intro_to_ai_ass4

April 25, 2023

1 DAT405/DIT407 Introduction to Data Science and AI

1.1 2023-2024, Reading Period 4

Name	Working Hours
Dimitrios Koutsakis	8
Bingcheng Chen	8

1.2 Assignment 4: Spam classification using Naïve Bayes

The exercise takes place in this notebook environment. Hints: You can execute certain linux shell commands by prefixing the command with `!`. You can insert Markdown cells and code cells. The first you can use for documenting and explaining your results the second you can use writing code snippets that execute the tasks required.

In this assignment you will implement a Naïve Bayes classifier in Python that will classify emails into spam and non-spam ("ham") classes. Your program should be able to train on a given set of spam and "ham" datasets. You will work with the datasets available at https://spamassassin.apache.org/old/publiccorpus/. There are three types of files in this location: - easy-ham: non-spam messages typically quite easy to differentiate from spam messages. - hard-ham: non-spam messages more difficult to differentiate - spam: spam messages

Execute the cell below to download and extract the data into the environment of the notebook – it will take a few seconds. If you chose to use Jupyter notebooks you will have to run the commands in the cell below on your local computer, with Windows you can use 7zip (https://www.7-zip.org/download.html) to decompress the data.

What to submit: Convert the notebook to a pdf-file and submit it. Make sure all cells are executed so all your code and its results are included. Double check the pdf displays correctly before you submit it.

[ ]:
```
#Download and extract data
# !wget https://spamassassin.apache.org/old/publiccorpus/20021010_easy_ham.tar.
 ↪bz2
# !wget https://spamassassin.apache.org/old/publiccorpus/20021010_hard_ham.tar.
 ↪bz2
# !wget https://spamassassin.apache.org/old/publiccorpus/20021010_spam.tar.bz2
# !tar -xjf 20021010_easy_ham.tar.bz2
# !tar -xjf 20021010_hard_ham.tar.bz2
```

```
# !tar -xjf 20021010_spam.tar.bz2
```

The data is now in the three folders `easy_ham`, `hard_ham`, and `spam`.

[ ]:
```
!ls -lah
```

```
'ls' is not recognized as an internal or external command,
operable program or batch file.
```

1.2.1 1. Preprocessing:

1.1 Look at a few emails from easy_ham, hard_ham and spam. Do you think you would be able to classify the emails just by inspection? How do you think a succesful model can learn the difference between the different classes of emails? Answer:

It's not hard for me to classify the emails, since spam emails are usually with the intention of promoting a product, service. there are several features that a spam email may have: - Suspicious sender: the email sent from a unfamiliar email address. - Deceptive subject: A spam email may use deceptive subject such as offering free products to attract receivers' clicks - Spam-like word: A spam email is more likely to use some specific words, such as dollar, invest, free.

A successful model can learn the difference between the different classes of emails through data science and AI, like what we have done in this assignment - using Naïve Bayes Classifiers: - First, feed the model with high quality datasets of different types of emails, which have already been labeled as spam or ham by human. - Second, train the model on these datasets, learning to identify patterns in the features of the emails that are associated with spam or ham. - Third, test the model on a separate set of dataset to evaluate its performance. - Fourth, iterate and optimise the model through the training and testing process, adjust the model's parameters and feed more high quality labeled datasets to the model.

[ ]:
```
import os
import random
import numpy as np

# Set directory path for easy_ham and spam
easy_ham_dir = 'easy_ham'
hard_ham_dir = 'hard_ham'
spam_dir = 'spam'

# Get the list of file names in each fold
easy_ham_files = [os.path.join(easy_ham_dir, file) for file in os.
 ↪listdir(easy_ham_dir)]
hard_ham_files = [os.path.join(hard_ham_dir, file) for file in os.
 ↪listdir(hard_ham_dir)]
spam_files = [os.path.join(spam_dir, file) for file in os.listdir(spam_dir)][1:]

# Choose 1 random file from each type
seed = 42
random.seed(seed)
```

```
np.random.seed(seed)

random_easy_ham = random.sample(easy_ham_files, 1)
random_hard_ham = random.sample(hard_ham_files, 1)
random_spam = random.sample(spam_files, 1)

# Print the name of the chosen files
print(random_easy_ham)
print(random_hard_ham)
print(random_spam)

# List of files to be read
read_list = random_easy_ham + random_hard_ham +random_spam

# Loop through the list and read each file
for file_name in read_list:
    with open(file_name, 'r', encoding='utf-8', errors='ignore') as file:
        # print contents
        print('############################## No.{}␣
 ↪email###############################'.format(read_list.index(file_name) + 1))
        print(file.read())
```

```
['easy_ham\\0457.dc1691cbb334cc33a1f1eb3060b8e02e']
['hard_ham\\0007.7f2ea3a532284cff3321e5ba159cdb50']
['spam\\0380.c4d530b5816543f4f1a23b8ce0d281f5']
############################## No.1 email###############################
From rpm-list-admin@freshrpms.net  Mon Aug 26 20:15:19 2002
Return-Path: <rpm-zzzlist-admin@freshrpms.net>
Delivered-To: yyyy@localhost.netnoteinc.com
Received: from localhost (localhost [127.0.0.1])
        by phobos.labs.netnoteinc.com (Postfix) with ESMTP id A7E4643F99
        for <jm@localhost>; Mon, 26 Aug 2002 15:15:18 -0400 (EDT)
Received: from phobos [127.0.0.1]
        by localhost with IMAP (fetchmail-5.9.0)
        for jm@localhost (single-drop); Mon, 26 Aug 2002 20:15:18 +0100 (IST)
Received: from egwn.net (auth02.nl.egwn.net [193.172.5.4]) by
    dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g7QJ8VZ05732 for
    <jm-rpm@jmason.org>; Mon, 26 Aug 2002 20:08:31 +0100
Received: from auth02.nl.egwn.net (localhost [127.0.0.1]) by egwn.net
    (8.11.6/8.11.6/EGWN) with ESMTP id g7QJ52J30477; Mon, 26 Aug 2002 21:05:02
    +0200
Received: from kamakiriad.com
    (IDENT:nZbdv/p4nmL0skumLgaQPfpaAEkGbyHy@cable-b-36.sigecom.net
    [63.69.210.36]) by egwn.net (8.11.6/8.11.6/EGWN) with ESMTP id
    g7QJ4uJ30412 for <rpm-list@freshrpms.net>; Mon, 26 Aug 2002 21:04:56 +0200
Received: from aquila.kamakiriad.local (aquila.kamakiriad.local
    [192.168.1.3]) by kamakiriad.com (8.11.6/8.11.0) with SMTP id g7QJ4me30195
    for <rpm-list@freshrpms.net>; Mon, 26 Aug 2002 14:04:49 -0500
```

From: Brian Fahrlander <kilroy@kamakiriad.com>
To: rpm-zzzlist@freshrpms.net
Subject: Re: New gkrellm 2.0.0, gtk2 version
Message-Id: <20020826140448.208e3da8.kilroy@kamakiriad.com>
In-Reply-To: <20020826191454.43e6c15f.matthias@egwn.net>
References: <20020826191454.43e6c15f.matthias@egwn.net>
X-Mailer: Sylpheed version 0.8.1 (GTK+ 1.2.10; i386-redhat-linux)
X-Message-Flag: : Shame on you! You know Outlook is how viruses are spread!
MIME-Version: 1.0
Content-Type: text/plain; charset=ISO-8859-1
Content-Transfer-Encoding: 8bit
X-Mailscanner: Found to be clean, Found to be clean
Sender: rpm-zzzlist-admin@freshrpms.net
Errors-To: rpm-zzzlist-admin@freshrpms.net
X-Beenthere: rpm-zzzlist@freshrpms.net
X-Mailman-Version: 2.0.11
Precedence: bulk
Reply-To: rpm-zzzlist@freshrpms.net
List-Help: <mailto:rpm-zzzlist-request@freshrpms.net?subject=help>
List-Post: <mailto:rpm-zzzlist@freshrpms.net>
List-Subscribe: <http://lists.freshrpms.net/mailman/listinfo/rpm-zzzlist>,
    <mailto:rpm-list-request@freshrpms.net?subject=subscribe>
List-Id: Freshrpms RPM discussion list <rpm-zzzlist.freshrpms.net>
List-Unsubscribe: <http://lists.freshrpms.net/mailman/listinfo/rpm-zzzlist>,
    <mailto:rpm-list-request@freshrpms.net?subject=unsubscribe>
List-Archive: <http://lists.freshrpms.net/pipermail/rpm-zzzlist/>
X-Original-Date: Mon, 26 Aug 2002 14:04:48 -0500
Date: Mon, 26 Aug 2002 14:04:48 -0500
X-Pyzor: Reported 0 times.
X-Spam-Status: No, hits=-8.2 required=7.0
        tests=EMAIL_ATTRIBUTION,IN_REP_TO,KNOWN_MAILING_LIST,
              QUOTED_EMAIL_TEXT,REFERENCES,SPAM_PHRASE_00_01
        version=2.40-cvs
X-Spam-Level:

On Mon, 26 Aug 2002 19:14:54 +0200, Matthias Saou <matthias@egwn.net> wrote:

> Hi all,
>
> I've repackaged the new gkrellm 2.0.0 which is now ported to gtk2
> (woohoo!). Unfortunately, the plugins are incompatible with the previous
> 1.2.x ones, and since not many/all have been ported yet, I prefer not to
> release the package on the main freshrpms.net site for now.
>
> For those of you who'd like to try it out, you can grab it here :
> http://ftp.freshrpms.net/pub/freshrpms/testing/gkrellm/
>
> I think the themes are still compatible, but haven't tried to install some

> with 2.0.0 yet.
> Last note, the above packages are for Valhalla. And yes, although GNOME 2
> is not in Valhalla, gtk2 and glib2 have been from the very beginning! ;-)

    Sweet, dude- I was really hoping it'd be out sooner or later; thanks a
bunch!

Brian Fahrlnder   Linux Zealot, Conservative, and Technomad
Evansville, IN   My Voyage: http://www.CounterMoon.com
ICQ 5119262

I've been complaining for years, and almost no one listened. "Windows is
just easier" you said. "I don't want to learn anything new", you said.
Tell me how easy THIS is:
http://www.guardian.co.uk/Archive/Article/0,4273,4477138,00.html

---
RPM-List mailing list <RPM-List@freshrpms.net>
http://lists.freshrpms.net/mailman/listinfo/rpm-list

############################## No.2 email################################
Return-Path: <2.20290.44-t9bsgc0tYwDu.1.b@ummail4.unitedmedia.com>
Received: from ummail1.unitedmedia.com (ummail1.unitedmedia.com [65.114.4.73])
        by dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g6AAGiT13749
        for <qqqqqqqqqq-dilbert@example.com>; Wed, 10 Jul 2002 11:16:44 +0100
Received: from umsan1 (10.1.1.75) by ummail1.unitedmedia.com (PowerMTA(TM)
v1.5); Wed, 10 Jul 2002 06:08:46 -0400 (envelope-from
<2.20290.44-t9bsgc0tYwDu.1.b@ummail4.unitedmedia.com>)
Message-ID: <24288927.1026296118194.JavaMail.root@umsan1>
Date: Wed, 10 Jul 2002 06:15:18 -0400 (EDT)
From: Daily Dilbert <2.20290.44-t9bsgc0tYwDu.1@ummail4.unitedmedia.com>
To: qqqqqqqqqq-dilbert@example.com
Subject: Your Daily Dilbert 07/10/2002
Mime-Version: 1.0
Content-Type: multipart/alternative;
boundary=23561619.1026296118170.JavaMail.root.umsan1

--23561619.1026296118170.JavaMail.root.umsan1
Content-Type: text/plain; charset=ISO-8859-1
Content-Transfer-Encoding: 7bit

E-mail error

You're subscribed to the HTML version of the Daily Dilbert,
which shows the comic strip as a graphic, but your mail system
either can't support HTML or is set up to remove HTML content. For

```
more information, contact your Internet service provider or mail
system administrator.

To change to a plain text subscription, modify your account
preferences at
http://www.dilbert.com/comics/dilbert/daily_dilbert/html/login.html

The plain text option appears toward the bottom of the modification
page.




--23561619.1026296118170.JavaMail.root.umsan1
Content-Type: text/html; charset=ISO-8859-1
Content-Transfer-Encoding: quoted-printable

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">

<HTML>
<HEAD>
<TITLE>Daily Dilbert</TITLE>
<meta http-equiv=3D"pragma" content=3D"no-cache">
<META HTTP-EQUIV=3D"expires" CONTENT=3D"0">
<meta http-equiv=3D"Cache-control" content=3D"no-cache">
</HEAD>

<BODY BGCOLOR=3D"#336699" text=3D"#000000" link=3D"#000000" vlink=3D"#00339=
9" alink=3D"#3333CC" MARGINWIDTH=3D"0" MARGINHEIGHT=3D"8" LEFTMARGIN=3D"0" =
TOPMARGIN=3D"8">

<TABLE WIDTH=3D"100%" BORDER=3D"0" CELLPADDING=3D"0" BGCOLOR=3D"#336699">
<TR>
<TD BGCOLOR=3D"#336699" WIDTH=3D"15"><IMG SRC=3D"http://www.comics.com/comi=
cs/dilbert/images/clear_dot.gif" WIDTH=3D"15" BORDER=3D"0" ALT=3D""></TD>
<TD BGCOLOR=3D"#336699">



<SCRIPT LANGUAGE=3D"JAVASCRIPT">
<!--=20
function rand(n)=20
{
seed =3D (0x015a4e35 * seed) % 0xffffffff;
return (seed >> 16) % n;
}
var now =3D new Date ();
```

```
var seed =3D now.getTime() % 0xffffffff;
var same =3D rand(1000);
// End hiding script from old browsers -->
</SCRIPT>

<SCRIPT Language=3D"JavaScript">
<!-- hide from old browsers

  var today =3D new Date()

//-->
</SCRIPT>



<!-- CONTENT AND SKYSCRAPER AD TABLE BEGIN -->
<TABLE CELLPADDING=3D"0" CELLSPACING=3D"0" BORDER=3D"0">
=09<TR VALIGN=3D"TOP">=09
=09=09<TD BGCOLOR=3D"#336699">

=09=09=09<!-- AD TABLE BEGIN -->
=09=09=09<TABLE WIDTH=3D"627" CELLPADDING=3D"0" CELLSPACING=3D"0" BORDER=3D=
"0">
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"468" BGCOLOR=3D"#336699"><IMG SRC=3D"http://www=
.comics.com/comics/dilbert/images/small_ad.gif" WIDTH=3D"20" HEIGHT=3D"11" =
BORDER=3D"0" ALT=3D"Advertisement"><BR><IMG SRC=3D"http://www.comics.com/co=
mics/dilbert/images/clear_dot.gif" WIDTH=3D"1" HEIGHT=3D"1" BORDER=3D"0" AL=
T=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"21" BGCOLOR=3D"#336699" ROWSPAN=3D"2"><IMG SRC=
=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"21"=
 HEIGHT=3D"72" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"120" BGCOLOR=3D"#336699"><IMG SRC=3D"http://www=
.comics.com/comics/dilbert/images/small_ad.gif" WIDTH=3D"20" HEIGHT=3D"11" =
BORDER=3D"0" ALT=3D"Advertisement"><BR><IMG SRC=3D"http://www.comics.com/co=
mics/dilbert/images/clear_dot.gif" WIDTH=3D"1" HEIGHT=3D"1" BORDER=3D"0" AL=
T=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"18" BGCOLOR=3D"#336699" ROWSPAN=3D"2"><IMG SRC=
=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"18"=
 HEIGHT=3D"72" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09
=09=09=09=09
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"468" BGCOLOR=3D"#336699">
=09=09=09=09=09<!-- BEGIN ENHANCED CREATIVE -->
=09=09=09=09=09<SCRIPT LANGUAGE=3D"JavaScript">
=09=09=09=09=09<!--
```

```
=09=09=09=09=09document.write ('<TABLE WIDTH=3D"468" CELLSPACING=3D"0" CELL=
PADDING=3D"0" BORDER=3D"0"><TR VALIGN=3D"TOP"><TD ALIGN=3D"LEFT"><ILAYER ID=
=3Dph1 VISIBILITY=3Dhidden width=3D"468" height=3D"60"></ILAYER></TD></TR><=
/TABLE>');
=09=09=09=09=09document.write ('<NOLAYER>');
=09=09=09=09=09document.write ('<IFRAME SRC=3D"http://ad.doubleclick.net/ad=
i/dilbert.email.com/email/;sz=3D468x60;ptile=3D1;ord=3D' + same + '?" name=
=3D"frame1" width=3D"468" height=3D"60" frameborder=3D"no" border=3D"0" MAR=
GINWIDTH=3D"0" MARGINHEIGHT=3D"0" SCROLLING=3D"no">');
=09=09=09=09=09document.write ('<A target=3D"_blank" HREF=3D"http://ummail4=
.unitedmedia.com:80/Click?q=3D84-CCDyQ60lyggM3vtZsnhhpYUoZsRR">');
=09=09=09=09=09document.write ('<IMG SRC=3D"http://ad.doubleclick.net/ad/di=
lbert.email.com/email/;abr=3D!ie;sz=3D468x60;ptile=3D1;ord=3D' + same + '?"=
 border=3D"0" WIDTH=3D"468" HEIGHT=3D"60"></A></IFRAME>');
=09=09=09=09=09document.write ('</NOLAYER>');=20
=09=09=09=09=09//-->
=09=09=09=09=09</SCRIPT>
=09=09=09=09=09<NOSCRIPT>
=09=09=09=09=09<A target=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:=
80/Click?q=3D99-4mayQAdljOrY53a2HHfS87c8B9RR"><IMG SRC=3D"http://ad.doublec=
lick.net/ad/dilbert.email.com/email/;abr=3D!ie;sz=3D468x60;ptile=3D1;ord=3D=
1986839?" BORDER=3D"0" WIDTH=3D"468" HEIGHT=3D"60"></a>
=09=09=09=09=09</NOSCRIPT>
=09=09=09=09=09<!-- END ENHANCED CREATIVE --></TD>
=09=09=09=09=09
=09=09=09=09=09
=09=09=09=09=09
=09=09=09=09=09<TD WIDTH=3D"120" BGCOLOR=3D"#336699"><IMG SRC=3D"http://www=
.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"1" HEIGHT=3D"1" B=
ORDER=3D"0" ALT=3D""><BR><A target=3D"_blank" HREF=3D"http://ummail4.united=
media.com:80/Click?q=3Dae-jOB6QMB3LJiS_0q8-1hV197zDRRR"><IMG SRC=3D"http://=
www.comics.com/comics/dilbert/daily_dilbert/images/dilbert_tshirt_120x60.gi=
f" WIDTH=3D"120" HEIGHT=3D"60" BORDER=3D"0" ALT=3D"Shop Dilbert"></A></TD>
=09=09=09=09</TR>
=09=09=09=09
=09=09=09=09
=09=09=09=09
=09=09=09=09
=09=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD WIDTH=3D"627" HEIGHT=3D"10" COLSPAN=3D"4"><IMG SRC=3D"ht=
tp://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"627" HEIG=
HT=3D"10" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09</TR>
=09=09=09=09</TABLE>
=09=09=09=09<!-- AD TABLE END -->
=09=09=09=09
=09=09=09=09<!-- HEADER TABLE BEGIN -->
=09=09=09=09<TABLE WIDTH=3D"627" CELLPADDING=3D"0" CELLSPACING=3D"0" BORDER=3D=
```

```
"0">
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"515" BGCOLOR=3D"#336699" ROWSPAN=3D"3"><A targe=
t=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3Dc3-x1sCQJL=
o-Qw0W2VZ57n0NcndxRRR"><IMG SRC=3D"http://www.comics.com/comics/dilbert/dai=
ly_dilbert/images/daily_dilbert_header_left.gif" WIDTH=3D"515" HEIGHT=3D"76=
" BORDER=3D"0" ALT=3D"Daily Dilbert"></A></TD>
=09=09=09=09=09<TD WIDTH=3D"90" BGCOLOR=3D"#336699"><IMG SRC=3D"http://www.=
comics.com/comics/dilbert/daily_dilbert/images/daily_dilbert_header_mid_top=
.gif" WIDTH=3D"90" HEIGHT=3D"55" BORDER=3D"0" ALT=3D"Daily Dilbert"></TD>
=09=09=09=09=09<TD WIDTH=3D"5" BGCOLOR=3D"#336699" ROWSPAN=3D"3"><IMG SRC=
=3D"http://www.comics.com/comics/dilbert/daily_dilbert/images/daily_dilbert=
_header_right.gif" WIDTH=3D"5" HEIGHT=3D"76" BORDER=3D"0" ALT=3D"Daily Dilb=
ert"></TD>
=09=09=09=09=09<TD WIDTH=3D"17" ROWSPAN=3D"3"><IMG SRC=3D"http://www.comics=
.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"17" HEIGHT=3D"10" BORDER=
=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09
=09=09=09=09<!-- ENTER DATE HERE -->
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"90" BGCOLOR=3D"#003366" HEIGHT=3D"13"><FONT FAC=
E=3D"helvetica, arial" SIZE=3D"1" COLOR=3D"#FFFFFF"><B>07/10/2002</B></FONT=
></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD WIDTH=3D"90" BGCOLOR=3D"#336699"><IMG SRC=3D"http://www.=
comics.com/comics/dilbert/daily_dilbert/images/daily_dilbert_header_mid_bot=
tom.gif" WIDTH=3D"90" HEIGHT=3D"8" BORDER=3D"0" ALT=3D"Daily Dilbert"></TD>
=09=09=09=09</TR>
=09=09=09</TABLE>
=09=09=09<!-- HEADER TABLE END -->
=09=09=09
=09=09=09<!-- WHITE LINE AND GREETING TABLE BEGIN -->
=09=09=09<TABLE WIDTH=3D"627" CELLPADDING=3D"0" CELLSPACING=3D"0" BORDER=3D=
"0">
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"1" HEIGHT=3D"2" BGCOLOR=3D"#003366"><IMG SRC=3D=
"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"1" HEI=
GHT=3D"2" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"605" HEIGHT=3D"2" BGCOLOR=3D"#FFFFFF" COLSPAN=
=3D"2"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"605" HEIGHT=3D"2" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"4" HEIGHT=3D"2" BGCOLOR=3D"#FFFFFF"><IMG SRC=3D=
"http://www.comics.com/comics/dilbert/daily_dilbert/images/white_line_right=
.gif" WIDTH=3D"4" HEIGHT=3D"2" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"17" HEIGHT=3D"48" ROWSPAN=3D"4"><IMG SRC=3D"htt=
p://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"17" HEIGHT=
```

```
=3D"48" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"1" HEIGHT=3D"37" BGCOLOR=3D"#003366" ROWSPAN=3D=
"2"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" =
WIDTH=3D"1" HEIGHT=3D"39" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"605" HEIGHT=3D"5" BGCOLOR=3D"#FFCC66" COLSPAN=
=3D"2"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"605" HEIGHT=3D"7" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"4" HEIGHT=3D"37" BACKGROUND=3D"http://www.comic=
s.com/comics/dilbert/daily_dilbert/images/ffcc66_right.gif" ROWSPAN=3D"2"><=
IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=
=3D"4" HEIGHT=3D"39" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD WIDTH=3D"10" HEIGHT=3D"32" BGCOLOR=3D"#FFCC66"><IMG SRC=
=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"10"=
 HEIGHT=3D"32" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"595" HEIGHT=3D"32" BGCOLOR=3D"#FFCC66"><FONT FA=
CE=3D"Arial,Verdana" SIZE=3D"1">Hi Justin, enjoy your daily comic from <A t=
arget=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3D02-qit=
9IljN-Yow111_XDuuQWJBb9RR">Dilbert.com</A>.<BR><B>If you like the Daily Dil=
bert, tell a friend! <A target=3D"_blank" HREF=3D"http://ummail4.unitedmedi=
a.com:80/Click?q=3D17-agRtI5XXfb4JSGFIzysGCUNA7RRR">Click here to send mail=
</A></B></FONT></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD WIDTH=3D"1" BGCOLOR=3D"#003366"><IMG SRC=3D"http://www.c=
omics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"1" HEIGHT=3D"7" BOR=
DER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"605" BGCOLOR=3D"#FFFFFF" COLSPAN=3D"2"><IMG SRC=
=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"605=
" HEIGHT=3D"7" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"4" HEIGHT=3D"7" BACKGROUND=3D"http://www.comics=
.com/comics/dilbert/daily_dilbert/images/ffffff_right.gif"><IMG SRC=3D"http=
://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"4" BORDER=
=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09</TABLE>
=09=09=09<!-- WHITE LINE AND GREETING TABLE END -->
=09=09=09
=09=09=09<!-- ONLINE STORE AND STRIP TABLE BEGIN -->
=09=09=09<TABLE WIDTH=3D"627" CELLPADDING=3D"0" CELLSPACING=3D"0" BORDER=3D=
"0">
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"1" HEIGHT=3D"348" BGCOLOR=3D"#003366" ROWSPAN=
=3D"5"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"1" HEIGHT=3D"348" BORDER=3D"0" ALT=3D""></TD>
```

=09=09=09=09=09<TD WIDTH=3D"4" HEIGHT=3D"348" BGCOLOR=3D"#FFFFFF" ROWSPAN=
=3D"5"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"4" HEIGHT=3D"348" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"6" HEIGHT=3D"83" BGCOLOR=3D"#FFFFFF" ROWSPAN=3D=
"4"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" =
WIDTH=3D"6" HEIGHT=3D"83" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"585" HEIGHT=3D"19" BGCOLOR=3D"#FFFFFF" COLSPAN=
=3D"7"><IMG SRC=3D"http://www.comics.com/comics/dilbert/daily_dilbert/image=
s/header_online_store.gif" WIDTH=3D"200" HEIGHT=3D"15" BORDER=3D"0" ALT=3D"=
">
<IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif=
" WIDTH=3D"4" HEIGHT=3D"4" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"9" HEIGHT=3D"83" BGCOLOR=3D"#FFFFFF" ROWSPAN=3D=
"4"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" =
WIDTH=3D"9" HEIGHT=3D"83" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"1" HEIGHT=3D"348" BGCOLOR=3D"#FFFFFF" ROWSPAN=
=3D"5"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"1" HEIGHT=3D"348" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"4" HEIGHT=3D"348" BACKGROUND=3D"http://www.comi=
cs.com/comics/dilbert/daily_dilbert/images/white_line_right.gif" ROWSPAN=3D=
"5"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" =
WIDTH=3D"4" HEIGHT=3D"348" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"17" HEIGHT=3D"349" ROWSPAN=3D"6"><IMG SRC=3D"ht=
tp://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"17" HEIGH=
T=3D"349" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD WIDTH=3D"200" HEIGHT=3D"59" BGCOLOR=3D"#FFFFFF" ROWSPAN=
=3D"2"><FONT FACE=3D"Arial,Verdana" SIZE=3D"1"><IMG SRC=3D"http://www.comic=
s.com/comics/dilbert/daily_dilbert/images/books.gif" WIDTH=3D"35" HEIGHT=3D=
"15" BORDER=3D"0" ALT=3D"Books">
<A target=3D"_blank" HREF=3D"http://umm=
ail4.unitedmedia.com:80/Click?q=3D57-YJKXII0tFp0ybw-F0u_u9RX_z9RR"><IMG SRC=
=3D"http://www.comics.com/comics/dilbert/ads/book_badge.gif" WIDTH=3D"169" =
HEIGHT=3D"60" BORDER=3D"0" ALT=3D"Another Day in Cubicle Paradise"></A>
=

Complete your collection while pretending to work!
<A target=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3D81=
-z1KaQ-KLrRxtx28pq_Rxt9CJZ9RR">> Shop</A></FONT></TD>
=09=09=09=09=09<TD WIDTH=3D"10" HEIGHT=3D"59" BGCOLOR=3D"#FFFFFF" ROWSPAN=
=3D"2"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"10" HEIGHT=3D"7" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"83" HEIGHT=3D"59" BGCOLOR=3D"#FFFFFF" ROWSPAN=
=3D"2"><A target=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click=
?q=3Dab-ZcXKQYp0Y4QKmwdXyRiFnw8ZwdRR"><IMG SRC=3D"http://www.comics.com/com=
ics/dilbert/daily_dilbert/images/thumb_dilbert_mints.gif" WIDTH=3D"83" HEIG=
HT=3D"57" BORDER=3D"0" ALT=3D"Dilbert Mints"></A></TD>
=09=09=09=09=09<TD WIDTH=3D"120" HEIGHT=3D"59" BGCOLOR=3D"#FFFFFF" ROWSPAN=
=3D"2"><FONT FACE=3D"Arial,Verdana" SIZE=3D"1"><IMG SRC=3D"http://www.comic=
s.com/comics/dilbert/daily_dilbert/images/mints.gif" WIDTH=3D"35" HEIGHT=3D=
"15" BORDER=3D"0" ALT=3D"Graphic (live connection required)">
<IMG SRC=

=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"120=
" HEIGHT=3D"1" BORDER=3D"0" ALT=3D"">
They're character-shaped, in colle=
ctible tins.
<B>> <A target=3D"_blank" HREF=3D"http://ummail4.unitedm=
edia.com:80/Click?q=3Dd5-p-SiQzL3ad2C6E2HCF4Y-4m5V9RR">Shop</A></B></FONT><=
/TD>
=09=09=09=09=09<TD WIDTH=3D"10" HEIGHT=3D"59" BGCOLOR=3D"#FFFFFF" ROWSPAN=
=3D"2"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"10" HEIGHT=3D"10" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"162" HEIGHT=3D"24" BGCOLOR=3D"#FFFFFF" COLSPAN=
=3D"2"><FONT FACE=3D"Arial,Verdana" SIZE=3D"1"><IMG SRC=3D"http://www.comic=
s.com/comics/dilbert/daily_dilbert/images/clothing_and_posters.gif" WIDTH=
=3D"121" HEIGHT=3D"15" BORDER=3D"0" ALT=3D"">
Get any Dilbert strip publ=
ished in the last 90 days printed on a
T-shirt, sweatshirt or poster. Se=
lected older strips are also available.
<B>> <A target=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Clic=
k?q=3Dff-sVHVQmste4RDbCdZxPhzoVUk5RRR">Shop</A></B>
<IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif" WIDT=
H=3D"148" HEIGHT=3D"20" BORDER=3D"0" ALT=3D""></FONT></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09<TD WIDTH=3D"14" HEIGHT=3D"35" BGCOLOR=3D"#FFFFFF"><IMG SRC=3D"=
http://www.comics.com/comics/dilbert/daily_dilbert/images/bullet.gif" WIDTH=
=3D"14" HEIGHT=3D"11" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"148" HEIGHT=3D"35" BGCOLOR=3D"#FFFFFF"><FONT FA=
CE=3D"Arial,Verdana" SIZE=3D"1"><A target=3D"_blank" HREF=3D"http://ummail4=
.unitedmedia.com:80/Click?q=3D14-5MiAIF-KaBxNkZ48ZgQpIXIl7dRR">View all pro=
ducts</A></FONT></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD WIDTH=3D"585" HEIGHT=3D"5" BGCOLOR=3D"#FFFFFF" COLSPAN=
=3D"7"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"585" HEIGHT=3D"5" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09<!-- COMIC STRIP HERE -->
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD WIDTH=3D"600" HEIGHT=3D"228" BGCOLOR=3D"#FFFFFF" ALIGN=
=3D"RIGHT" COLSPAN=3D"10"><FONT FACE=3D"helvetica, arial" SIZE=3D"1"><A tar=
get=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3D29-SgTrI=
jM2Hozdm5nNyDEptl6PrdRR"><IMG SRC=3D"http://www.comics.com/comics/dilbert/d=
aily_dilbert/images/bullet.gif" WIDTH=3D"14" HEIGHT=3D"11" ALT=3D"" BORDER=
=3D"0">Buy This Strip</A>   <A HREF=3D"http://www.comics.com=
/webmail/SendAStrip?AppName=3DSendAStrip&ComicName=3D/comics/dilbert/&Attac=
hments=3D/comics/dilbert/archive/images/dilbert2003482820710.gif&EmailDate=
=3DJuly-10-2002"><IMG SRC=3D"http://www.comics.com/comics/dilbert/daily_dil=
bert/images/bullet.gif" WIDTH=3D"14" HEIGHT=3D"11" ALT=3D"" BORDER=3D"0">Se=
nd This Strip to a Friend</A>   <A target=3D"_blank" HREF=3D=
"http://ummail4.unitedmedia.com:80/Click?q=3D53-n8CdINnRpyDM-psjH-xb_IfwI9R=
R"><IMG SRC=3D"http://www.comics.com/comics/dilbert/daily_dilbert/images/bu=

```
llet.gif" WIDTH=3D"14" HEIGHT=3D"11" ALT=3D"" BORDER=3D"0">Month of Dilbert=
</A>  </FONT><BR><IMG SRC=3D"http://www.comics.com/comics/dilbert=
/images/clear_dot.gif" WIDTH=3D"600" HEIGHT=3D"2" BORDER=3D"0" ALT=3D""><BR=
><A target=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3D7=
d-FFBqIDLUd2DsXI4UytqS-9z3UdRR"><IMG SRC=3D"http://www.comics.com/comics/di=
lbert/archive/images/dilbert2003482820710.gif" BORDER=3D"0" ALT=3D"You need=
 to go online to see today's strip and other cool features on the Daily Dil=
bert. "></A><BR><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/cle=
ar_dot.gif" WIDTH=3D"600" HEIGHT=3D"8" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09</TABLE>
=09=09=09<!-- ONLINE STORE AND STRIP TABLE END -->
=09=09=09
=09=09=09<!-- FOOTER TABLE BEGIN -->
=09=09=09<TABLE WIDTH=3D"610" CELLPADDING=3D"0" CELLSPACING=3D"0" BORDER=3D=
"0">
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"610" HEIGHT=3D"1" BGCOLOR=3D"#003366" COLSPAN=
=3D"8"><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gi=
f" WIDTH=3D"610" HEIGHT=3D"1" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"1" BGCOLOR=3D"#003366"><IMG SRC=3D"http://www.c=
omics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"1" HEIGHT=3D"4" BOR=
DER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"10" BGCOLOR=3D"#FFFFFF"><IMG SRC=3D"http://www.=
comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"10" HEIGHT=3D"4" B=
ORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"478" BGCOLOR=3D"#FFFFFF"><IMG SRC=3D"http://www=
.comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"478" HEIGHT=3D"4"=
 BORDER=3D"0" ALT=3D""><BR><FONT FACE=3D"helvetica, arial" SIZE=3D"1"><A ta=
rget=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3Da8-02QZ=
QQZO8FsQPNXDtcsrwMJWwsRR">Unsubscribe</A> | <A target=3D"_blank" HREF=3D"ht=
tp://ummail4.unitedmedia.com:80/Click?q=3Dd2-THH_QULylbSNFPV3Ge2S5G1mkRRR">=
Modify Your Subscription</A> | <A target=3D"_blank" HREF=3D"http://ummail4.=
unitedmedia.com:80/Click?q=3Dfc-WpomQVw_4c4WGxZa9p0FA0wF5dRR">Request a New=
 Subscription</A> | <a href=3D"mailto:dilberthelp@unitedmedia.com">Report S=
ubscription Problems</A><BR>
<A target=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3D26=
-KqZVIiMmfPuWOSdBBNk-Z1dXjsRR">Subscribe to the Dilbert Newsletter/Join Dog=
bert's New Ruling Class</A> |<BR>
<A target=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3D50=
-M_6BIR0rsFUiFr_AS5nxjtaOIRRR">Dilbert on Your Site/Intranet</A> | <A targe=
t=3D"_blank" HREF=3D"http://ummail4.unitedmedia.com:80/Click?q=3D7a-EETAIPB=
-Be1flYzhU0ckfggJpsRR">Advertising Info</A> | <A target=3D"_blank" HREF=3D"=
http://ummail4.unitedmedia.com:80/Click?q=3Da4-HG_sQ6p9oUtGwrK46HL35O3CssRR=
">Dilbert in Your Publication</A></FONT></TD>
=09=09=09=09=09<TD WIDTH=3D"45" BGCOLOR=3D"#FFFFFF"><IMG SRC=3D"http://www.=
```

```
comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"45" HEIGHT=3D"4" B=
ORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"62" BGCOLOR=3D"#FFFFFF"><IMG SRC=3D"http://www.=
comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"62" HEIGHT=3D"4" B=
ORDER=3D"0" ALT=3D""><BR><A target=3D"_blank" HREF=3D"http://ummail4.united=
media.com:80/Click?q=3Dce-SdMCQMsow2JPkSs3u_zGKPsvo9RR"><IMG SRC=3D"http://=
www.comics.com/comics/dilbert/daily_dilbert/images/comics_com_logo.gif" WID=
TH=3D"62" HEIGHT=3D"42" BORDER=3D"0" ALT=3D"Comics.com"></A></TD>
=09=09=09=09=09<TD WIDTH=3D"10" BGCOLOR=3D"#FFFFFF"><IMG SRC=3D"http://www.=
comics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"10" HEIGHT=3D"4" B=
ORDER=3D"0" ALT=3D""></TD>
=09=09=09=09=09<TD WIDTH=3D"4" BACKGROUND=3D"http://www.comics.com/comics/d=
ilbert/daily_dilbert/images/white_line_right.gif"><IMG SRC=3D"http://www.co=
mics.com/comics/dilbert/images/clear_dot.gif" WIDTH=3D"4" HEIGHT=3D"4" BORD=
ER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">=09
=09=09=09=09=09<TD WIDTH=3D"610" COLSPAN=3D"8"><IMG SRC=3D"http://www.comic=
s.com/comics/dilbert/daily_dilbert/images/daily_dilbert_footer.gif" WIDTH=
=3D"610" HEIGHT=3D"11" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09</TABLE>
=09=09=09<!-- FOOTER TABLE END -->
=09=09=09
=09=09</TD>
=09=09<TD>
=09=09
=09=09=09
=09=09
=09=09=09<!-- SKYSCRAPER TABLE BEGIN -->
=09=09=09<table border=3D"0" cellspacing=3D"0" cellpadding=3D"0" height=3D"=
600">
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/=
small_ad.gif" WIDTH=3D"20" HEIGHT=3D"11" BORDER=3D"0" ALT=3D"Advertisement"=
><BR><IMG SRC=3D"http://www.comics.com/comics/dilbert/images/clear_dot.gif"=
 WIDTH=3D"2" HEIGHT=3D"2" BORDER=3D"0" ALT=3D""></TD>
=09=09=09=09</TR>
=09=09=09=09<TR VALIGN=3D"TOP">
=09=09=09=09=09<TD><a href=3D"http://www.partner2profit.com/redir.cfm?ccode=
=3D1D74C2EA&pcode=3DEF39755C" target=3D"_blank"><img src=3D"http://www.flow=
go.com/images/bt/p2p.gif?ccode=3D1D74C2EA&pcode=3DEF39755C" width=3D120 hei=
ght=3D600 border=3D0></a></TD>
=09=09=09=09</TR>
=09=09=09</table>
=09=09=09<!-- SKYSCRAPER TABLE END -->
=09=09=09
=09=09=09
```

```
=09=09=09
=09=09</TD>
=09</TR>
</TABLE>
<!-- CONTENT AND SKYSCRAPER AD TABLE END -->



<script language=3D"JavaScript">
<!--
document.writeln ('<layer SRC=3D"http://ad.doubleclick.net/adl/dilbert.emai=
l.com/email/;sz=3D468x60;ord=3D' + same + '?" visibility=3D"hidden" width=
=3D468 height=3D60 onload=3D"moveToAbsolute(ph1.pageX,ph1.pageY);clip.width=
=3D468;clip.height=3D60;visibility=3D\'show\';"></layer>');
//-->
</script>

</TD>
</TR>
</TABLE>



</BODY>
</HTML>


<IMG HEIGHT=3D1 WIDTH=3D1 SRC=3D"http://ummail4.unitedmedia.com:80/Click?q=
=3D0d-3a2unDr3RbASR2VR56fIrkpY">
--23561619.1026296118170.JavaMail.root.umsan1--



############################ No.3 email################################
From fo0kozh4k44705211@hotmail.com  Thu Sep 19 11:15:10 2002
Return-Path: <fo0kozh4k44705211@hotmail.com>
Delivered-To: zzzz@localhost.jmason.org
Received: from localhost (jalapeno [127.0.0.1])
        by zzzzason.org (Postfix) with ESMTP id 4745D16F03
        for <zzzz@localhost>; Thu, 19 Sep 2002 11:15:09 +0100 (IST)
Received: from jalapeno [127.0.0.1]
        by localhost with IMAP (fetchmail-5.9.0)
        for zzzz@localhost (single-drop); Thu, 19 Sep 2002 11:15:09 +0100 (IST)
Received: from webnote.net (mail.webnote.net [193.120.211.219]) by
    dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g8J4OrC04796 for
    <zzzz@jmason.org>; Thu, 19 Sep 2002 05:24:53 +0100
Received: from mail.kti.nsc.ru (IDENT:root@ns.kti.nsc.ru [194.226.170.3])
    by webnote.net (8.9.3/8.9.3) with ESMTP id FAA00636 for
```

```
    <zzzz@example.com>; Thu, 19 Sep 2002 05:25:23 +0100
Received: from unspecified.host ([194.226.170.51]) by mail.kti.nsc.ru
    (8.12.5/8.12.5) with ESMTP id g8J2Utmi031901; Thu, 19 Sep 2002 11:23:52
    +0700 (NOVST)
Received: from 172.190.77.230 ([172.190.77.230]) by 127.0.0.1 (WinRoute
    Pro 4.2.2) with SMTP; Tue, 17 Sep 2002 15:44:14 +0700
Message-Id: <00006b1a2e12$00000c87$00006191@mx06.hotmail.com>
To: <dabadboy1@yahoo.com>
Cc: <toufrom@aol.com>, <moflava4us@aol.com>, <pabbott793@aol.com>,
        <whitepln88@aol.com>, <zzzz@example.com>, <sandrab107@aol.com>,
        <reeb@simrax.com>
From: "Mackenzie" <fo0kozh4k44705211@hotmail.com>
Subject: Toners and inkjet cartridges for less…          BSJMCOIK
Date: Tue, 17 Sep 2002 01:46:06 -1900
MIME-Version: 1.0
Reply-To: fo0kozh4k44705211@hotmail.com
Content-Type: text/html; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

<HR>
<html>
<div bgcolor=3D"#FFFFCC">

  <p align=3D"center"><a
href=3D"http://www.webbasedmailing.com"><img border=3D"0"
src=3D"http://www.webbasedmailing.com/Toners2goLogo.jpg"
width=3D"349" height=3D"96"></a></p>
<p align=3D"center"><font size=3D"6" face=3D"Arial MT
Black"><i>Tremendous Savings</i>
on Toners, </font></p>
<p align=3D"center"><font size=3D"6" face=3D"Arial MT
Black">
Inkjets, FAX, and Thermal Replenishables!!</font></p>
<p><a href=3D"http://www.webbasedmailing.com">Toners 2 Go
</a>is your secret
weapon to lowering your cost for <a
href=3D"http://www.webbasedmailing.com">High Quality,
Low-Cost</a> printer
supplies!  We have been in the printer
replenishables business since 1992,
and pride ourselves on rapid response and outstanding
customer service. 
What we sell are 100% compatible replacements for
Epson, Canon, Hewlett Packard,
Xerox, Okidata, Brother, and Lexmark; products that
meet and often exceed
original manufacturer's specifications.</p>
<p><i><font size=3D"4">Check out these
```

```
prices!</font></i></p>
<p>        Epson Stylus
Color inkjet cartridge
(SO20108):     Epson's Price:
$27.99  <font face=3D"Arial MT Black">  
Toners2Go price: $9.95!</font></p>
  <p>   <font face=3D"Arial MT
Black">    </font>HP
LaserJet 4 Toner Cartridge
(92298A):            =
;
HP's
Price:
$88.99           
<font face=3D"Arial MT Black">Toners2Go
  price: $41.75!</font></p>
<p align=3D"center"> </p>
<p>Come visit us on the web to check out our hundreds
of similar bargains at <a
href=3D"http://www.webbasedmailing.com">Toners
2 Go</a>!
</p>

<hr>

  request to be excluded by visiting <a
href=3D"http://www.interactive-mailing.com/datacenter.htm">HERE</a></span>=
</font><br>

</div>
</html>
michaell
```

1.2 Note that the email files contain a lot of extra information, besides the actual message. Ignore that for now and run on the entire text (in the optional part further down can experiment with filtering out the headers and footers). We don't want to train and test on the same data (it might help to reflect on why if you don't recall). Split the spam and the ham datasets in a training set and a test set. (`hamtrain`, `spamtrain`, `hamtest`, and `spamtest`). Use only the easy_ham part as ham data for quesions 1 and 2. Answer 1.1:

[ ]:
```python
from sklearn.model_selection import train_test_split

def read_files(file_list):
    # files = os.listdir(file_fold)
    messages = []
    for file in file_list:
        with open(file, 'r', encoding='utf-8', errors='ignore') as f:
            content = f.read()
            messages.append((content, "ham" if "ham" in file else "spam"))

    return messages


easy_ham_messages = read_files(easy_ham_files)
hard_ham_messages = read_files(hard_ham_files)
spam_messages = read_files(spam_files)

# Split the ham and spam files into training and testing datasets
hamtrain, hamtest = train_test_split(easy_ham_messages, test_size=0.2,␣
 ↪random_state=42)
spamtrain, spamtest = train_test_split(spam_messages, test_size=0.2,␣
 ↪random_state=42)
ham_hardtrain, ham_hardtest = train_test_split(hard_ham_messages, test_size=0.
 ↪2, random_state=42)

# Print the number of the traning/testing datasets
print("Size of hamtrain:", len(hamtrain))
print("Size of hamtest:", len(hamtest))
print("Size of spamtrain:", len(spamtrain))
print("Size of spamtest:", len(spamtest))
```

```
Size of hamtrain: 2040
Size of hamtest: 511
Size of spamtrain: 400
Size of spamtest: 100
```

1.2.2 2.1 Write a Python program that:

1. Uses the four datasets from Question 1 (`hamtrain`, `spamtrain`, `hamtest`, and `spamtest`)
2. Trains a Naïve Bayes classifier (use the scikit-learn library) on `hamtrain` and `spamtrain`, that classifies the test sets and reports True Positive and False Negative rates on the `hamtest` and `spamtest` datasets. Use `CountVectorizer` (Documentation here) to transform the email texts into vectors. Please note that there are different types of Naïve Bayes Classifier in scikit-learn (Documentation here). Test two of these classifiers that are well suited for this problem:

- Multinomial Naive Bayes

- Bernoulli Naive Bayes.

Please inspect the documentation to ensure input to the classifiers is appropriate before you start coding.

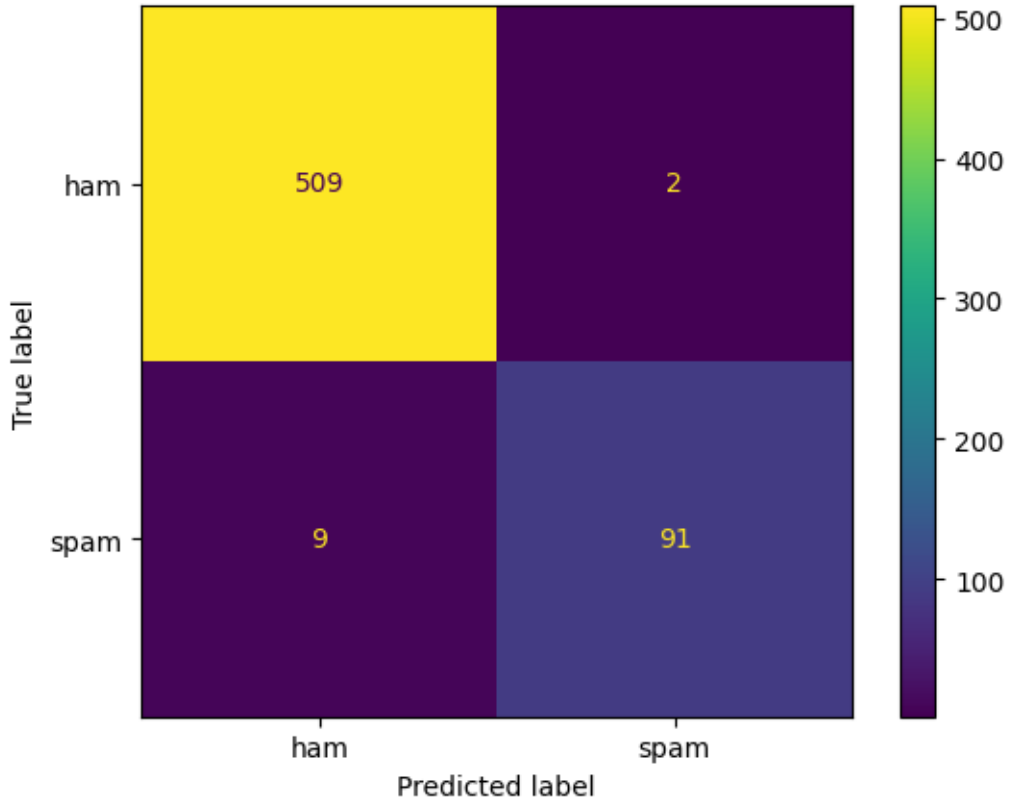
[ ]:
```
from sklearn.feature_extraction.text import CountVectorizer

X_train, y_train = zip(*(hamtrain+spamtrain))
X_test, y_test = zip(*(hamtest+spamtest))

# Create a Vectorizer Object
vectorizer = CountVectorizer()

# convert text to numerical data
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)

# Printing the identified Unique words along with their indices
# print("Vocabulary: ", vectorizer.vocabulary_)
```

2.1.1 Multinomial Naive Bayes

[ ]:
```
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt

plt.rcParams['figure.dpi'] = 100

# create Multinomial Naive Bayes model object and fit it
mnb = MultinomialNB()
mnb.fit(X_train, y_train)

# predit X_test dataset
y_pred = mnb.predict(X_test)

# reports True Positive and False Negative rates on the `hamtest` and␣
 ↪`spamtest` datasets.
tn1, fp1, fn1, tp1 = confusion_matrix(y_test, y_pred).ravel()
print("True Positive Rate:", tp1 / (tp1 + fn1))
print("False Negative Rate:", fn1 / (tp1 + fn1))

# accuracy score
accuracy_1 = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy_1:.3f}")

# plot the comfusionMatrix
cm = confusion_matrix(y_test, y_pred, labels=mnb.classes_)
```

```
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=mnb.classes_)
disp.plot()

plt.show()
```

```
True Positive Rate: 0.91
False Negative Rate: 0.09
Accuracy: 0.982
```



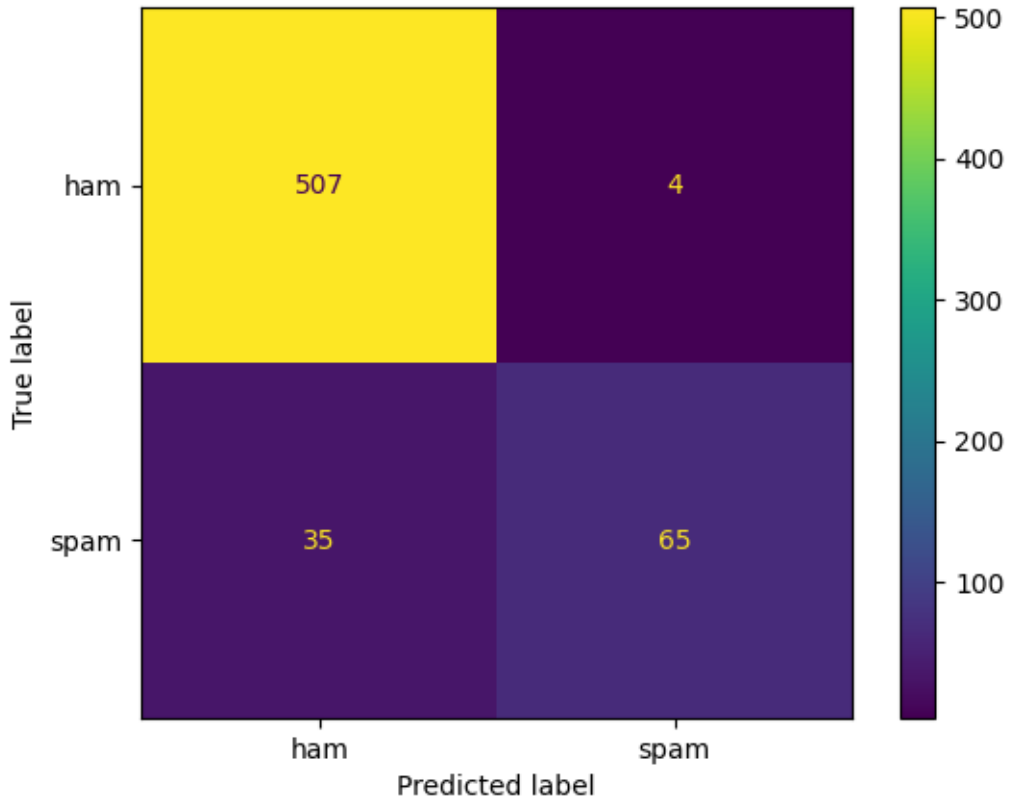
2.1.2 Bernoulli Naive Bayes

[ ]:
```
from sklearn.naive_bayes import BernoulliNB

# create Bernoulli Naive Bayes model object and fit it
bnb = BernoulliNB(force_alpha=True, binarize=0.0)
bnb.fit(X_train, y_train)

# predit X_test dataset
y_pred = bnb.predict(X_test)

# reports True Positive and False Negative rates on the `hamtest` and␣
↪`spamtest` datasets.
```

```
tn2, fp2, fn2, tp2 = confusion_matrix(y_test, y_pred).ravel()
print("True Positive Rate:", tp2 / (tp2 + fn2))
print("False Negative Rate:", fn2 / (tp2 + fn2))

# accuracy score
accuracy_2 = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy_2:.3f}")

# plot the comfusionMatrix
cm = confusion_matrix(y_test, y_pred, labels=bnb.classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=bnb.classes_)
disp.plot()
```

```
True Positive Rate: 0.65
False Negative Rate: 0.35
Accuracy: 0.936
```

[ ]: `<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1630fb3b2d0>`



1.2.3 2.2 Answer the following questions:

a) What does the CountVectorizer do? Answer 2.2.a

Countvectorizer is used to convert text to a numerical matrix, in which each row represents a document and each column represents a unique word in the vocabulary. The values in the matrix represent the frequency of the corresponding word.
Usually, we use it for preprocessing the text dataset for machine learning algorithms that require numerical input data.

b) What is the difference between Multinomial Naive Bayes and Bernoulli Naive Bayes
Answer 2.2.b - In Multinomial Naive Bayes, the converted matrix has discrete values, however in Bernoulli Naive Bayes model, the discrete values would be converted to 0 and 1 based on threshold set. - Multinomial Naive Bayes works well with frequency-based features, such as frequency of a word in a document, Bernoulli Naive Bayes works well with binary features, such as whether a word appears in a document or not. - In this case, after comparing the performance of two models, Multinomial Naive Bayes works better than Bernoulli Naive Bayes.

1.2.4 3.1 Run the two models:

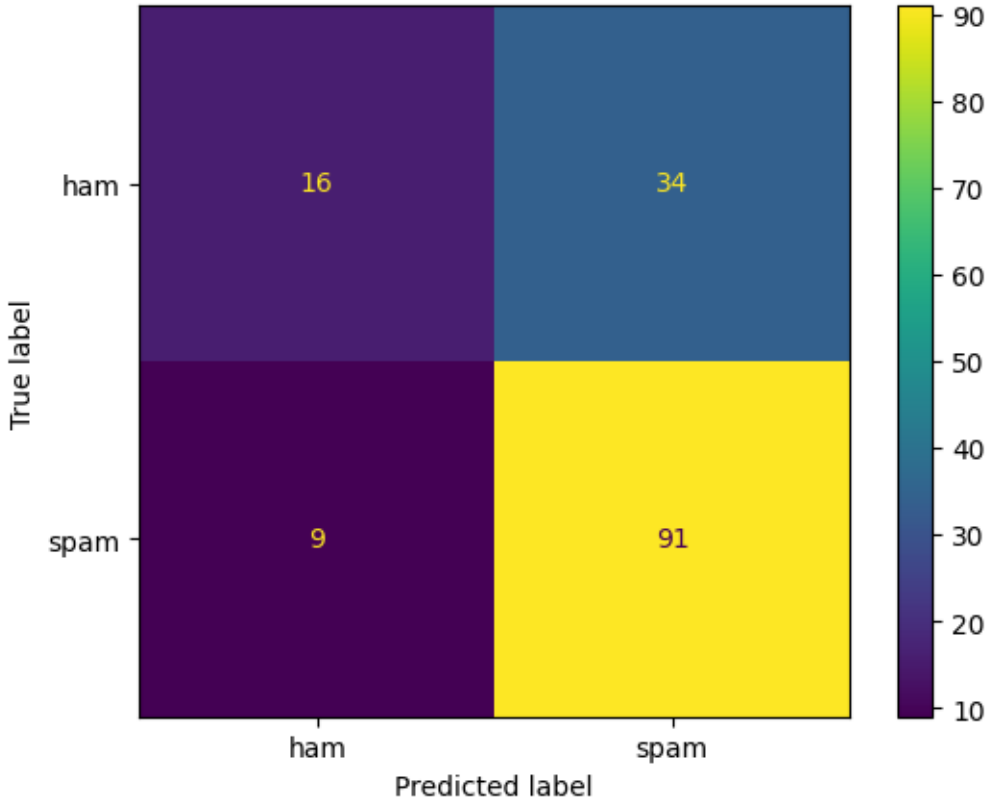
Run (don't retrain) the two models from Question 2 on spam versus hard-ham. Does the performance differ compared to question 2 when the model was run on spam versus easy-ham? If so, why?

3.1.1 Multinomial Naive Bayes (on spam versus hard-ham, don't retrain)

[ ]:
```
# reconstruct the testing dataset, using ham_hardtest+spamtest
X_test_hard = vectorizer.transform([message[0] for message in␣
 ↪ham_hardtest+spamtest])
y_test_hard = [message[1] for message in ham_hardtest+spamtest]

# predit X_test_hard dataset
y_pred_hard = mnb.predict(X_test_hard)

# reports True Positive and False Negative rates on the `hamtest` and␣
 ↪`spamtest` datasets.
tn3, fp3, fn3, tp3 = confusion_matrix(y_test_hard, y_pred_hard).ravel()
print("True Positive Rate:", tp3 / (tp3 + fn3))
print("False Negative Rate:", fn3 / (tp3 + fn3))

# accuracy score
accuracy_3 = accuracy_score(y_test_hard, y_pred_hard)
print(f"Accuracy: {accuracy_3:.3f}")

# plot the comfusionMatrix
cm = confusion_matrix(y_test_hard, y_pred_hard, labels=mnb.classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=mnb.classes_)
disp.plot()

plt.show()
```

```
True Positive Rate: 0.91
False Negative Rate: 0.09
```

```
Accuracy: 0.713
```

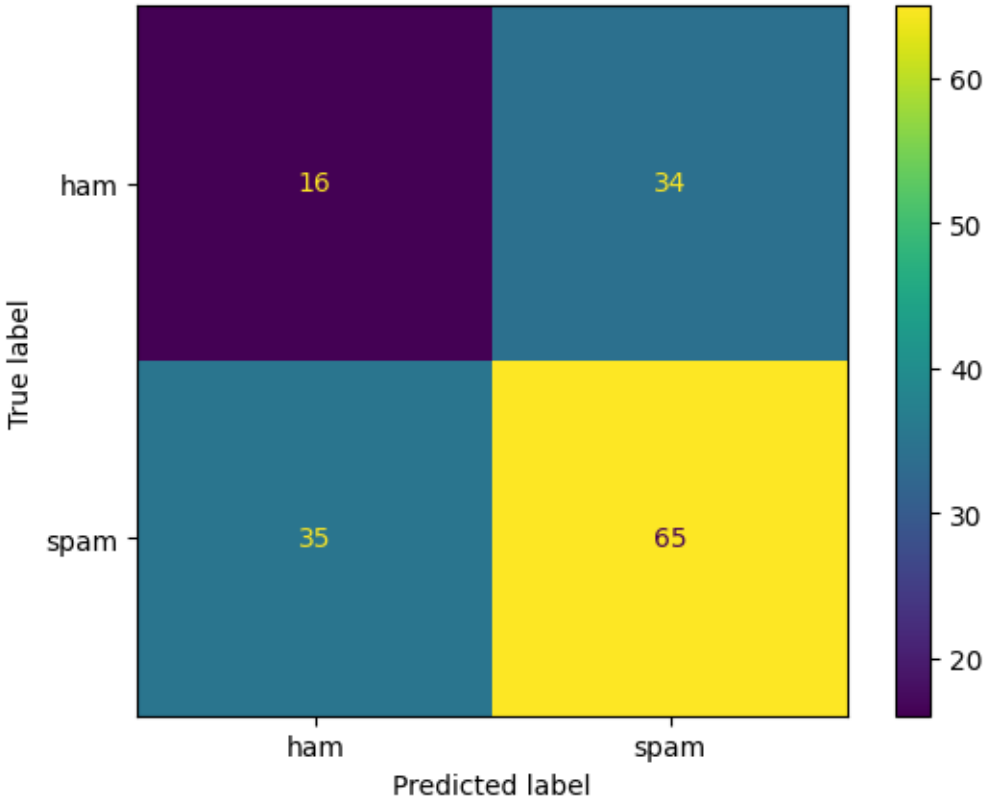


3.1.2 Bernoulli Naive Bayes (on spam versus hard-ham, don't retrain)

[ ]:
```
# predit X_test_hard dataset
y_pred_hard = bnb.predict(X_test_hard)

# reports True Positive and False Negative rates on the `hamtest` and␣
↪`spamtest` datasets.
tn4, fp4, fn4, tp4 = confusion_matrix(y_test_hard, y_pred_hard).ravel()
print("True Positive Rate:", tp4 / (tp4 + fn4))
print("False Negative Rate:", fn4 / (tp4 + fn4))

# accuracy score
accuracy_4 = accuracy_score(y_test_hard, y_pred_hard)
print(f"Accuracy: {accuracy_4:.3f}")

# plot the comfusionMatrix
cm = confusion_matrix(y_test_hard, y_pred_hard, labels=bnb.classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=bnb.classes_)
disp.plot()
```

```
plt.show()
```

```
True Positive Rate: 0.65
False Negative Rate: 0.35
Accuracy: 0.540
```



Answer 3.1:

When both models were run on spam versus hard-ham, their performance was worse than that achieved in question 2 when the model was run on spam versus easy-ham.

- For Multinomial Naive Bayes, the accuracy decreased from 0.982 to 0.713.
- For Bernoulli Naive Bayes, the accuracy decreased from 0.936 to 0.540.

This is because the hard-ham dataset was not included in the model's training process, this dataset contains some 'new information' that the model did not know, thus the model's performance dropped a lot on prediting hard-ham dateset.

In both models, the true positive rate and false negative rate remained unchanged, as the model itself was not altered and the spam testing dataset remained the same. As a result, we would expect to see same results between the two models.

1.2.5 3.2 Retrain

Retrain new Multinomial and Bernolli Naive Bayes classifers on the combined (easy+hard) ham and spam. Now evaluate on spam versus hard-ham as in 3.1. Also evaluate on spam versus easy-ham. Compare the performance with question 2 and 3.1. What do you observe?

[ ]:
```
# Create a Vectorizer Object
vectorizer2 = CountVectorizer()

# convert text to numerical data
X_train = vectorizer2.fit_transform([message[0] for message in␣
 ↪hamtrain+ham_hardtrain+spamtrain])
y_train = [message[1] for message in hamtrain+ham_hardtrain+spamtrain]

# convert spam versus hard-ham dataset as in question 3.1.
X_test_hard_spam = vectorizer2.transform([message[0] for message in␣
 ↪ham_hardtest+spamtest])
y_test_hard_spam = [message[1] for message in ham_hardtest+spamtest]

# convert spam versus easy-ham dataset as in question 2.
X_test_easy_spam = vectorizer2.transform([message[0] for message in␣
 ↪hamtest+spamtest])
y_test_easy_spam = [message[1] for message in hamtest+spamtest]
```

3.2.1 Multinomial Naive Bayes (retrain)

[ ]:
```
# create Multinomial Naive Bayes model object and fit it
mnb2 = MultinomialNB()
mnb2.fit(X_train, y_train)

print('####### Spam vs. hard-ham, Multinomial Naive Bayes retrain, comparing␣
 ↪with question 3.1 #######')
# predit spam versus hard-ham dataset comparing with question 3.1.
y_pred_hard_mnb = mnb2.predict(X_test_hard_spam)

# reports True Positive and False Negative rates
tn5, fp5, fn5, tp5 = confusion_matrix(y_test_hard_spam, y_pred_hard_mnb).ravel()
print("True Positive Rate:", tp5 / (tp5 + fn5))
print("False Negative Rate:", fn5 / (tp5 + fn5))

# accuracy score
accuracy_5 = accuracy_score(y_test_hard_spam, y_pred_hard_mnb)
print(f"Accuracy: {accuracy_5:.3f}")


print('\n####### spam vs. easy-ham, Multinomial Naive Bayes retrain, comparing␣
 ↪with question 2 #######')
# predit spam versus easy-ham dataset comparing with question 2.
```

```
y_predit_easy_mnb = mnb2.predict(X_test_easy_spam)

# reports True Positive and False Negative rates
tn6, fp6, fn6, tp6 = confusion_matrix(y_test_easy_spam, y_predit_easy_mnb).
 ↪ravel()
print("True Positive Rate:", tp6 / (tp6 + fn6))
print("False Negative Rate:", fn6 / (tp6 + fn6))

# accuracy score
accuracy_6 = accuracy_score(y_test_easy_spam, y_predit_easy_mnb)
print(f"Accuracy: {accuracy_6:.3f}")
```

```
####### Spam vs. hard-ham, Multinomial Naive Bayes retrain, comparing with
question 3.1 #######
True Positive Rate: 0.94
False Negative Rate: 0.06
Accuracy: 0.940

####### spam vs. easy-ham, Multinomial Naive Bayes retrain, comparing with
question 2 #######
True Positive Rate: 0.94
False Negative Rate: 0.06
Accuracy: 0.990
```

3.2.2 Bernoulli Naive Bayes (retrain)

[ ]:
```
# create Bernoulli Naive Bayes model object and fit it
bnb2 = BernoulliNB(force_alpha=True, binarize=0.0)
bnb2.fit(X_train, y_train)

print('####### pam vs. hard-ham, Bernoulli Naive Bayes retrain, comparing with␣
 ↪question 3.1 #######')
# predit spam versus hard-ham dataset comparing with question 3.1.
y_pred_hard_bnb = bnb2.predict(X_test_hard_spam)

# reports True Positive and False Negative rates
tn7, fp7, fn7, tp7 = confusion_matrix(y_test_hard_spam, y_pred_hard_bnb).ravel()
print("True Positive Rate:", tp7 / (tp7 + fn7))
print("False Negative Rate:", fn7 / (tp7 + fn7))

# accuracy score
accuracy_7 = accuracy_score(y_test_hard_spam, y_pred_hard_bnb)
print(f"Accuracy: {accuracy_7:.3f}")


print('\n####### spam vs. easy-ham, Bernoulli Naive Bayes retrain, comparing␣
 ↪with question 2 #######')
# predit spam versus easy-ham dataset comparing with question 2.
```

```
y_predit_easy_bnb = bnb2.predict(X_test_easy_spam)

# reports True Positive and False Negative rates
tn8, fp8, fn8, tp8 = confusion_matrix(y_test_easy_spam, y_predit_easy_bnb).
 ↪ravel()
print("True Positive Rate:", tp8 / (tp8 + fn8))
print("False Negative Rate:", fn8 / (tp8 + fn8))

# accuracy score
accuracy_8 = accuracy_score(y_test_easy_spam, y_predit_easy_bnb)
print(f"Accuracy: {accuracy_8:.3f}")
```

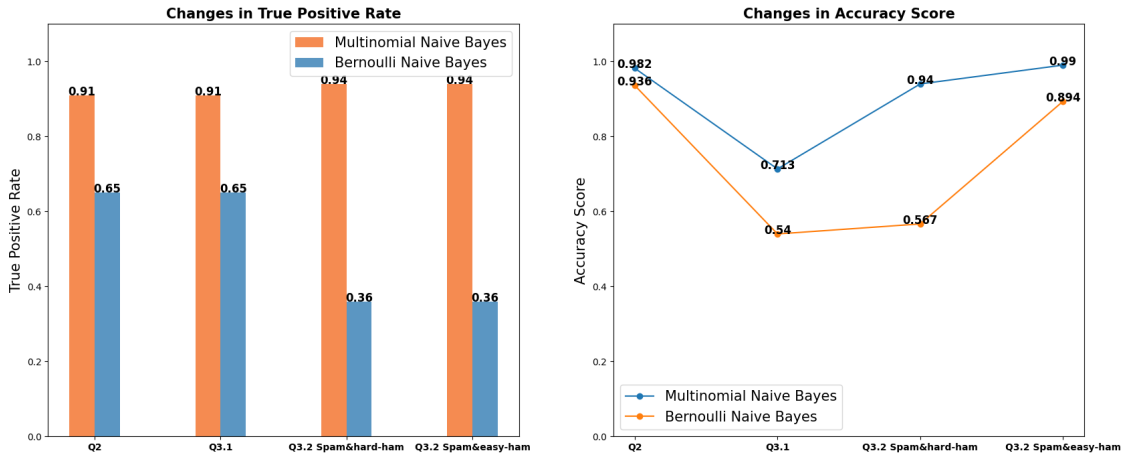
```
####### pam vs. hard-ham, Bernoulli Naive Bayes retrain, comparing with question
3.1 #######
True Positive Rate: 0.36
False Negative Rate: 0.64
Accuracy: 0.567

####### spam vs. easy-ham, Bernoulli Naive Bayes retrain, comparing with
question 2 #######
True Positive Rate: 0.36
False Negative Rate: 0.64
Accuracy: 0.894
```

Plot the varies between different scenarios

[ ]:
```
def addlabels(x,y,rows,cloumns,index):
    for j in range(len(x)):
        plt.subplot(rows,cloumns,index).text(x[j], y[j], y[j], ha =␣
 ↪'center',weight = 'bold', fontsize=12)
```

[ ]:
```
import numpy as np

multi_tp = [tp1/(tp1+fn1), tp3/(tp3+fn3), tp5/(tp5+fn5), tp6/(tp6+fn6)]
Bernou_tp = [tp2/(tp2+fn2), tp4/(tp4+fn4), tp7/(tp7+fn7), tp8/(tp8+fn8)]

multi_accuracy = [accuracy_1, accuracy_3, accuracy_5, accuracy_6]
Bernou_accuracy = [accuracy_2,accuracy_4, accuracy_7, accuracy_8]

X = ['Q2', 'Q3.1', 'Q3.2 Spam&hard-ham', 'Q3.2 Spam&easy-ham']
X_axis = np.arange(len(X))

fig, (ax1, ax2) = plt.subplots(1, 2,figsize = (20, 8))
ax1.bar(X_axis - 0.1, multi_tp, 0.2, label = 'Multinomial Naive Bayes',␣
 ↪color='#F58B51')
ax1.bar(X_axis + 0.1, Bernou_tp, 0.2, label = 'Bernoulli Naive Bayes',␣
 ↪color='#5B96C2')
ax1.set_xticks(X_axis, X, weight='bold')
```

```
ax1.set_ylabel('True Positive Rate',fontsize=15)
addlabels(X_axis - 0.1, multi_tp, 1, 2, 1)
addlabels(X_axis + 0.1, Bernou_tp, 1, 2, 1)
ax1.legend(fontsize=15)
ax1.set_ylim([0, 1.1])
ax1.set_title('Changes in True Positive Rate', fontsize=15, weight='bold')

ax2.plot(X,multi_accuracy, label='Multinomial Naive Bayes', marker = 'o')
ax2.plot(X,Bernou_accuracy, label='Bernoulli Naive Bayes', marker = 'o')
ax2.set_ylabel('Accuracy Score', fontsize=15)
ax2.legend(fontsize=15)
ax2.set_title('Changes in Accuracy Score', fontsize=15, weight='bold')
addlabels(X_axis,np.round(multi_accuracy,3), 1, 2, 2)
addlabels(X_axis,np.round(Bernou_accuracy,3), 1, 2, 2)
ax2.set_xticks(X_axis, X, weight='bold')
ax2.set_ylim([0, 1.1])

plt.show()
```



Answer 3.2:

As we can seen from the figure above:

- After combining (easy+hard) ham and spam datasets to retrain the model in question 3, the true positive rate of Multinomial Naive Bayes model increased, however, the true positive rate of Bernoulli Naive Bayes model decreased.
- When evaluating on spam versus hard-ham as in question 3.1, we observed that the new Multinomial Naive Bayes model's accuracy increased a lot from 0.713 to 0.940 comapred with the old model in question 3.1. However, the new Bernoulli Naive Bayes model's performance did not increase obviously, from 0.540 to 0.567.
- when evaluating on spam versus easy-ham as in question 2, the performance of two new models is similar to the two old ones in question 2, with new Multinomial Naive Bayes model's accuracy reached 0.990, and the new Bernoulli Naive Bayes model reached 0.894.

1.2.6 3.3 Further improvements

Do you have any suggestions for how performance could be further improved? You don't have to implement them, just present your ideas.

Answer 3.3:

- As mentioned in question 1, the email files contain a lot of nonsensical information, which lead to multiple noise columns when we convert text to numerical matrix using countvectorizer, if we can improve feature selection techniques to identify the most informative features, like filtering out the the headers and footers, it will help improve the performance.
- Some parameters can have a significant impact on the performance of the model, we can improve the model's performance by tuning these parameters, like alpha, binarize in Bernoulli Naive Bayes.
- Increase the size and quality of training dataset, this can improve the generalization performance of the model.
- Try other classification algorithms such as logistic regression, decision trees, or support vector machines (SVMs) to classify spam emails if Naive Bayes model cannot provide satisfying result.