DAT405/DIT407 Introduction to Data Science and AI

2022-2023, Reading Period 4

Assignment 4: Spam classification using Naïve Bayes

Authors: Kevin To and Filip Cederqvist

Work load: 13 h each

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.

In [1]:
```
# Download and extract data
# !wget https://spamassassin.apache.org/old/publiccorpus/20021010_easy_ham.t
# !wget https://spamassassin.apache.org/old/publiccorpus/20021010_hard_ham.t
# !wget https://spamassassin.apache.org/old/publiccorpus/20021010_spam.tar.b
# !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`.

In [2]:
```
# !ls -lah
```

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?

In [3]:
```
import os
import pandas as pd

easy_ham_files = os.listdir('./datasets/easy_ham/')
hard_ham_files = os.listdir('./datasets/hard_ham/')
spam_files = os.listdir('./datasets/spam/')

easy_ham = []
hard_ham = []
spam = []

# Putting all the mails in a list of dictionaries for easy conversion to dat
for file in easy_ham_files:
  f = open('./datasets/easy_ham/' + file,  encoding = "ISO-8859-1")
  easy_ham.append({"msg": f.read(), "label": "ham"})

for file in hard_ham_files:
  f = open('./datasets/hard_ham/' + file,  encoding = "ISO-8859-1")
  hard_ham.append({"msg": f.read(), "label": "ham"})

for file in spam_files:
  f = open('./datasets/spam/' + file,  encoding = "ISO-8859-1")
  spam.append({"msg": f.read(), "label": "spam"})

msg_easy_ham = list(easy_ham[0].values())[0]
msg_hard_ham = list(hard_ham[0].values())[0]
msg_spam = list(spam[0].values())[0]

print(msg_easy_ham)
print("-"*50 + "NEW MAIL" + "-"*50)
print(msg_hard_ham)
print("-"*50 + "NEW MAIL" + "-"*50)
print(msg_spam)
```

```
From rssfeeds@jmason.org  Mon Sep 30 13:43:46 2002
Return-Path: <rssfeeds@example.com>
Delivered-To: yyyy@localhost.example.com
Received: from localhost (jalapeno [127.0.0.1])
        by jmason.org (Postfix) with ESMTP id AE79816F16
        for <jm@localhost>; Mon, 30 Sep 2002 13:43:46 +0100 (IST)
Received: from jalapeno [127.0.0.1]
        by localhost with IMAP (fetchmail-5.9.0)
        for jm@localhost (single-drop); Mon, 30 Sep 2002 13:43:46 +0100 (IS
T)
Received: from dogma.slashnull.org (localhost [127.0.0.1]) by
    dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g8U81fg21359 for
    <jm@jmason.org>; Mon, 30 Sep 2002 09:01:41 +0100
Message-Id: <200209300801.g8U81fg21359@dogma.slashnull.org>
To: yyyy@example.com
From: gamasutra <rssfeeds@example.com>
Subject: Priceless Rubens works stolen in raid on mansion
Date: Mon, 30 Sep 2002 08:01:41 -0000
Content-Type: text/plain; encoding=utf-8
Lines: 6
X-Spam-Status: No, hits=-527.4 required=5.0
        tests=AWL,DATE_IN_PAST_03_06,T_URI_COUNT_0_1
        version=2.50-cvs
X-Spam-Level:

URL: http://www.newsisfree.com/click/-1,8381145,215/
Date: 2002-09-30T03:04:58+01:00

*Arts:* Fourth art raid on philanthropist's home once targeted by the IRA an
d
Dublin gangster Martin Cahill.



------------------------------------------------------NEW MAIL----------------
---------------------------------
Return-Path: <Online#3.19846.2a-726zgP3UI7kTO9RR.1.b@newsletter.online.com>
Received: from ABV-SFO1-ACMTA6.CNET.COM (abv-sfo1-acmta6.cnet.com [206.16.1.
169])
        by dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g6BNM8J12116
        for <qqqqqqqqqq-zdnet@example.com>; Fri, 12 Jul 2002 00:22:08 +0100
Received: from abv-sfo1-ac-agent5 (206.16.0.240) by ABV-SFO1-ACMTA6.CNET.COM
(PowerMTA(TM) v1.5); Thu, 11 Jul 2002 16:24:20 -0700 (envelope-from <Online#
3.19846.2a-726zgP3UI7kTO9RR.1.b@newsletter.online.com>)
Message-ID: <5645232.1026429716737.JavaMail.root@abv-sfo1-ac-agent5>
Date: Thu, 11 Jul 2002 16:21:56 -0700 (PDT)
From: "CNET News.com Investor" <Online#3.19846.2a-726zgP3UI7kTO9RR.1@newslet
ter.online.com>
To: qqqqqqqqqq-zdnet@example.com
Subject: NEWS.COM INVESTOR: Battered stocks regain ground
Mime-Version: 1.0
Content-Type: text/html; charset=ISO-8859-1
Content-Transfer-Encoding: 7bit
X-Mailer: Accucast (http://www.accucast.com)
X-Mailer-Version: 2.8.4-2

<html>
<head>
<title>CNET Investor Dispatch</title>
</head>
<body bgcolor="#eeeeee" link="#0000ff" vlink="#0000ff">
<div align="center">
<a href="top"></a>
<table width=612 bgcolor="#ffffff" cellpadding=0 cellspacing=0 border=0>
```

```
<tr valign=top><td width=442 colspan=4><img src="http://www.cnet.com/i/nl/in
v/hd.gif" width=442 height=60 alt="CNET Investor Dispatch"><br>
<table bgcolor="#ffcc00" width=442 cellpadding=0 cellspacing=0 border=0>
<tr><td bgcolor="#000000" colspan=3><img src="http://www.cnet.com/b.gif" wid
th=1 height=1></td></tr>
<tr><td bgcolor="#000000" width=1 rowspan=2><img src="http://www.cnet.com/b.
gif" width=1 height=10></td>
<td width=10 rowspan=2><img src="http://www.cnet.com/b.gif" width=1 height=1
0></td>
<td width=430><img src="http://www.cnet.com/b.gif" width=1 height=10></td></
tr>
<tr><td bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width=1 heigh
t=1></td></tr>
</table></td>
<td width=160 rowspan=2><img src="http://www.cnet.com/i/nl/dd_b.gif" width=1
60 height=18><br>

<!-- ad -->
<iframe src="http://www.zdnet.com/include/ads/ifc/RGROUP=2560" scrolling="n
o" frameborder="0" hspace="0" vspace="0" height="600" width="160" marginheig
ht="0" marginwidth="0">
<script language="JavaScript" src="http://www.zdnet.com/include/ads/js/RGROU
P=2560">
</script>
</iframe>
<!-- ad -->

<br>

<!-- lookup -->
<table width=160 bgcolor="#cccccc" cellpadding=1 cellspacing=0 border=0>
<tr><td><table width="100%" bgcolor="#ffffff" cellpadding=9 cellspacing=0 bo
rder=0>
<tr><td>
<form name="lookupForm" method="get" action="http://investor.cnet.com/invest
or/quotes/quote-process/0-0.html">
<font face="arial, helvetica" color="#333334" size="-1"><b>Quote Lookup</b><
br>Enter symbol:</font><br>
<input type="text" name="symbol" size=10 value="">
<input name="target" type="hidden" value="detailquote">
<input type="submit" value="Go!"><br>
&#183; <font face="ms sans serif, geneva" size="-2"><a href="http://clickthr
u.online.com/Click?q=3f-4sAEIqMW45Ukpn8Hcoh57BymNzeR" >Symbol Lookup</a><br>
Quotes delayed 20+ minutes</font>
</td></form></tr>
</table></td></tr>
</table>
<!-- /lookup -->

<p>

<!-- nav -->
<img src="http://www.cnet.com/i/nl/inv/rhc_sect.gif" width=160 height=20 alt
="Investor Sections"><br>
<font face="arial, helvetica" size="-1">
&#149; <b><a href="http://clickthru.online.com/Click?q=55-TAHaIKJ8Xqi8kriqm8
6ccJ_2d2ZR" >My Portfolio</a></b><br>
&#149; <b><a href="http://clickthru.online.com/Click?q=6a-z_30IeWiUkrRvFBzce
CGFySNP8dR" >Broker Reports</a></b><br>
&#149; <b><a href="http://clickthru.online.com/Click?q=7f-bn6KIIP67D3GxYE9IO
3JewkhBpuR" >IPOs</a></b><br>
&#149; <b><a href="http://clickthru.online.com/Click?q=94-rOVNQaI7DcMlbBqJ5e
lUjfr5JQyR" >Splits</a></b><br>
&#149; <b><a href="http://clickthru.online.com/Click?q=a9-IlbqQpUxJluCjrqz2J
```

```
QZv91zVMlR" >Messages</a></b>
</font>
<!-- /nav -->

<p>

<!-- tech winners -->
<img src="http://www.cnet.com/i/nl/inv/rhc_win.gif" width=160 height=20 alt
="Tech Winners"><br>
<table width=160 cellpadding=0 cellspacing=0 border=0>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_up.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=be-B7ixQCpTZPSgNfzbnY9y52GfwH4R" >MRAE</a
></td>
<td><font face="ms sans serif, geneva" size="-2">3.9</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#009933>27.87%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_up.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=d3-GZHkQJPai3zCH6GK74p01K7fy3rR" >LEXG</a
></td>
<td><font face="ms sans serif, geneva" size="-2">5.15</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#009933>27.16%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_up.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=e9-gdzpQxVzOsNdL8fTRnU3P83P5XlR" >BNHNA</
a></td>
<td><font face="ms sans serif, geneva" size="-2">14.95</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#009933>24.63%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_up.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=fe-xkIgQwgwyCntBcLeE4rbg6amtI4R" >AETH</a
></td>
<td><font face="ms sans serif, geneva" size="-2">3.48</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#009933>22.97%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_up.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=13-NxSZI3XWrk9CAcMssDLfzpHOCerR" >GGUY</a
></td>
<td><font face="ms sans serif, geneva" size="-2">3.1</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#009933>21.57%</font></font></td></tr>

</table>
<!-- /tech winners -->

<br>

<!-- tech losers -->
<img src="http://www.cnet.com/i/nl/inv/rhc_lose.gif" width=160 height=20 alt
="Tech Losers"><br>
```

```
<table width=160 cellpadding=0 cellspacing=0 border=0>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_down.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=28-A2wXIRd4OHYp_crEL2t1VlTK8yiR" >GDYS</a
></td>
<td><font face="ms sans serif, geneva" size="-2">6.951</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#cc0000>-32.78%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_down.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=3d-oi9NIbz4iUtIaYRvwlGWUTnaNtPR" >APHT</a
></td>
<td><font face="ms sans serif, geneva" size="-2">4.07</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#cc0000>-26.67%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_down.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=52-LUXoIrJtzdNpUwFMWz8AtIQwjjnR" >AUTN</a
></td>
<td><font face="ms sans serif, geneva" size="-2">2.3</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#cc0000>-26.52%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_down.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=67-PnDkIOEDCi6swAuWhKPAKa1CPIuR" >MHCO</a
></td>
<td><font face="ms sans serif, geneva" size="-2">2.74</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#cc0000>-21.71%</font></font></td></tr>

<tr><td><img src="http://home.cnet.com/Images/Investor/arrow_down.gif" width
="7" height="9" alt="" border="0" /> </td>
<td nowrap height=20><font face="ms sans serif, geneva" size="-2"><a href="h
ttp://clickthru.online.com/Click?q=7c-7NCZIZWc-BjephAfLHDx1984BOsR" >PMFG</a
></td>
<td><font face="ms sans serif, geneva" size="-2">9.5</td>
<td nowrap align=right><font face="ms sans serif, geneva" size="-2"><font co
lor=#cc0000>-21.22%</font></font></td></tr>

</table>
<!-- /tech losers -->

<p>
<!-- ad -->
<!-- Vertical Brick -->
<IMG alt="Also from CNET" height=20 src="http://home.cnet.com/Ads/Media/Imag
es/RHC_ALSOfromCNETnet.gif" width=150 NOSEND="1"><BR>
<TABLE border=0 cellPadding=2 cellSpacing=0 width=150><TBODY><TR vAlign=top>
<TD><FONT face="arial, helvetica" size=-1><B>
<a href="http://clickthru.online.com/Click?q=92-ubRVQTzTtqolHMIzeFOPq1Fsnyn
R" >Live tech help NOW!</a><p>
<a href="http://clickthru.online.com/Click?q=a7-fR_2Q8BwDGVqv-ILF5vLk7KCgFe
R" >April's tech award</a><p>
<a href="http://clickthru.online.com/Click?q=bc-zkPJQ6BdhioQyZwmobF8XEDxwcy
R" >1 million open jobs</a><p>
<a href="http://clickthru.online.com/Click?q=d1-ulOzQpErnMSx69710-kl0aSxkZl
```

```
R" >News.com: Top CIOs</a> <p>
<a href="http://clickthru.online.com/Click?q=e6-qG8rQwyqVGvpdecXGugH6NQYvU4
R" >ZDNet: PeopleSoft</a>
</B></FONT></TD>
</TR></TBODY></TABLE></P>
<!-- End Vertical Brick-->
<!-- ad -->
</td>
<td width=9 rowspan=2><img src="http://www.cnet.com/i/nl/dd_c.gif" width=9 h
eight=60><br>
<table width=9 cellpadding=0 cellspacing=0 border=0>
<tr><td bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width=1 heigh
t=1></td></tr>
<tr><td bgcolor="#ffcc00"><img src="http://www.cnet.com/b.gif" width=1 heigh
t=10></td></tr>
<tr><td bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width=1 heigh
t=1></td></tr>
</table></td>
<td width=1 bgcolor="#000000" rowspan=2><table width=1 cellpadding=0 cellspa
cing=0 border=0>
<tr><td bgcolor="#eeeeee"><img src="http://www.cnet.com/b.gif" width=1 heigh
t=60></td></tr>
</table></td></tr>

<tr valign=top><td width=1 bgcolor="#000000"><img src="http://www.cnet.com/
b.gif" width=1 height=1></td>
<td width=10 bgcolor="#ffcc00"><img src="http://www.cnet.com/b.gif" width=10
height=1></td>
<td width=1 bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width=1 h
eight=1></td>
<td width=430 bgcolor="#ffffff"><table cellpadding=0 cellspacing=0 border=0
width=430>
<tr valign=top><td width=10><img src="http://www.cnet.com/b.gif" width=10 he
ight=1></td>
<td width=412><font face="arial, helvetica" size="-1"><img src="http://www.c
net.com/b.gif" width=1 height=8><br><font size="-2">July 11, 2002</font><br>

<!-- /index -->

<table width="100%" cellpadding=0 cellspacing=0 border=0>
<tr><td colspan=4 bgcolor="#cccccc"><img src="http://www.cnet.com/b.gif" wid
th=1 height=1></td></tr>

<tr><td width="25%"><img src="http://home.cnet.com/Images/Investor/arrow_dow
n.gif" width="7" height="9" alt="" border="0" /> <font face="ms sans serif,
geneva" size="-2"><a href="http://clickthru.online.com/Click?q=fb-aWErQkgq8D
fIcxXoFokcTbTDoVrR" >DJIA</a><br>8801.53  <font color=#cc0000>-11.97</f
ont></td>

<td width="25%"><img src="http://home.cnet.com/Images/Investor/arrow_up.gif"
width="7" height="9" alt="" border="0" /> <font face="ms sans serif, geneva"
size="-2"><a href="http://clickthru.online.com/Click?q=10-YU9qI9NwnpRSHODPff
ACEH5fCAiR" >NASDAQ</a><br>1374.43  <font color=#009933>28.42</font></t
d>

<td width="25%"><img src="http://home.cnet.com/Images/Investor/arrow_up.gif"
width="7" height="9" alt="" border="0" /> <font face="ms sans serif, geneva"
size="-2"><a href="http://clickthru.online.com/Click?q=25-inepIbdXN8Llvx2ZTs
cR7jeQrFPR" >S&P 500</a><br>927.37  <font color=#009933>6.90</font></td
>

<td width="25%"><img src="http://home.cnet.com/Images/Investor/arrow_up.gif"
width="7" height="9" alt="" border="0" /> <font face="ms sans serif, geneva"
size="-2"><a href="http://clickthru.online.com/Click?q=3a-QAmgIrzD8m3Woym7JR
```

```
1rqAFZcCnR" >CNET TECH</a><br>995.16  <font color="#cc0000"><font
color=#009933>29.87</font></font>

</td></tr>
<tr><td colspan=4><img src="http://www.cnet.com/b.gif" width=1 height=3></td
></tr>

<!-- VISION PROMO -->
<tr><td colspan=4 bgcolor="#cccccc"><img src="http://www.cnet.com/b.gif" wid
th=1 height=1></td></tr>
<tr><td colspan=4><img src="http://www.cnet.com/b.gif" width=1 height=3></td
></tr>
<tr><td colspan=4><font face="arial, helvetica" size="-1"><b>CNET News.com V
ision Series</b><br>Read News.com's exclusive interviews of 10 top CIOs.<br>
<a href="http://clickthru.online.com/Click?q=4f-Iar4IIU8lztgSEJ8g-rrsLKFK0u
R" >Vision Series home</a>
</td></tr>
<tr><td colspan=4><img src="http://www.cnet.com/b.gif" width=1 height=3></td
></tr>
<!-- /VISION PROMO -->

<tr><td colspan=4 bgcolor="#cccccc"><img src="http://www.cnet.com/b.gif" wid
th=1 height=1></td></tr>
</table>
<!-- /index -->

<br>
<table width=93 align=left cellpadding=0 cellspacing=0 border=0>
<tr><td><a href="?tag=dd.inv.dht.hed.0"><img src="http://www.cnet.com/i/ne/p
t/2001/03/pts_0313nov.jpg" width=85 height=110 border=0></a></td></tr>
</table>
<font size=3><b>Battered stocks regain ground</b></font><br>
Led by a surge in chip titan Intel, beaten-down technology shares rose Thurs
day.
<p>
Despite a Merrill Lynch downgrade of Yahoo to "reduce/sell", the Internet po
rtal rose after posting better-than-expected second-quarter results and also
helped lift the sector. CNET's Tech index gained 29.87 points, or 3.11 perce
nt, to close at 995.16. The tech-laden Nasdaq composite index tacked on 28.4
2 points, or 2.11 percent, to close at 1,374.43--after hitting a five-year l
ow Wednesday.
<p>
Broader markets were mixed amid news wholesale prices rose slightly and drug
firm Bristol-Myers Squibb is being probed by the Securities and Exchange Com
mission. The Dow Jones industrial average fell 12.0 points, or 0.14 percent,
to close at 8,801.5. The S&P 500 added 6.9 points, or 0.75 percent, to close
at 927.37, after it too reached a five-year low Wednesday.
<br><br>
</font></td>
<td width=8><img src="http://www.cnet.com/b.gif" width=8 height=1></td></tr>
</table>

<table cellpadding=10 cellspacing=0 border=0>
<tr><td>
<!-- top news -->
<table width=400 cellpadding=0 cellspacing=0 border=0>
<tr><td><img src="http://www.cnet.com/i/nl/inv/sub_news.gif" width=192 heigh
t=19 alt="Top Investor News"></td>
<td width="100%"><img src="http://www.cnet.com/b.gif" width=1 height=8></td>
</tr>
<tr bgcolor="#cccccc"><td colspan=2><img src="http://www.cnet.com/b.gif" wid
th=1 height=1></td></tr>
<tr><td colspan=2><img src="http://www.cnet.com/b.gif" width=1 height=5></td
></tr>
```

```
</table>
<font face="arial, helvetica" size="-1">
<font size=3><b><a href="http://clickthru.online.com/Click?q=65-K82zIDPtMvF-
RG6D70VOf0m8P9ZR" >With rebates, HP angles for PC sales </a></b></font><br>
Hewlett-Packard is trying to beat the heat with summer deals on PCs.
For retail buyers in the United States, the company is offering several new
rebate or promotional programs, designed to stimulate demand for its PCs, go
ing into the back-to-school shopping season. Some of the deals include a new
$200 mail-in rebate on desktops, while others offer instant $50 rebates.<br>
<B>HEWLETT-PACKARD CO 15.23 -0.13% </B>
<p>
<font size=3><b><a href="http://clickthru.online.com/Click?q=7a-EE_AIPBjBPYe
1kShU6coezkBZXdR" >Shareholders sue PayPal, eBay </a></b></font><br>
The PayPal-eBay merger has hit an early snag: Two shareholder lawsuits have
been filed against the companies seeking to block the deal.
The lawsuits, each filed in Delaware Chancery Court earlier this week on beh
alf of PayPal shareholders, charge that the deal represents a breach of the
companies' fiduciary duty to those shareholders and that the price eBay is p
aying for PayPal is unfair and inadequate, the companies said in separate re
gulatory filings on Thursday.<br><B>EBAY INC 60.38 2.63% </B>
<p>
<font size=3><b><a href="http://clickthru.online.com/Click?q=8f-OOYvQ8KKOQMi
Cj0Xvsg9tyQFWveR" >SAP cuts forecast, shows Q2 loss </a></b></font><br>
Europe's biggest software maker, SAP AG, shocked the market on Thursday by u
nexpectedly cutting its sales forecast and missing analysts' forecast for th
e second quarter.

The company said it would post a net loss of 235 million euros ($232 millio
n) and said it was taking a non-recurring accounting charge of 414 million e
uros to cover impairments of minority investments in the United States, incl
uding a 318 million euros charge from its 20 percent stake in Commerce One.<
br><B>SAP AKTIENGESELLSCHAFT ADS 19.70 -7.73% </B>
<!-- /top news -->
<p>

<p>
<!-- rtsq -->
<font color="#999999"><b>Also from CNET</b></font><br>
<table width=400 cellpadding=0 cellspacing=0 border=0>
<tr bgcolor="#cccccc"><td><img src="http://www.cnet.com/b.gif" width=1 heigh
t=1></td></tr>
<tr><td><img src="http://www.cnet.com/b.gif" width=1 height=5></td></tr>
</table>
Real-time stock quotes from CNET News.com Investor.<br>30-day <a href="htt
p://clickthru.online.com/Click?q=a4-HGTsQ6paohXSOe2463LvvI2rgYyR" >free tria
l</a>!
<!-- /rtsq -->
<p>
<!-- broker reports -->
<table width=400 cellpadding=0 cellspacing=0 border=0>
<tr><td><img src="http://www.cnet.com/i/nl/inv/sub_brok.gif" width=192 heigh
t=19 alt="Today's Broker Reports"></td>
<td width="100%"><img src="http://www.cnet.com/b.gif" width=1 height=8></td>
</tr>
<tr bgcolor="#cccccc"><td colspan=2><img src="http://www.cnet.com/b.gif" wid
th=1 height=1></td></tr>
<tr><td colspan=2><img src="http://www.cnet.com/b.gif" width=1 height=5></td
></tr>
</table>
<font face="arial, helvetica" size="-1">
<font size=3><b><a href="http://clickthru.online.com/Click?q=b9-WWcsQ-Badhgf
Bp7h8FdMkmoOwzlR" >Merrill Lynch downgrades Yahoo in report 7/11/02 </a></b>
</font><br>
Analysts Justin Baldauf and Tim Gernitis cut their intermediate-term rating
```

```
on the Internet giant from "neutral" to "reduce/sell" and their long-term ra
ting from "strong buy" to "neutral" after seeing Yahoo's second-quarter resu
lts. Although the results look solid at first, Baldauf and Gernitis say Yaho
o's growth was driven entirely by new initiatives. They also argue the fir
m's improved outlook seems driven completely by the Overture deal extension.
Eventually, these initiatives will result in much lower growth rates, the an
alysts say. They argue Yahoo's stock price isn't supported by the firm's fun
damentals.<br><B>YAHOO INC 12.92 5.99% </B><br>
<a href="http://clickthru.online.com/Click?q=ce-Sd0CQMsxwzzzEDkn3ufz1Jey2fP4
R" >Visit the Brokerage Center</a>
<!-- /broker reports -->

<p>

<!-- ceo wealth meter -->
<table width=400 cellpadding=0 cellspacing=0 border=0>
<tr><td><img src="http://www.cnet.com/i/nl/inv/sub_ceo.gif" width=192 height
=19 alt="CEO Wealth Meter"></td>
<td width="100%"><img src="http://www.cnet.com/b.gif" width=1 height=8></td>
</tr>
<tr bgcolor="#cccccc"><td colspan=2><img src="http://www.cnet.com/b.gif" wid
th=1 height=1></td></tr>
<tr><td colspan=2><img src="http://www.cnet.com/b.gif" width=1 height=5></td
></tr>
</table>
<font face="arial, helvetica" size="-1">
<font size=3><b><a href="http://clickthru.online.com/Click?q=e4-ABYFQVkWaEHb
v1jI81u0PT_Qv1yR" >EMC lures HP exec to technology post </a></b></font><br>
EMC has hired Mark Lewis away from Hewlett-Packard to become chief technolog
y officer and vice president of new ventures, the high-end storage specialis
t said Thursday.
Lewis had been head of Compaq Computer's high-end storage division, then bec
ame head of worldwide storage marketing after the company's $19 billion merg
er with HP. He replaces Jim Rothnie, who is retiring and becoming CTO emerit
us.<br><B>EMC CORP 8.15 4.49% </B><br>
<a href="http://clickthru.online.com/Click?q=f9-mtKTQxOWLWRlxGNSM_svdGheo-l
R" >Visit the CEO Wealth Meter</a><br clear=all>
<!-- /ceo wealth meter -->

<p>

</font><br>
</td></tr>
</table>
</td></tr>
</table>
<table width=612 cellpadding=0 cellspacing=0 border=0>
<tr><td width=1 bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width
=1 height=1></td>
<td width=10 bgcolor="#ffcc00"><img src="http://www.cnet.com/b.gif" width=10
height=1></td>
<td width=1 bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width=1 h
eight=1></td>
<td width=599 bgcolor="#cccccc"><img src="http://www.cnet.com/b.gif" width=1
height=1></td>
<td width=1 bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width=1 h
eight=1></td></tr>
</table>
<table width="612" bgcolor="#FFFFFF" cellpadding="0" cellspacing="0" border
="0">
<tr><td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" wid
th="1" height="1"></td>
<td width="10" bgcolor="#ffcc00"><img src="http://www.cnet.com/b.gif" width
="10" height="1"></td>
```

```
<td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width
="1" height="1"></td>
<td width="12"><img src="http://www.cnet.com/b.gif" width="12" height="1"></
td>
<td width="575"><img src="http://www.cnet.com/b.gif" width="1" height="10"><
br>
<!-- Leads Module -->
<img src="http://www.cnet.com/i/gl/new-cc.gif" alt="Crucial Clicks" width="1
20" height="19" border="0"><br>
<table width="100%"  cellpadding="0" cellspacing="0" border="0" bgcolor="#ff
ffff">
<tr bgcolor="#cccccc"><td colspan="7"><img src="http://home.cnet.com/b.gif"
width="1" height="1" border="0"></td></tr>
<tr>
<td bgcolor="#cccccc" rowspan="3"><img src="http://home.cnet.com/b.gif" heig
ht=1 width=1></td>
<td rowspan="3"><img src="http://home.cnet.com/b.gif" height=1 width=5></td>
<td colspan="3"><img src="http://home.cnet.com/b.gif" height=5 width=1></td>
<td rowspan="3"><img src="http://home.cnet.com/b.gif" height=1 width=5></td>
<td bgcolor="#cccccc" rowspan="3"><img src="http://home.cnet.com/b.gif" heig
ht=1 width=1></td>
</tr>
<tr valign="top">
<td width="100%">
                <table width="100%"  cellpadding="0" cellspacing="0" border
="0">
                <tr valign="top">
                <td width="85"><a href="http://clickthru.online.com/Click?q=
0e--eOTIiQzhArLSHv4r8dJx98lbK4R" ><img src="http://a.r.tv.com/cnet.1d/i/pg/0
21502_pointcamera.jpg" width="85" height="110" border="0" alt="" /></a></td>
        <td width="5"><img src="http://home.cnet.com/b.gif" height=1 width=5
></td>
        <td width="100%"><font face="arial, helvetica" size="-1"><a href="ht
tp://clickthru.online.com/Click?q=23-sFcGINM80Cnlnq0iAv2Y9ZZvQKrR" ><b>Digic
ams for summer shutterbugs</b></a><br />Going on vacation, or just headed to
the beach? Indulge your summer snapshot habit with one of our picks.</font>
</td>
                </tr>
                <tr><td colspan="3"><img src="http://home.cnet.com/b.gif" he
ight=3 width=1><br /><font face="arial, helvetica" size="-1">
        &#149; <a href="http://clickthru.online.com/Click?q=38-bLkmIRM2LJWcN
kXnhJiSUxuD-DiR" >5-megapixel shoot-out</a><br />
        &#149; <a href="http://clickthru.online.com/Click?q=4d--URsIKZDQ-uQ0
UQdaB0S22bbKAZR" >Leica Digilux 1: street shooter"s digicam.</a><br />
        </font></td></tr>
        </table>
</td>
<td><img src="http://home.cnet.com/b.gif" height=1 width=5></td>
<td width="180">
        <table cellpadding="0" cellspacing="0" width="180" border="0" bgcolo
r="#ffffef">
        <tr><td colspan="5" bgcolor="#666666"><img src="http://home.cnet.co
m/b.gif" width="1" height="1" alt="" border="0" /></td></tr>
        <tr>
        <td bgcolor="#666666"><img src="http://home.cnet.com/b.gif" width
="1" height="1" alt="" border="0" /></td>
        <td><img src="http://home.cnet.com/b.gif" width="10" height="1" alt
="" border="0" /></td>
        <td><font face="ms sans serif, geneva" size="-2"><img src="http://ho
me.cnet.com/b.gif" width="1" height="5" alt="" border="0" /><br />
        <font face="arial, helvetica" color="#666666" size="-1"><b>Most popu
lar products</b></font><br /><img src="http://home.cnet.com/b.gif" width="1"
height="5" alt="" border="0" /><br />
        <b>Digital cameras</b><br /><img src="http://home.cnet.com/b.gif" wi
```

```
dth="1" height="7" alt="" border="0" /><br />
        1. <a href="http://clickthru.online.com/Click?q=62-4oBiIePgZqOKG4dG9
TmIKJ15BRdR" >Canon PowerShot G2</a><br /><img src="http://home.cnet.com/b.g
if" width="1" height="3" alt="" border="0" /><br />
        2. <a href="http://clickthru.online.com/Click?q=77-u6h6IIpsqtLro70x3
1g9NPmdZPuR" >Canon PowerShot S40</a><br /><img src="http://home.cnet.com/b.
gif" width="1" height="3" alt="" border="0" /><br />
        3. <a href="http://clickthru.online.com/Click?q=8d-nbL1Q7no54XmsTkIi
SupcptbPMPR" >Canon PowerShot S30</a><br /><img src="http://home.cnet.com/b.
gif" width="1" height="3" alt="" border="0" /><br />
        4. <a href="http://clickthru.online.com/Click?q=a2-6-M7QXZxTvbFipI9R
0CGk_HtsonR" >Canon PowerShot A40</a><br /><img src="http://home.cnet.com/b.
gif" width="1" height="3" alt="" border="0" /><br />
        5. <a href="http://clickthru.online.com/Click?q=b7-FOgPQ8JxgOlolIL0N
pbnpTbdL3eR" >Nikon Coolpix 995</a><br /><img src="http://home.cnet.com/b.gi
f" width="1" height="6" alt="" border="0" /><br />
        <img src="http://a.r.tv.com/cnet.1d/i/fd/yl_arrow.gif" width="13" he
ight="13" border="0" alt="" align="absmiddle" /> <b><a href="http://clickthr
u.online.com/Click?q=cc-h4fQQ0OqtVbcqIC7qeZDozn6f5yR" >See all most popular
cameras</a></b><br />
        <img src="http://home.cnet.com/b.gif" width="1" height="10" alt="" b
order="0" /><br />
        </font></td>
        <td><img src="http://home.cnet.com/b.gif" width="10" height="1" alt
="" border="0" /></td>
        <td bgcolor="#666666"><img src="http://home.cnet.com/b.gif" width
="1" height="1" alt="" border="0" /></td>
        </tr>
        <tr><td colspan="5" bgcolor="#666666"><img src="http://home.cnet.co
m/b.gif" width="1" height="1" alt="" border="0" /></td></tr>
        </table>
</td>
</tr>
<tr valign="top"><td colspan="3"><img src="http://home.cnet.com/b.gif" heigh
t=5 width=1></td></tr>
<tr><td colspan="7" bgcolor="#cccccc"><img src="http://home.cnet.com/b.gif"
height=1 width=1></td></tr>
</table><br />
<!-- /Leads Module -->
</td><td width="12"><img src="http://www.cnet.com/b.gif" width="12" height
="1"></td>
<td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width
="1" height="1"></td></tr>
</table>
<!-- ### footer ### -->
<table width="612" bgcolor="#000000" cellpadding="0" cellspacing="0" border
="0">
<tr><td width="1"><img src="http://www.cnet.com/b.gif" width="1" height="1">
</td>
<td width="10" bgcolor="#ffcc00"><img src="http://www.cnet.com/b.gif" width
="10" height="1"></td>
<td width="1"><img src="http://www.cnet.com/b.gif" width="1" height="1"></td
>
<td width="599" bgcolor="#cccccc"><img src="http://www.cnet.com/b.gif" width
="1" height="1"></td>
<td width="1"><img src="http://www.cnet.com/b.gif" width="1" height="1"></td
></tr>
</table>
<table width="612" bgcolor="#eeeeee" cellpadding="0" cellspacing="0" border
="0">
<tr><td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" wid
th="1" height="1"></td>
<td width="10" bgcolor="#ffcc00"><img src="http://www.cnet.com/b.gif" width
="10" height="1"></td>
```

```
<td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width
="1" height="1"></td>
<td width="12"><img src="http://www.cnet.com/b.gif" width="12" height="1"></
td>
<td width="575"><img src="http://www.cnet.com/b.gif" width="1" height="10"><
br>
      <a href="http://clickthru.online.com/Click?q=e1-6RvNQxk2WvmDOAlPhsYK3A
wi9qlR" >
      </a>
      <table width="100%" border="0" cellpadding="5" align="center" borderco
lor="#000000" bgcolor="#CCCCCC">
        <tr bgcolor="#ffcc00">
          <td bgcolor="#CCCCCC"><font face="Arial, Helvetica" size=-1><b><fo
nt size="2" color=blue>NEW!</font></b>
            <font size="2">CNET professional e-mail publishing for just $24.
95/month.
            <b><a href="http://clickthru.online.com/Click?q=f6-QktIQMwrDgXcL
AfYYRGOT9_Jmp4R" >FREE for 30 days. Click
            here!</a></b></font></font></td>
        </tr>
      </table>
      <p><a href="http://clickthru.online.com/Click?q=0b-0NI-INQ4dZR1DaTGftm
ZCtvxYhrR" ><img src="http://a.r.tv.com/cnet.1d/i/nl/ft.gif" width="375" hei
ght="18" border="0" alt="Sign up for more free newsletters from CNET!"></a><
br>
        <!-- subscription management -->
        <font face="ms sans serif, geneva" size="-2"> The e-mail address for
your
        subscription is qqqqqqqqqq-zdnet@example.com</font></p>
      <p><font face="ms sans serif, geneva" size="-2">

<A NOTRACK HREF='http://clickthru.online.com/Click?q=21-UPZOzEdWyZZDaQmwrqLf
8L42jnb7L9RR'>Unsubscribe</A> | <a href='http://nl.com.com/servle
t/url_login?email=qqqqqqqqqq-zdnet@example.com&brand=cnet'>Manage My Subscri
ptions</a> | <a href="http://clickthru.online.com/Click?q=36-wg7-I
iI4oUHco2_MxvH3QainXQ4R" >FAQ</A> | <a href="http://clickthru.onli
ne.com/Click?q=4b-k_l7IBUPPQwjTRaP9VA35c54JMcR" >Advertise</A><p>
Please send any questions, comments, or concerns to <a href="mailto:dis
patchfeedback@news.com">dispatchfeedback@news.com</A>.<br></font>
</font>

<!-- /subscription management-->

<img src="http://www.zdnet.com/b.gif" width="1" height="8"></td>
<td width="12"><img src="http://www.cnet.com/b.gif" width="12" height="1"></
td>
<td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width
="1" height="1"></td></tr>
</table>
<table width="612" bgcolor="#000000" cellpadding="0" cellspacing="0" border
="0">
<tr><td width="1"><img src="http://www.cnet.com/b.gif" width="1" height="1">
</td>
<td width="10" bgcolor="#ffcc00"><img src="http://www.cnet.com/b.gif" width
="10" height="1"></td>
<td width="601"><img src="http://www.cnet.com/b.gif" width="1" height="1"></
td></tr>
</table>
<table width="612" bgcolor="#ffcc00" cellpadding="0" cellspacing="0" border
="0">
<tr><td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" wid
th="1" height="1"></td>
<td width="10"><img src="http://www.cnet.com/b.gif" width="10" height="1"></
td>
```

```
<td width="37"><a href="http://clickthru.online.com/Click?q=60-oqItIsEAuhURJ
kELN8S_eSgFWVRR" ><img src="http://www.cnet.com/i/dp/smrb.gif" width="37" he
ight="37" border="0"></a></td>
<td width="563" nowrap><font face="arial, helvetica" size="-1">
<a href="http://clickthru.online.com/Click?q=75-whwtIKUAIuvS8ED_rjups7LLZ5Z
R" ><font color="#000000">Price comparisons</font></a> |
<a href="http://clickthru.online.com/Click?q=8a-3y3dQXndaN53J2EO3Nt-tKChBwn
R" ><font color="#000000">Product reviews</font></a> |
<a href="http://clickthru.online.com/Click?q=40-6Pp4Idn8DzxUeaceVXCHnpqL0wR
R" ><font color="#000000">Tech news</font></a> |
<a href="http://clickthru.online.com/Click?q=55-TAHaIKJ8Xqi8ylrqm86ccJ_2d2Z
R" ><font color="#000000">Downloads</font></a> |
<a href="http://clickthru.online.com/Click?q=6a-z_30IeWiUkrR9cqzceCGFySNP8d
R" ><font color="#000000">All CNET services</font></a>
</font></td>
<td width="1" bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width
="1" height="1"></td></tr>
</table>
<table width="612" cellpadding="0" cellspacing="0" border="0">
<tr><td bgcolor="#000000"><img src="http://www.cnet.com/b.gif" width="1" hei
ght="1"></td></tr>
<tr><td height="25"><font face="ms sans serif, geneva" size="-2"><table widt
h=100% border=0 cellspacing=2 cellpadding=1> <tr valign=bottom> <td width=7
5% height=31> <p></font> <br><b><font face=Arial, Helvetica, sans-serif size
=2>    Copyright 2002 CNET Networks, Inc. All rights reserved.    </font></b
> </p></td><td height=31 valign=top> <div align=right> <img src="http://gser
v-cnet.zdnet.com/clear/outbound.gif?APPID=2&EMID=25136482&NL=e426&ISSUE=2002
-07-11" height=1 width=1>    </div></td></tr><tr>  <td colspan=2><font face=
Arial, Helvetica, sans-serif size=2>    </font></td></tr></table>
.</font></td></tr>
</table>
<!-- ### /footer ### -->

</body>
</html>
<IMG HEIGHT=1 WIDTH=1 SRC="http://clickthru.online.com/Click?q=81-z1Ka-xKW8d
-4uEJI7_9st9QKU9RR">


------------------------------------------------NEW MAIL-----------------
--------------------------------
From 102192086381143-17090200005-example.com?zzzz@bounce.tilw.net  Tue Sep 1
7 17:23:29 2002
Return-Path: <102192086381143-17090200005-example.com?zzzz@bounce.tilw.net>
Delivered-To: zzzz@localhost.jmason.org
Received: from localhost (jalapeno [127.0.0.1])
        by zzzzason.org (Postfix) with ESMTP id 938AE16F03
        for <zzzz@localhost>; Tue, 17 Sep 2002 17:23:27 +0100 (IST)
Received: from jalapeno [127.0.0.1]
        by localhost with IMAP (fetchmail-5.9.0)
        for zzzz@localhost (single-drop); Tue, 17 Sep 2002 17:23:27 +0100 (I
ST)
Received: from webnote.net (mail.webnote.net [193.120.211.219]) by
    dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g8HGF3C17155 for
    <zzzz@jmason.org>; Tue, 17 Sep 2002 17:15:03 +0100
Received: from sonic1.tilw.net (sonic1.tilw.net [209.164.4.167]) by
    webnote.net (8.9.3/8.9.3) with SMTP id RAA27422 for <zzzz@example.com>;
    Tue, 17 Sep 2002 17:15:31 +0100
From: CopyYourDVD <atomica2020@hotmail.com>
Subject: Friend, Copy ANY DVD or Playstation Game with this software......
To: zzzz@example.com
X-Owner: atomicDOT;mp*qhwqrwhlqf!frp;8;
MIME-Version: 1.0
X-RMD-Text: yes
```

```
Date: Tue, 17 Sep 2002 09:15:32 PST
X-Mailer: 2.0-b55-VC_IPA [Aug 20 2002, 12:25:33]
Message-Id: <17090200005$102192086381143$1159552220$0@sonic1.tilw.net>
Content-Type: multipart/alternative; boundary="------------103227703017104"

--------------103227703017104
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit

Friend,Now you can copy DVD's and Games
http://www.terra.es/personal9/iop1008/

BACKUP DVD VIDEO's WITH YOUR CD-R BURNER

With 321 studio's software, you can now copy
any DVD and Playstation Game.  Never buy another
backup DVD movie again.  Just copy it!

This is the first time this software is being made
available to the public. All the software you need
to burn your own DVD Video, is included in 321 Studio's
software package DVD Copy Plus!  The movies will play
in a standard DVD player. With detailed, easy to follow,
step-by-step instructions, you can BURN your own DVD
Video using nothing more than your DVD-ROM
and CD-R drives. Purchase a copy! Click below.

http://www.terra.es/personal9/iop1008/

Order today and receive!

*Step by Step Interactive Instructions
*All Software Tools Included On CD
*No DVD Burner Required
*FREE Live Technical Support
*30 Day Risk Free Trial Available

With DVD Copy Plus you can backup Your DVD Movies with
the same 74min or 80min CD-R's you've used in the past
to create audio CD's. Our software compresses the large
DVD files on your standard DVD to VCD, SVCD, and DivX
much the same way the popular MP3 format compresses audio.
Order today and start burning
http://www.terra.es/personal9/iop1008/

Thank You,

CopymyDVD


http://inglesa.net/unsub.php?client=atomicDOT



We take your privacy very seriously and it is our policy never to send
unwanted email messages. This message has been sent to zzzz@example.com
because you originally joined one of our member sites or you signed up
with a party that has contracted with atomicDOT. Please
http://tilw.net/unsub.php?client=atomicDOT&msgid=17090200005
to Unsubscribe (replying to this email WILL NOT unsubscribe you).
```

```
TRCK:atomicDOT;mp*qhwqrwhlqf!frp;8;



--------------103227703017104
Content-Type: text/html; charset=us-ascii
Content-Transfer-Encoding: 7bit

<HTML><HEAD><TITLE>Backup your DVD's</TITLE>
<META http-equiv=Content-Type content="text/html; charset=iso-8859-1">
<META content="MSHTML 6.00.2600.0" name=GENERATOR></HEAD>


<DIV></DIV>
<BLOCKQUOTE
style="PADDING-RIGHT: 0px; PADDING-LEFT: 5px; MARGIN-LEFT: 5px; BORDER-LEFT:
#000000 2px solid; MARGIN-RIGHT: 0px">
  <DIV></DIV>
  <TABLE cellSpacing=0 cellPadding=0 width=525 align=center bgColor=#ffffff
  border=0>

    <TBODY>
    <TR>
      <TD width="33"><IMG height=1 src="images/spacer.gif" width=33 border=0
></TD>
      <TD width="174"><IMG height=1 src="images/spacer.gif" width=174 border
=0></TD>
      <TD width="360"><IMG height=1 src="images/spacer.gif" width=360 border
=0></TD>
      <TD width="33"><IMG height=1 src="images/spacer.gif" width=33 border=0
></TD>
    </TR>
    <TR>
      <TD colSpan=4><a href="http://www.terra.es/personal9/iop1008/"><IMG he
ight=253 src="http://www.terra.es/personal9/iop1008/images/mailer8_r1_c1.gi
f"
        width=600 useMap=#Map border=0></a></TD>
    </TR>
    <TR>
      <TD background=http://www.terra.es/personal9/iop1008/images/mailer8_r2
_c1.gif height=181> </TD>
      <TD vAlign=top align=left colSpan=2 height=181>
        <P><FONT face="Verdana, Arial, Helvetica, sans-serif" size=2>With DV
DCopyPlus
          and PowerCDR5.0 you can copy and burn your</FONT></P>
        <UL>
          <LI><FONT color=#003399><B><FONT
          face="Verdana, Arial, Helvetica, sans-serif" size=2>DVD Movies</FO
NT></B></FONT>
          <LI><FONT color=#003399><B><FONT
          face="Verdana, Arial, Helvetica, sans-serif" size=2>Playstation </
FONT></B></FONT>
          <LI><FONT color=#003399><B><FONT
          face="Verdana, Arial, Helvetica, sans-serif" size=2>MP3s, AVIs and
all
            other multimedia Files</FONT></B></FONT>
          <LI><FONT color=#003399><B><FONT
          face="Verdana, Arial, Helvetica, sans-serif" size=2>Software, Musi
c
            CDs (perfect RAW data duplication)</FONT></B></FONT>
          <LI><FONT color=#003399><B><FONT
          face="Verdana, Arial, Helvetica, sans-serif" size=2>NEW! Burn to C
D-R
```

```
            or DVD-R</FONT></B></FONT></LI>
          </UL>
          <P><FONT face="Verdana, Arial, Helvetica, sans-serif" size=2>Protect
your
            investments by backing up your CD, DVD, MP3, Data and Game collect
ions
            with burned copies that really work! Separately, DVDCopyPlus and P
owerCDR5.0
            are over $100 worth of software that you can purchase bundled righ
t
            now for just $49.99!</FONT></P>
          <P align=center><FONT face="Verdana, Arial, Helvetica, sans-serif"
          size=2><B><FONT size=5>THAT'S 50% OFF!<BR>
            </FONT><A
          href="http://www.terra.es/personal9/iop1008/">MORE DETAILS</A> <A
          href="http://www.terra.es/personal9/iop1008/">ORDER NOW</A></B></FON
T></P>
        </TD>
        <TD background=http://www.terra.es/personal9/iop1008/images/mailer8_r2
_c4.gif height=181> </TD>
      </TR>
      <TR>
        <TD colSpan=4><IMG height=24 src="http://www.terra.es/personal9/iop100
8/images/mailer8_r3_c1.gif"
          width=600 border=0 name=mailer8_r3_c1></TD>
      </TR>
      <TR>
        <TD height="18"></TD>
        <TD></TD>
        <TD></TD>
        <TD></TD>
      </TR>
      </TBODY>
    </TABLE>
    <MAP name=Map><AREA shape=POLY
      coords=244,119,270,129,280,114,297,125,318,113,327,129,355,120,352,133,3
85,131,381,140,398,145,398,152,405,151,397,160,383,165,409,172,381,180,404,1
93,384,195,382,207,384,212,364,210,351,213,353,224,327,216,317,230,297,219,2
78,232,269,218,243,225,238,210,211,210,209,202,199,193,216,180,186,172,216,1
64,192,151,195,143,211,136,211,132,231,135,243,130,243,121
      href="http://www.terra.es/personal9/iop1005/"></MAP><FONT
    face=Verdana,Geneva,Arial,Helvetica,sans-serif color=#000000 size=1>
    <P> </P></FONT></BLOCKQUOTE></BODY></HTML>
<br><br><br><br>
<p><font size="1" face="Arial">We take your privacy very seriously and it is
our policy never to send unwanted email messages.  This message has been sen
t to zzzz@example.com because you originally joined one of our member sites
or you signed up with a party that has contracted with atomicDOT.  Please <a
href="http://tilw.net/unsub.php?client=atomicDOT&msgid=17090200005" TARGET
="_blank">Click Here</a></font> to Unsubscribe  <font size="1" face="Arial">
(replying to this email WILL NOT unsubscribe you).  <br><br><br><br>

<>
</body></html><br><br>

<img src="http://tilw.net/counter.php?client=atomicDOT&email=zzzz@example.co
m&msgid=17090200005" width="1" height="1" border="0" alt="">

<br><br><br>
<font size="-5">TRCK:atomicDOT;mp*qhwqrwhlqf!frp;8;</font>



--------------103227703017104--
```

Answer 1.1: We think we would be able to classify the mails by looking at them but it would take a long time. We think a model could classify the mails successfully to some degree, but some not with 100% accuracy

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.

In [4]:
```
# Converting list to dataframe
df_easy_ham = pd.DataFrame(easy_ham)
df_hard_ham = pd.DataFrame(hard_ham)
df_spam = pd.DataFrame(spam)

df = pd.concat([df_easy_ham, df_hard_ham, df_spam], ignore_index=True)

print('# of Easy_Ham Mails:', + len(df_easy_ham))
print('# of Hard_Ham Mails:', + len(df_hard_ham))
print('# of Spam Mails:', + len(df_spam))
df.sample(5)
```

```
# of Easy_Ham Mails: 2551
# of Hard_Ham Mails: 250
# of Spam Mails: 501
```

Out[4]:

	msg	label
457	From fork-admin@xent.com Thu Sep 26 11:04:44 ...	ham
152	From fork-admin@xent.com Wed Aug 28 10:50:41 ...	ham
3193	From mortgage_quotes_fast@nationwidemortgage.u...	spam
1561	From rssfeeds@jmason.org Thu Oct 10 12:32:34 ...	ham
1932	From quinlan@pathname.com Wed Aug 28 10:45:34...	ham

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.

In [5]:
```
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB, BernoulliNB
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix, accuracy_score
import seaborn as sns
import matplotlib.pyplot as plt
```

In [6]:
```
def data(df_ham, df_spam):
    # Splitting the data into train and test
    hamtrain, hamtest = train_test_split(df_ham, test_size=0.3, random_state
    spamtrain, spamtest = train_test_split(df_spam, test_size=0.3, random_st

    # Merging the train and test dataframes
    X_train = pd.concat([hamtrain, spamtrain], ignore_index=True)
    X_test = pd.concat([hamtest, spamtest], ignore_index=True)

    # Shuffling the data
    X_train = X_train.sample(frac=1).reset_index(drop=True)
    X_test = X_test.sample(frac=1).reset_index(drop=True)

    # Assigning the labels to y_train and y_test
    y_train = X_train.label
    y_test = X_test.label

    X_train = X_train.msg
    X_test = X_test.msg

    return X_train, X_test, y_train, y_test
```

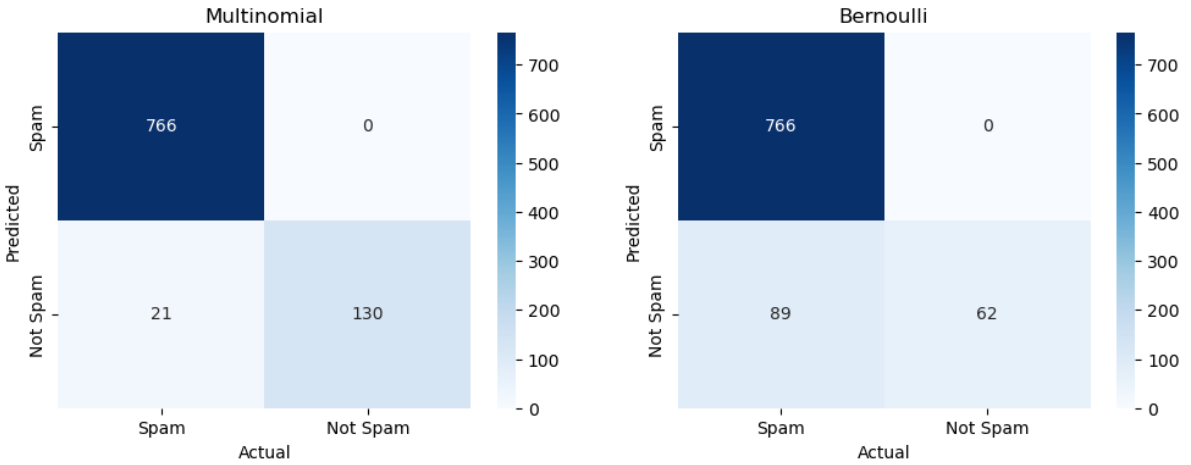
In [7]:
```
def train_model(X_train, y_train, X_test, y_test, vect, model):
    # vectorize the text
    vect.fit(X_train)

    # transform the text into matrixes
    X_train_vector = vect.transform(X_train)
    X_test_vector = vect.transform(X_test)

    # Train the model
    model.fit(X_train_vector, y_train)
    y_pred = model.predict(X_test_vector)
    acc = accuracy_score(y_test, y_pred)

    # Confusion Matrix
    cm = confusion_matrix(y_test, y_pred)

    TN = cm[0][0]
    FN = cm[1][0]
    TP = cm[1][1]
    FP = cm[0][1]

    return cm, TN, FN, TP, FP, acc
```

In [8]:
```
def plot_cm(MNB_cm, BNB_cm, mTP, mFN, mTN, mFP, MNM_acc, bTP, bFN, bTN, bFP,
    # Visualizing the confusion matrix
    fig, ax = plt.subplots(1,2)
    fig.set_size_inches(12, 4)
    labels=['Spam', 'Not Spam']

    sns.heatmap(MNB_cm, annot=True, fmt='d', xticklabels=labels, yticklabels
```

```
    ax[0].set_ylabel('Predicted')
    ax[0].set_xlabel('Actual')
    ax[0].set_title('Multinomial')

    sns.heatmap(BNB_cm, annot=True, fmt='d', xticklabels=labels, yticklabels
    ax[1].set_ylabel('Predicted')
    ax[1].set_xlabel('Actual')
    ax[1].set_title('Bernoulli')

    # Scores TPR = TP/TP+FN, TNR = TN/FP+TN
    print('Multinomial Accuracy: ', round(MNM_acc,3), "%")
    print('True Spam: ', round(mTP /(mFP+mTP),3))
    print('True Ham: ', round(mTN /(mTN+mFN),3))

    print('\nBernoulli Accuracy: ', round(BNB_acc,3), "%")
    print('True Spam: ', round(bTP /(bFP+bTP),3))
    print('True Ham: ', round(bTN /(bTN+bFN),3))
```

In [9]:
```
MNB_model = MultinomialNB()
BNB_model = BernoulliNB()
vect = CountVectorizer()

X_train, X_test, y_train, y_test = data(df_easy_ham, df_spam)

MNB_cm, mTP, mFN, mTP, mFP, m_acc = train_model(X_train, y_train, X_test, y_
BNB_cm, bTP, bFN, bTP, bFP, b_acc = train_model(X_train, y_train, X_test, y_

print("Easy Ham VS Spam with only easy ham training data \n")

plot_cm(MNB_cm, BNB_cm, mTP, mFN, mTP, mFP, m_acc, bTP, bFN, bTP, bFP, b_acc
```

```
Easy Ham VS Spam with only easy ham training data

Multinomial Accuracy:  0.977 %
True Spam:  1.0
True Ham:  0.861

Bernoulli Accuracy:  0.903 %
True Spam:  1.0
True Ham:  0.411
```



2.2) Answer the following questions:

a) What does the CountVectorizer do?

CountVectorizer converts the text strings into a matrix the have words as coulums and files as rows, where each cell is the cont of a word within the text file.

b) What is the difference between Multinomial Naive Bayes and Bernoulli Naive Bayes

From the confusion matrixes, we can also note that the Multinomial Naive Bayes classifier performs better than the Bernoulli Naive Bayes classifier, as it doesn't missclassify as much spam emails while also having a higher accuracy score. However, it is worth noting that both classifiers are good at classifying ham, e.g non-spam, emails, as they both have equally high true positive rate.

A difference between the classifiers is that Bernoulli Naive Bayes can be viewed as binary where it only cares if a word is used or not in the file, whereas the Mulitnomial Navie Bayes takes into account how many times each word is used.

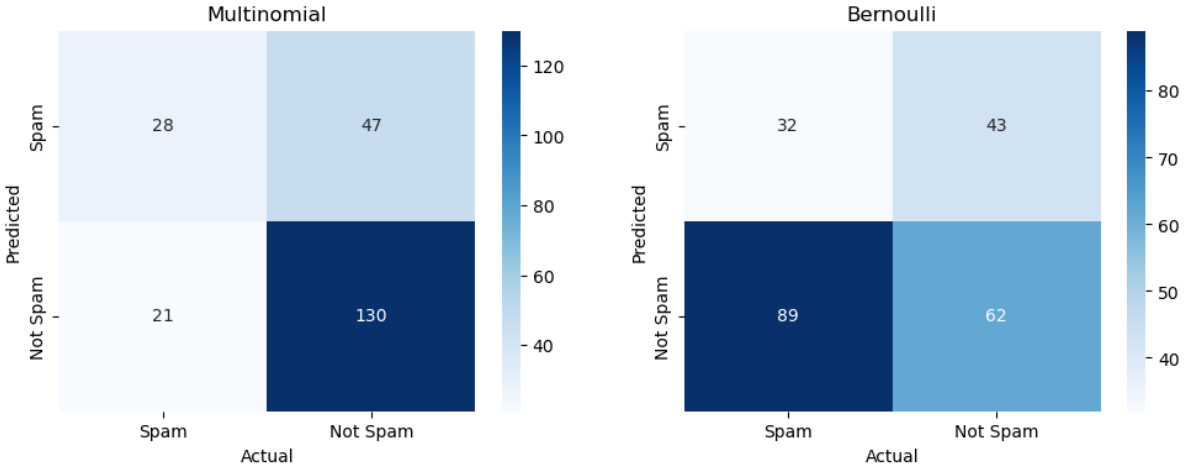
3.1 Run the two models:

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?

In [10]:
```
# Gettig relevant data
X_train, X_test, y_train, y_test = data(df_hard_ham, df_spam)

# transform the text into matrixes
X_train_vector = vect.transform(X_train)
X_test_vector = vect.transform(X_test)

# Using models that were trained on easy ham and spam
MNB_pred = MNB_model.predict(X_test_vector)
BNB_pred = BNB_model.predict(X_test_vector)

# Accuracy
MNB_acc = accuracy_score(y_test, MNB_pred)
BNB_acc = accuracy_score(y_test, BNB_pred)

# Confusion Matrix
MNB_cm = confusion_matrix(y_test, MNB_pred)
BNB_cm = confusion_matrix(y_test, BNB_pred)

mTN = MNB_cm[0][0]
mFN = MNB_cm[1][0]
mTP = MNB_cm[1][1]
mFP = MNB_cm[0][1]

bTN = BNB_cm[0][0]
bFN = BNB_cm[1][0]
bTP = BNB_cm[1][1]
bFP = BNB_cm[0][1]
```

In [11]:
```
print("Hard Ham VS Spam with only easy ham training data \n")

plot_cm(MNB_cm, BNB_cm, mTP, mFN, mTP, mFP, MNB_acc, bTP, bFN, bTP, bFP, BNE
```

```
Hard Ham VS Spam with only easy ham training data

Multinomial Accuracy:  0.699 %
True Spam:  0.734
True Ham:  0.861

Bernoulli Accuracy:  0.416 %
True Spam:  0.59
True Ham:  0.411
```



Answer 3.1:

When the same models is used, but the data is changed to Spam and Hard ham, the overall performance drops. A reason for this is that the Hard Ham emails are more similar to spam emails than easy ham emails. Seeing as the previous model was trained on only easy ham emails, it is not difficult to see why it performs much worse.

3.2 Retrain

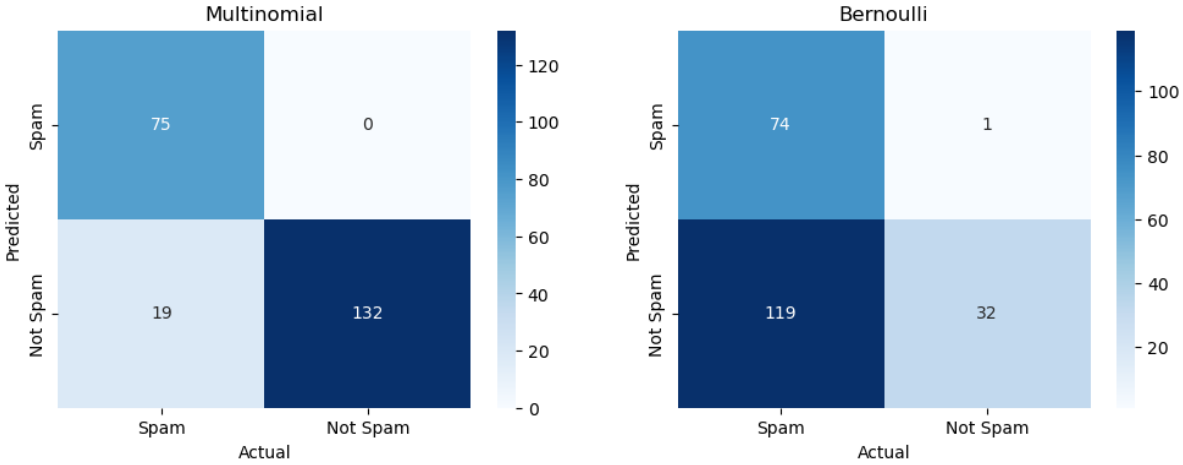
Retrain new Multinomial and Bernolli Naive Bayes classifers on the combined (easy+hard) ham and spam. Now evaluate on spam versus hard-ham as in 3.1. Also evaluate on spam versus easy-ham. Compare the performance with question 2 and 3.1. What do you observe?

In [12]:
```
MNB_model = MultinomialNB()
BNB_model = BernoulliNB()
new_vect = CountVectorizer()

# Merging easy and hard ham
df_easy_hard = pd.concat([df_easy_ham, df_hard_ham], ignore_index=True)
df_easy_hard = df_easy_hard.sample(frac=1).reset_index(drop=True)

X_train, X_test, y_train, y_test = data(df_easy_hard, df_spam)

# Training the new models with easy and hard ham + spam
_ = train_model(X_train, y_train, X_test, y_test, new_vect, MNB_model)
_ = train_model(X_train, y_train, X_test, y_test, new_vect, BNB_model)
```

In [13]:
```
# Spam vs hard ham
X_train, X_test, y_train, y_test = data(df_hard_ham, df_spam)

# transform the text into matrixes
X_train_vector = new_vect.transform(X_train)
X_test_vector = new_vect.transform(X_test)
```

```
# Using models that were trained on easy ham and spam
MNB_pred = MNB_model.predict(X_test_vector)
BNB_pred = BNB_model.predict(X_test_vector)

MNB_acc = accuracy_score(y_test, MNB_pred)
BNB_acc = accuracy_score(y_test, BNB_pred)

# Confusion Matrix
MNB_cm = confusion_matrix(y_test, MNB_pred)
BNB_cm = confusion_matrix(y_test, BNB_pred)

mTN = MNB_cm[0][0]
mFN = MNB_cm[1][0]
mTP = MNB_cm[1][1]
mFP = MNB_cm[0][1]

bTN = BNB_cm[0][0]
bFN = BNB_cm[1][0]
bTP = BNB_cm[1][1]
bFP = BNB_cm[0][1]

print("Hard Ham VS Spam with easy and hard ham training data")
plot_cm(MNB_cm, BNB_cm, mTP, mFN, mTP, mFP, MNB_acc, bTP, bFN, bTP, bFP, BNB
```

```
Hard Ham VS Spam with easy and hard ham training data
Multinomial Accuracy:  0.916 %
True Spam:  1.0
True Ham:  0.874

Bernoulli Accuracy:  0.469 %
True Spam:  0.97
True Ham:  0.212
```



Answer 3.2a:

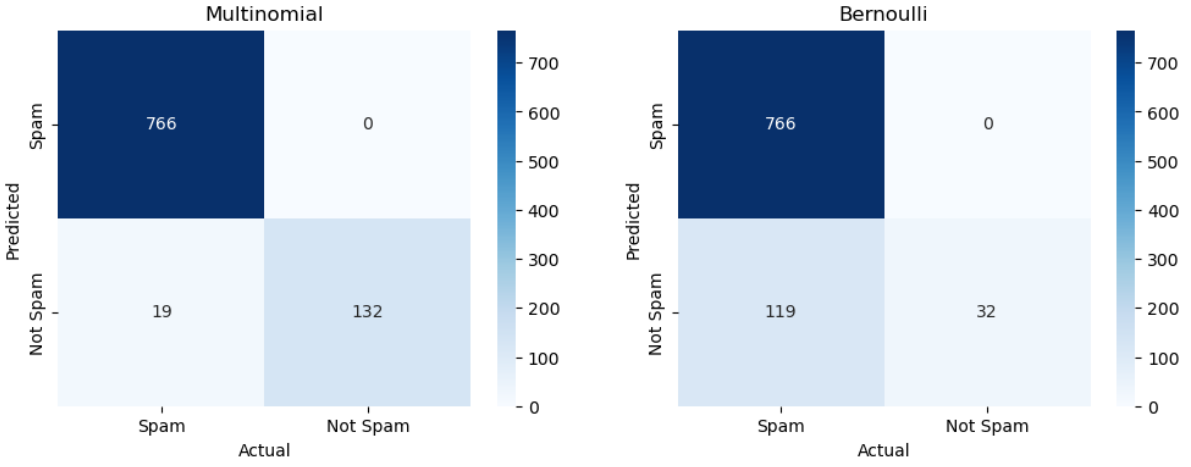
When comparing the performance of the new model, which is trained on both easy and hard ham emails, to the result in 3.1, we can see that the performance increases massively for the Multinomial model, from 0.70% to 0.916%. However, for the Bernoulli model we don't see a big difference as the performance only increases from 0.416% to 0.478%. Looking at the confusion matrix for the Bernoulli model, we can see that it is better at detecting true spam, while being worse at detecting true ham.

A conclusion we can draw from this is that the Multinomial model benefits greatly from being trained on both easy and hard ham emails. This is not the case with the Bernoulli

model. Instead, it has a more difficult time distinguishing between ham and spam emails when trained on both easy and hard ham emails.

In [14]:
```
# Spam vs easy ham
X_train, X_test, y_train, y_test = data(df_easy_ham, df_spam)

# transform the text into matrixes
X_train_vector = new_vect.transform(X_train)
X_test_vector = new_vect.transform(X_test)

# Using models that were trained on easy ham and spam
MNB_pred = MNB_model.predict(X_test_vector)
BNB_pred = BNB_model.predict(X_test_vector)

MNB_acc = accuracy_score(y_test, MNB_pred)
BNB_acc = accuracy_score(y_test, BNB_pred)

# Confusion Matrix
MNB_cm = confusion_matrix(y_test, MNB_pred)
BNB_cm = confusion_matrix(y_test, BNB_pred)

mTN = MNB_cm[0][0]
mFN = MNB_cm[1][0]
mTP = MNB_cm[1][1]
mFP = MNB_cm[0][1]

bTN = BNB_cm[0][0]
bFN = BNB_cm[1][0]
bTP = BNB_cm[1][1]
bFP = BNB_cm[0][1]

print("Easy Ham VS Spam with easy and hard ham training data \n")
plot_cm(MNB_cm, BNB_cm, mTP, mFN, mTP, mFP, MNB_acc, bTP, bFN, bTP, bFP, BNB
```

```
Easy Ham VS Spam with easy and hard ham training data

Multinomial Accuracy:  0.979 %
True Spam:  1.0
True Ham:  0.874

Bernoulli Accuracy:  0.87 %
True Spam:  1.0
True Ham:  0.212
```



Answer 3.2b:

In the second case, where we compare the result from Q2 (easy ham VS spam with only easy ham training data) to Q3.2, we can't see mcuh of a difference from neither models.

The accuracy for Multinomial model is pretty much the same, changing from 0.977% to 0.979%. For the Bernoulli model, the accuracy actually decreases from 0.903% to 0.869%. This is not that difficult to explain, as in Q2, the Bernoulli model is only trained on easy ham emails. In Q3.2 it is trained on hard ham emails as well, which are more similar to spam emails than easy ham emails. Therefore, it is likely that the model is more confused when it comes to classifying spam emails.

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:

To make the models perform even better, we suggest to remove all unnecessary information in the emails such as headers and footers. If these sections were removed, it would make the model more accurate, as it will only focus on the actual content of the emails. Furthermore, another idea is to remove all words that are not relevant to the classification, such as "filling words" like "the", "a", "an", etc. This would also make the model more accurate, as it would only focus on the words that are relevant to the classification.