com>
Date: Thu, 22 Aug 2002 17:23:28 +0100
Subject: Re: [zzzzteana] Which Muppet Are You?
Reply-To: zzzzteana@yahoogroups.com
Content-Type: text/plain; charset=US-ASCII
Co
<------------------------------ body ------------------------------------------------->
ntent-Transfer-Encoding: 7bit

> Apols if this has been posted before:
>
> http://www.pinkpaperclips.net/subs/quiz2.html
>
So, anyone who isn't Beaker?

TimC
...

<------------------------------ End -------------------------------------------------->
+------------------------------------ mail 22 ---------------------------+
Test result:
state 1: 0 ~ 2234
state 2:2235 ~ 3643
<------------------------------ header ------------------------------------------------>
From ilug-admin@linux.ie  Thu Aug 22 17:45:53 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 08BC143F99
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:45:53 -0400 (EDT)
...
X-Mailman-Version: 1.1
Precedence: bulk
List-Id: Irish Linux Users' Group <ilug.linux.ie>
X-Beenthere: ilug@linux.ie

On Thu, Aug 22, 2002 at 05:13:01PM +0100,
<------------------------------ body ------------------------------------------------->
 Fergal Moran mentioned:
> In a nutshell - Solaris is Suns own flavour of UNIX.

 Though I'm sure that this nice person would like a bit more detail.
```

```
 Solaris is quite different to Linux, though these days you can make
solaris act a lot like linux with an extra CD of GNU tools Sun ship with
solaris. It is based on the SysV unix family, so it's quite similar to
other unixen like HPUX and SCO.
...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 23 --------------------------+
Test result:
state 1: 0 ~ 2168
state 2:2169 ~ 3750
<------------------------------ header ----------------------------------------------->
From ilug-admin@linux.ie  Thu Aug 22 17:45:54 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id E8B1343F9B
        for <zzzz@localhost>; Thu, 22 Aug 2002 12:45:53 -0400 (EDT)
...
Precedence: bulk
List-Id: Irish Linux Users' Group <ilug.linux.ie>
X-Beenthere: ilug@linux.ie

John P. Looney wrote:
> On Thu, Aug 22, 2002 at 05:13:01PM +0100,
<------------------------------ body ------------------------------------------------>
 Fergal Moran mentioned:
>
>>In a nutshell - Solaris is Suns own flavour of UNIX.
>
>
>  Though I'm sure that this nice person would like a bit more detail.
>
>  Solaris is quite different to Linux, though these days you can make
> solaris act a lot like linux with an extra CD of GNU tools Sun ship with
...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 24 --------------------------+
Test result:
state 1: 0 ~ 2560
state 2:2561 ~ 3701
<------------------------------ header ----------------------------------------------->
From lejones@ucla.edu  Thu Aug 22 18:29:58 2002
Return-Path: <lejones@ucla.edu>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 89B2943F99
```

```
        for <zzzz@localhost>; Thu, 22 Aug 2002 13:29:49 -0400 (EDT)
...
List-Unsubscribe: <mailto:zzzzteana-unsubscribe@yahoogroups.com>
Date: Thu, 22 Aug 2002 10:19:48 -0700
Subject: Re: [zzzzteana] Which Muppet Are You?
Reply-To: zzzzteana@yahoogroups.com
Content-Type: text/plain; charset=US-ASCII
Co
<------------------------------ body ---------------------------------------------------->
ntent-Transfer-Encoding: 7bit

Hey, it's not easy being green.

leslie

Leslie Ellen Jones, Ph.D.
Jack of All Trades and Doctor of Folklore
lejones@ucla.edu
...

<------------------------------ End ----------------------------------------------------->
+------------------------------------ mail 25 ---------------------------+
Test result:
state 1: 0 ~ 2319
state 2:2320 ~ 3238
<------------------------------ header -------------------------------------------------->
From ilug-admin@linux.ie  Fri Aug 23 11:07:47 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 6F82C4416B
        for <zzzz@localhost>; Fri, 23 Aug 2002 06:06:31 -0400 (EDT)
...
Precedence: bulk
List-Id: Irish Linux Users' Group <ilug.linux.ie>
X-Beenthere: ilug@linux.ie


> On Thu, 22 Aug 2002,
<------------------------------ body ---------------------------------------------------->
 John P. Looney wrote:
> >  Sun's hardware in general is more reliable,
> ROFL. not in our experience.

Well at least our Caps-Lock keys work:

peter@staunton.ie said:
> Another problem. I have a Dell branded keyboard and if I hit Caps-Lock
```

```
> twice, the whole machine crashes (in Linux, not Windows) - even the on/
...

<------------------------------ End ---------------------------------------------->
+------------------------------------ mail 26 ---------------------------+
Test result:
state 1: 0 ~ 2027
state 2:2028 ~ 4467
<------------------------------ header ------------------------------------------->
From fork-admin@xent.com  Fri Aug 23 11:08:20 2002
Return-Path: <fork-admin@xent.com>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 93D8944160
        for <zzzz@localhost>; Fri, 23 Aug 2002 06:06:44 -0400 (EDT)
...
List-Subscribe: <http://xent.com/mailman/listinfo/fork>, <mailto:fork-request@xent.com?subject=
List-Id: Friends of Rohit Khare <fork.xent.com>
List-Unsubscribe: <http://xent.com/mailman/listinfo/fork>,
    <mailto:fork-request@xent.com?subject=unsubscribe>
List-Archive: <http://xent.com/pipermail/fork/>
Date: Thu, 22 Aug 2002 11:11:57 -0700
<------------------------------ body --------------------------------------------->



>                                  You have multiple generations of
> peasants/squatters that cultivate and live on the lands almost as a
> human parts of the property package.

When I'd read that "getting legal title
can take 20 years", when I believe that
...

<------------------------------ End ---------------------------------------------->
+------------------------------------ mail 27 ---------------------------+
Test result:
state 1: 0 ~ 1771
state 2:1772 ~ 3148
<------------------------------ header ------------------------------------------->
From ilug-admin@linux.ie  Fri Aug 23 11:07:42 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 762374415C
        for <zzzz@localhost>; Fri, 23 Aug 2002 06:06:30 -0400 (EDT)
...
Subject: [ILUG] Newbie seeks advice - Suse 7.2
```

```
Sender: ilug-admin@linux.ie
Errors-To: ilug-admin@linux.ie
X-Mailman-Version: 1.1
Precedence: bulk
Li
<------------------------------ body ------------------------------------------------>
st-Id: Irish Linux Users' Group <ilug.linux.ie>
X-Beenthere: ilug@linux.ie

Folks,

my first time posting - have a bit of Unix experience, but am new to Linux.


Just got a new PC at home - Dell box with Windows XP. Added a second hard disk
...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 28 ---------------------------+
Test result:
state 1: 0 ~ 2225
state 2:2226 ~ 2541
<------------------------------ header ----------------------------------------------->
From fork-admin@xent.com  Fri Aug 23 11:08:26 2002
Return-Path: <fork-admin@xent.com>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id BA70647C68
        for <zzzz@localhost>; Fri, 23 Aug 2002 06:06:46 -0400 (EDT)
...
List-Unsubscribe: <http://xent.com/mailman/listinfo/fork>,
    <mailto:fork-request@xent.com?subject=unsubscribe>
List-Archive: <http://xent.com/pipermail/fork/>
Date: Thu, 22 Aug 2002 15:25:24 -0400 (EDT)

On Thu, 22 Aug 2002,
<------------------------------ body ------------------------------------------------>
 Joseph S. Barrera III wrote:
--]Why wait until you're dead? I'm sure there's enough carbon in
--]the fat from your typical liposuction job to make a decent diamond.

So thats why I keep seeing DeBeers agents hovering around me.

-tom(diamonds in the folds of my flesh)wsmf


...
```

```
<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 29 ---------------------------+
Test result:
state 1: 0 ~ 2344
state 2:2345 ~ 2890
<------------------------------ header ----------------------------------------------->
From fork-admin@xent.com  Fri Aug 23 11:08:30 2002
Return-Path: <fork-admin@xent.com>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id C348B44163
        for <zzzz@localhost>; Fri, 23 Aug 2002 06:06:47 -0400 (EDT)
...
List-Subscribe: <http://xent.com/mailman/listinfo/fork>, <mailto:fork-request@xent.com?subject=
List-Id: Friends of Rohit Khare <fork.xent.com>
List-Unsubscribe: <http://xent.com/mailman/listinfo/fork>,
    <mailto:fork-request@xent.com?subject=unsubscribe>
List-Archive: <http://xent.com/pipermail/fork/>
Date: Thu, 22 Aug 2002 16:30:07 -0300
<------------------------------ body ------------------------------------------------>

Joseph S. Barrera III wrote:

> Chris Haun wrote:
>
>> A LifeGem is a certified, high quality diamond created from the
>> carbon of your loved one as a memorial to their unique and wonderful
>> life.
>
...

<------------------------------ End ------------------------------------------------>
+------------------------------------ mail 30 ---------------------------+
Test result:
state 1: 0 ~ 2173
state 2:2174 ~ 5160
<------------------------------ header ----------------------------------------------->
From ilug-admin@linux.ie  Fri Aug 23 11:07:51 2002
Return-Path: <ilug-admin@linux.ie>
Delivered-To: zzzz@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 7419C4416C
        for <zzzz@localhost>; Fri, 23 Aug 2002 06:06:33 -0400 (EDT)
...
    UAA19403
Sender: ilug-admin@linux.ie
Errors-To: ilug-admin@linux.ie
X-Mailman-Version: 1.1
```

```
Precedence: bulk
Li
<----------------------------- body ------------------------------------------------->
st-Id: Irish Linux Users' Group <ilug.linux.ie>
X-Beenthere: ilug@linux.ie

Update on this for anyone that's interested, and because I like closed
threads... nothing worse than an infinite while loop, is there?

I ended up formatting a floppy on my flatmate's (un-networked) P100 running
FAT16 Win95, and mcopied the contents of the bootdisk across. Now I have a
FAT16 Win98 install running alongside Slackware, and can play Metal Gear
...

<----------------------------- End -------------------------------------------------->
```

In most of condition, the algorithm can split the mail well. But we can also find that if there a mail included at the begining of the mail body, the algorithm will meet difficulty to find the bound between the body and the header, we can find this condition in **mail14**, **mail18** and **mail23**. Sometimes we can find the algorithm find the bound in the middle of line. it's not reasonable. So I think we should condiser escape characters as a character in order to resolve this problem.

1.5 Question5. How would you model the problem if you had to segment the mails in more than two parts (for example : header, body, signature) ?

In this case, we firstly have to recalculate the transition matrix and emission matrix on the basis of the samples. The transition matrix will hence become a 3x3 matrix of this form:

$$\begin{bmatrix} p_{11} & p_{12} & 0 \\ 0 & p_{22} & p_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

The transition matrix will hence become a 256x3 matrix: and the initial vector will become :

$$^{T} = (1, 0, 0)$$

.

1.6 6. How would you model the problem of separating the portions of mail included, knowing that they always start with the character ">".

In this case, the model would have four states:header, body_text, body_included, and signature. So we need a transition matrix 4x4 of this form:

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & 0 \\ 0 & p_{22} & p_{23} & p_{24} \\ 0 & p_{32} & p_{33} & p_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and the initial vector :

$$^{T} = (1, 0, 0, 0)$$

We also need to recalculate the emission matrix.

We can also traite the lines instead of character.Beacause most of the bound between different part is not in the middle of line. if a line start with ">", the conditional probability of this line in 'mail include state' will increase. We can also apply the bigram to increase accuracy rate.