Data 100/200 Homework 10 Written

Jackie Hu

TOTAL POINTS

11 / 15

QUESTION 1

1 Question 1 2 / 6

+ **6 pts** Responded to all three three parts thoughtfully.

✓ **+ 2 pts** **Thoughtful response to part 1.**

+ **2 pts** Thoughtful response to part 2.

+ **2 pts** Thoughtful response to part 3.

+ **0 pts** Blank/No effort

QUESTION 2

2 Question 2 6 / 6

✓ **+ 3 pts** **Chooses appropriate plot that includes title, axis labels**

✓ **+ 3 pts** **Describes plot and implications with respect to features**

- **1 pts** Plot does not include title and/or axis labels

- **1 pts** Doesn't include implications of visualization

+ **0 pts** Incorrect/blank

QUESTION 3

3 Question 3 3 / 3

✓ **+ 3 pts** **Correct axes and reasonable curve**

+ **2 pts** Reasonable curve

+ **1 pts** Correctly labelled axes

+ **0 pts** Blank/incorrect

Notebook

April 22, 2021

0.0.1 Question 1: Feature/Model Selection Process

In this following cell, describe the process of improving your model. You should use at least 2-3 sentences each to address the follow questions:

1. How did you find better features for your model?
2. What did you try that worked or didn't work?
3. What was surprising in your search for good features?

[13]:
```
pd.set_option('display.max_rows', None)
```

Useful Functions

[14]:
```
def find_the_count_in_email(df, pattern):
    word_count_list = []

    for line in df['email']:
        l = re.findall(pattern, line)
        word_count = len(l)
        word_count_list.append(word_count)
    return pd.Series(word_count_list)
```

[15]:
```
def find_the_count_in_subject(df, pattern):
    word_count_list = []

    for line in df['subject']:
        l = re.findall(pattern, line)
        word_count = len(l)
        word_count_list.append(word_count)
    return pd.Series(word_count_list)
```

Number of characters in the subject / body

[ ]:
```
train
```

[17]:
```
ham = train[train['spam'] == 0]
```

[18]:
```
spam = train[train['spam'] == 1]
```

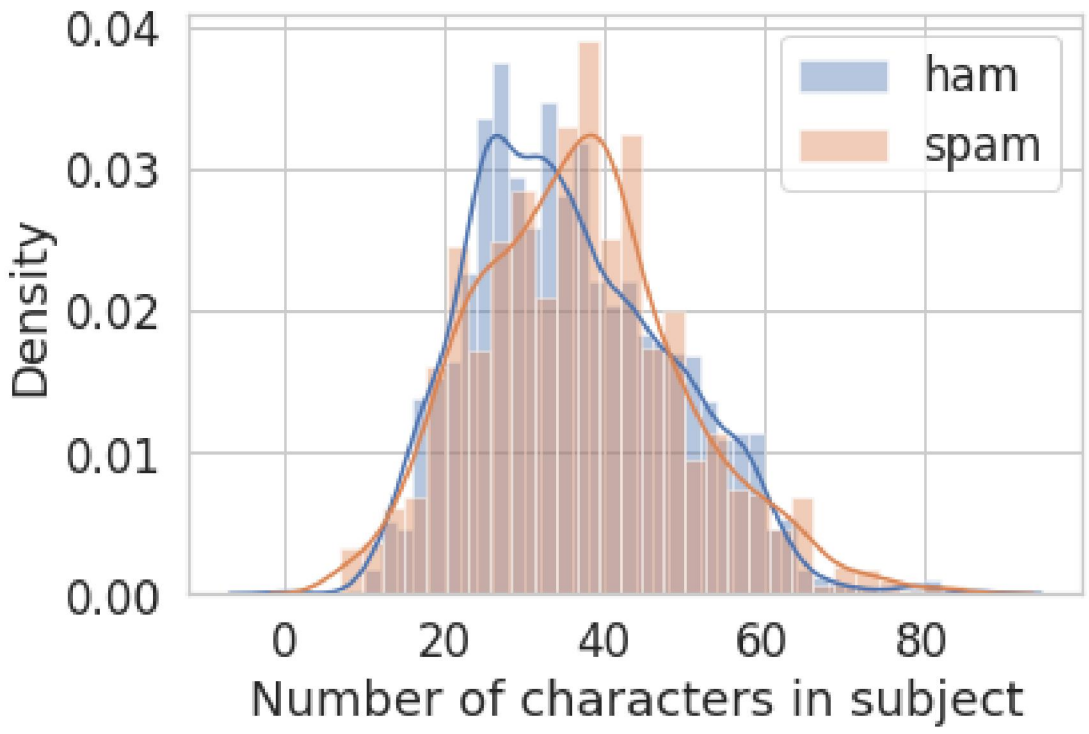
[19]:
```
#number of characteds in subject
def num_char_in_df(df, col):
    char_num_list = []
    #col = str(col)
    char = df[col].str.findall('\w')
    for entry in char:
        char_num_list.append(len(entry))
        #print(len(entry))
    return char_num_list
```

[20]:
```
train_subject_char_h = num_char_in_df(ham, 'subject')
train_body_char_h = num_char_in_df(ham, 'email')
```

[21]:
```
train_subject_char_s = num_char_in_df(spam, 'subject')
train_body_char_s = num_char_in_df(spam, 'email')
```

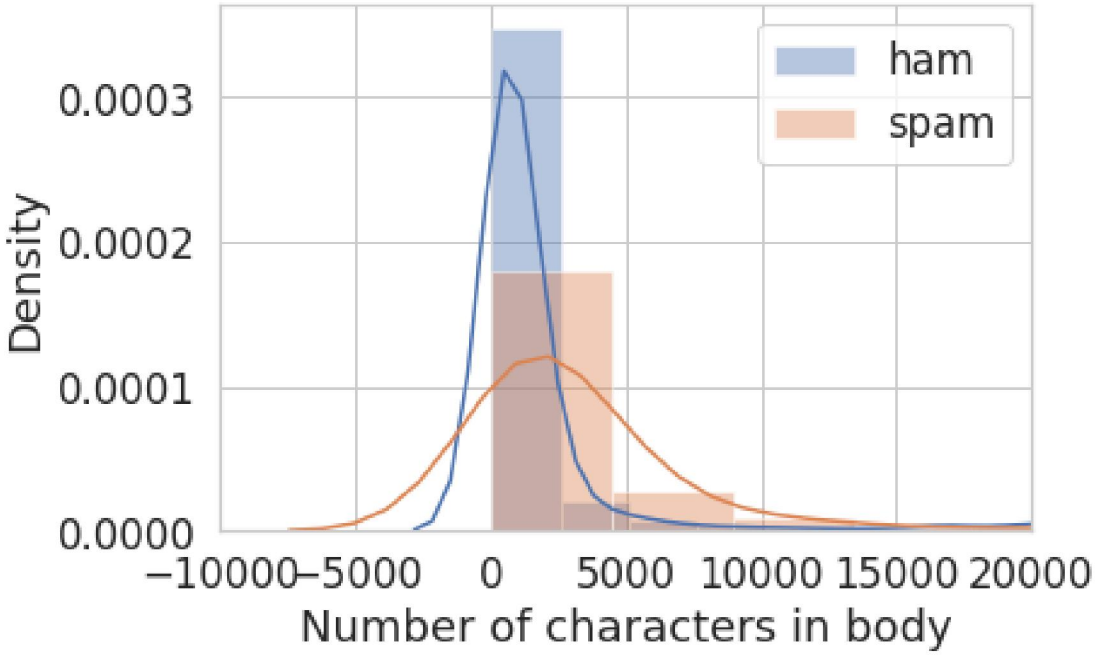
[22]:
```
sns.distplot(train_subject_char_h, label = 'ham')
sns.distplot(train_subject_char_s, label = 'spam')
plt.xlabel('Number of characters in subject')
plt.legend()
```

[22]: <matplotlib.legend.Legend at 0x7f5e7b9c23a0>



[23]:
```
sns.distplot(train_body_char_h, label = 'ham')
sns.distplot(train_body_char_s, label = 'spam')
plt.xlabel('Number of characters in body')
plt.xlim(-10000, 20000)
plt.legend()
```

[23]:
```
<matplotlib.legend.Legend at 0x7f5e7ba19a60>
```



- similar distribution in the 'number of characters in the subject'; however, body seems to contain less characters in the spam comparing to ham.
- can be used as useful feature

[24]:
```
pattern_ch = r"\w"
```

Number of words in the subject / body

[25]:
```
import re
```

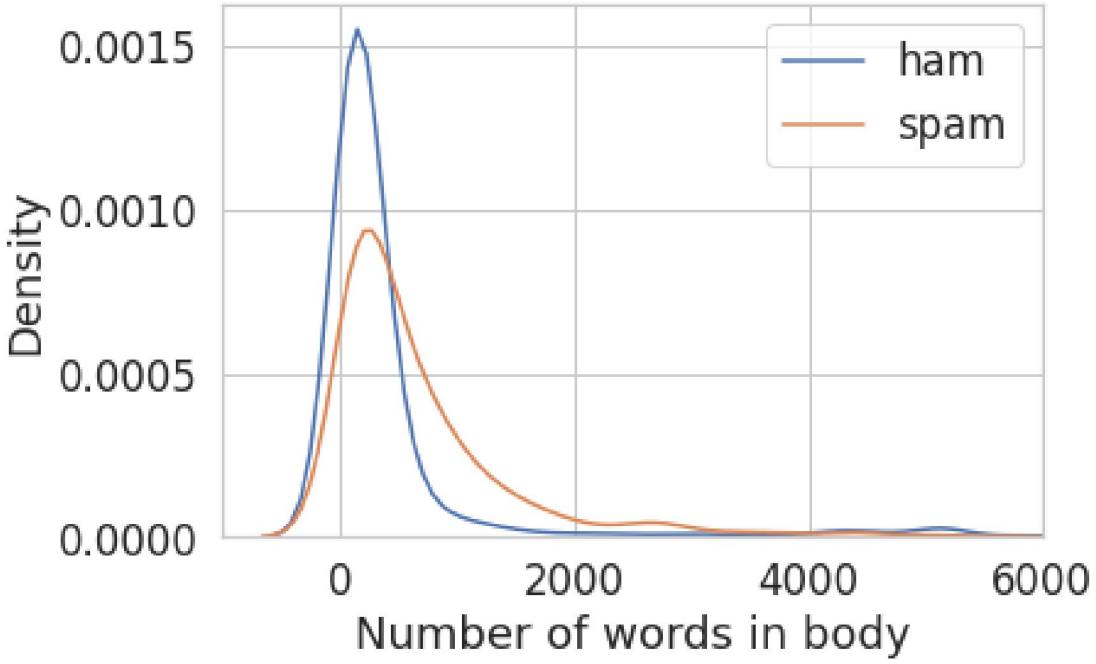
[26]:
```
pattern_w = r'\w+'
```

[27]:
```
ham_word_count_list = find_the_count_in_email(ham, pattern_w)
spam_word_count_list = find_the_count_in_email(spam, pattern_w)
```

[28]:
```
sns.distplot(ham_word_count_list, label = 'ham', hist=False)
sns.distplot(spam_word_count_list, label = 'spam', hist=False)
plt.xlim(-1000, 6000)
```

```
plt.xlabel('Number of words in body')
plt.legend()
```

[28]:
```
<matplotlib.legend.Legend at 0x7f5e783eb760>
```



- the number of words in ham email body is more than in spam email body, but this feature is similar to finding the number of characters in body, which didn't really add much addition information to the prediction model.

Use of html tag

[29]:
```
train['email'][4]
```

[29]:
```
'<html>\n <head>\n <title>tech update today</title>\n </head>\n <body
style="margin:8px 9px 9px 12px" bgcolor="#ffffff"
background="http://techupdate.zdnet.com/techupdate/i/bg_232850.gif"
link="#003399" alink="#cc0000" vlink="#666699">\n \n <div align="center">\n \n
<!-- main -->\n <a name="top"></a>\n <table width=612 bgcolor="#232850"
cellpadding=0 cellspacing=0 border=0>\n <tr valign=bottom>\n <td width=440
colspan=4><a href="http://clickthru.online.com/click?q=0e--
eotiiqzhxizsnm4r8djx98lbk4r" ><img
src="http://www.zdnet.com/techupdate/i/itnewsletter_today.gif" width="440"
height="60" border="0"  alt="tech update today"></a></td>\n <td width=160
align=center valign=top rowspan=2 bgcolor="#ffffff">\n         <img
src="http://www.zdnet.com/techupdate/i/gl_ad2.gif" width="160" height="16"
border="0"  alt="advertisement"><br>\n         <!--tower -->\n        <iframe
```

src="http://www.zdnet.com/include/ads/ifc/rgroup=2766" scrolling="no"
frameborder="0" hspace="0" vspace="0" height="600" width="160" marginheight="0"
marginwidth="0">\n <script language="javascript"
src="http://www.zdnet.com/include/ads/js/rgroup=2766">\n </script>\n </iframe>\n
<!-- tower -->\n \n \n \n </td>\n <td width=12 colspan=3
valign=bottom><img src="http://www.zdnet.com/techupdate/i/gl_corner2.gif"
width="12" height="6" border="0" alt=""></td>\n </tr>\n <tr valign=top>\n <td
width=1 bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif" width=1
height=1></td>\n <td width=10 bgcolor="#1e5c99"><img
src="http://www.zdnet.com/b.gif" width=10 height=1></td>\n <td width=1
bgcolor="#000000"><img src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n
<td width=428 bgcolor="#ffffff">\n \n <table cellpadding=0 cellspacing=0
border=0>\n <tr valign=top>\n <td width=7 rowspan=2><img
src="http://www.zdnet.com/b.gif" width=7 height=1></td>\n <td
width=248><img src="http://www.zdnet.com/b.gif" width=1 height=5>
\n
<img src="http://www.zdnet.com/techupdate/i/dots248.gif" width="248"
height="5">
\n <a
href="http://clickthru.online.com/click?q=23-sfcginm8kmpanomiav2y9zzvqkrr" ><img
src="http://www.zdnet.com/techupdate/i/news_tech_header.gif" width="248"
height="24" vspace="6" border=0 alt="tech update today">
 \n
<!--date-->\n <img src="http://www.zdnet.com/techupdate/i/dots248.gif"
width="248" height="5">
\n <font face="ms sans
serif, geneva" size="-2" color="#666666">vital signs for july 15,
2002
 \n <!--date-->\n \n <img
src="http://www.zdnet.com/techupdate/i/dots248.gif" width="248"
height="5">
\n <img src="http://www.zdnet.com/b.gif" width=1
height=5>
\n <!--lede art-->\t\t\n \n <table cellpadding=0 cellspacing=0
border=0 align=left>\n \n \n <tr valign=top><td width="90"><img
src="http://image.com.com/techupdate/i/dfarber.jpg" width="85" height="110"
border="0"></td></tr>\n </table>\n \n <!--lede art-->\n \n \n dan farber
\n \n \n
\n <a
href="http://clickthru.online.com/click?q=38-blkmirm2py08xjfnhjisuxud-dir" >\n
unplugged: fbi cio darwin john
\n \n \n in the twilight of
his career, an it veteran takes on the formidable task of transforming the
fbi\'s antiquated technology infrastructure. he\'s being sworn in today.\n \n \n
\n
\n <img src="http://www.zdnet.com/techupdate/i/arrow_blue.gif"
width="8" height="10" border="0">\n \t <a
href="http://clickthru.online.com/click?q=4e-0y95inzasfvhi6jiklstn8n_jj9r" >\n
\t read my interview with darwin john\n
\n \n \n
<img src="http://www.zdnet.com/b.gif" width=1 height=8>
\n </td>\n
<td width=5 rowspan=2><img src="http://www.zdnet.com/b.gif" width=5
height=1></td>\n <td width=5 bgcolor="#fffff3" rowspan=2><img
src="http://www.zdnet.com/b.gif" width=5 height=1></td>\n <td width=160
bgcolor="#fffff3" rowspan=2><img src="http://www.zdnet.com/b.gif" width=1
height=7>
\n <font face="verdana, arial" size="-2">\n \n
<!---- headlines---->\n \n latest from zdnet news
<img

```
src="http://www.zdnet.com/b.gif" width=1 height=5><br>\n \n \n <img
src="http://www.zdnet.com/b.gif" height=7 width=1><br>\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=63-tpkjihwsgv_ryftg_9-hwkp_wecr" >yopy
brings linux to wider audience</a><br>\n \n <img
src="http://www.zdnet.com/b.gif" height=7 width=1><br>\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=78-d5asidudupsde5r5vnmun5xsporr"
>activists to isps: don\'t be a stoolpigeon</a><br>\n <img
src="http://www.zdnet.com/b.gif" height=7 width=1><br>\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=8d-nbl1q7notycugupiisupcptbpmpr" >new
athlon to launch in stormy seas\n </a><br>\n <img
src="http://www.zdnet.com/b.gif" height=7 width=1><br>\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=a2-6-m7qxzx3gj1yak9r0cgk_htsonr"
>wireless companies ripe for consolidation</a><br>\n <img
src="http://www.zdnet.com/b.gif" height=7 width=1><br>\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=b7-fogpq8jxs54h15z0npbnptbdl3er" >w3c
boosts web services language\n </a><br>\n  <img src="http://www.zdnet.com/b.gif"
height=7 width=1><br>\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=cc-h4fqq0oq5aj8c567qezdozn6f5yr"
>cisco hooks up airport \'hot spots\'</a><br>\n <img
src="http://www.zdnet.com/b.gif" height=7 width=1><br>\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=e2-j1opqhyzy9mcjdbenbezdihv9knr"
>palm, handspring win patent spat</a><br>\n \n \n \n <img
src="http://www.zdnet.com/b.gif" height=7 width=1><br>\n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=f7-n0qnqlgzfy038fnlrzigir6imoer" >\n
<b>more enterprise news</b></a><br>\n \n <img src="http://www.zdnet.com/b.gif"
height=7 width=1><br>\n \n <img src="http://www.zdnet.com/b.gif" height=7
width=1><br>\n <img src="http://www.zdnet.com/b.gif" height=6 width=1><br>\n
</font>\n </td>\n <td width=3 bgcolor="#fffff3" rowspan=2><img
src="http://www.zdnet.com/b.gif" width=3 height=1></td>\n         </tr>\n
<tr><td><img src="http://www.zdnet.com/techupdate/i/dots249.gif" width=248
height=5></td></tr>\n         </table><img src="http://www.zdnet.com/b.gif"
```

```
width=1 height=10><br>\n  <!---- headlines---->   \n  \n  \n <!---editors
choice--->    \n <table cellpadding="7" cellspacing="0" border="0">\n \n <tr>\n
<td>\n <font face="arial, helvetica" size="-1">      \n \n \n \n \n <!---
david\'s picks --->\n \n \n \n \n <font size="+1" color="#990000"><b>david\'s
picks</b></font>\n <br>\n                 <table width="100%" cellpadding="0"
cellspacing="0" border="0">\n                     <tr>\n <td bgcolor="#cccccc">\n
<img src="http://www.zdnet.com/b.gif" width="1" height="0"></td>\n </tr>\n
<tr>\n <td><img src="http://www.zdnet.com/b.gif" width="1" height="4">\n </td>\n
</tr>\n                 </table>\n \t\t\t    <table cellpadding="0"
cellspacing="0" border="0" width="85" align="left">\n <tr valign=top>\n
<td>\n <img src="http://www.zdnet.com/techupdate/i/dberlind.gif" width="85"
height="110" border="0">\n <br>\n \n <div align="center"><font face="ms sans
serif, geneva" size="-2">\n dan<br>farber</font></div>\n <img
src="http://www.zdnet.com/techupdate/i/dots75.gif" width=75 height=5>\n
</td>\n \n                 <td width=10>\n <img src="http://www.zdnet.com/b.gif"
width=10 height=1></td>\n \n                 </tr>\n </table>\n <!--story 1-->\n
\n \n \n <font size="+1"><b>it sees new profits in old
glory</b></a></font><br>\n \n \n \n \n can crm be tweaked to track terrorists?
as the fight against <b>terrorism</b> heats up, the government turns to the tech
industry to come up with tools to secure the country against attack. it\'s an
opportunity for patriotism--and for profits. (nearly $38 billion is slated for
homeland security.) <b>software companies</b>  are sending tens of thousands of
proposals to government agencies, some of which are so new they don\'t have the
budget to hire people to evaluate them all.\n \n \n \n <br><img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=0d-3aguibr2qb7tcefnv0t0qky6b7pr" >\n
<b>read more the zdnet news focus</b></a>\n \n <p>\n \n <!--story 1-->\n  \n \n
\n <font size="+1"><b>are spam blocklists going too far?</b></a></font><br>\n \n
many businesses are turning to spam-filtering tools to regain\n control of their
e-mail boxes. but legitimate e-mails are being \n deleted, too, now that
<b>sometimes-indiscriminate blacklists</b> have\n become a key weapon in the war
against unsolicited bulk e-mail.\n  companies\n subscribe to the lists, bouncing
any traffic directed to their\n servers that originates from those addresses. is
<b>overkill</b> the\n answer to stopping spam? what is your company doing to
stop\n junk e-mail?\n \n \n <br><img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=22-xnrrirda8qrryu0uo4i9yc99ncnr" >\n
<b>read the full story</b></a>\n \n <p>\n \n \n \t    \n \n \n   <!-- story 2
-->\n \n \n   <table cellpadding="0" cellspacing="0" border="0" width="85"
align="left">\n <tr valign=top>\n \n <td>\n   <img
src="http://techupdate.zdnet.com/techupdate/i/wayne_rash85x110.jpg" width="85"
height="110" border="0" align="left">\n </td>\n \n                   <td
width=5>\n <img src="http://www.zdnet.com/b.gif" width=5  height=1></td>\n \n
</tr>\n </table>\n \n <font size="+1">\n   <b>attacks are on the
rise</b></a></font><br>\n \n no, you\'re not paranoid. <b>wayne rash</b> warns
```

that the number of\n attacks on internet-connected networks is increasing, with no \n sign of letting up. here\'s what you can do to protect your\n business against malicious attacks.\n \n
\n <img src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10" border="0">\n \n read the full story\n \n \n \n <p>\n \n <!-- story 3 -->\n \n \n \n gartner unfolds sun\'s server roadmap
\n \n \n sun\'s midframe and high-end servers will see only evolutionary\n change through 2006--faster processors, support for pci-x i/o, \n and infiniband, according to gartner. more significant\n developments will come through software, with moves toward\n providing public utility-style availability and capacity. but gartner\n adds note of caution concerning sun\'s credibility.\n \n \n \n
<img src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10" border="0">\n \n read the full story\n \n <p>\n \n \n   \t   \t \n \n \n \n \n \n \n <!--story 4-->\n \n \n \n \'my favorite pda is the nokia 9290\'
\n \n tired of carrying around both pda and phone, reader john maravilla says he spent a year "looking for a decent integrated pda/phone" before finding the nokia 9290. then he read my "what to look for in your next pda/phone" and wrote to reassure me that i\'m not crazy for liking a pda resembling a cell phone from the reagan era.\n \n
<img src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10" border="0">\n \n read his full letter\n \n \n
<img src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10" border="0">\n \n "what to look for in your next pda/phone"\n \n <p>\n \n \n \n <!-- story 5-->\n   \n a hulk of a system
\n \n how much does a cutting-edge pc cost these days? plenty--but sys performance 2533 delivers a lot in return. it\'s packed to the rafters with high-end components, and boasts the highest office-productivity score zdnet reviewers have seen.\n \n
<img src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10" border="0">\n \n read the full review

\n <img src="http://www.zdnet.com/enterprise/graphics/071202sys2533.jpg" width="400" height="206" border="0">
\n \n <p>\n \n \n <!-- story 6 -->\n \n <!--cell phone personality test
\n \n whether you\'re a total gear head or you like to keep things simple, \n take our cell phone personality test to find out which models suit you best.\n \n \n
<img

```
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a
href="http://clickthru.online.com/click?q=b5-50fcq7u79nplufr1phq15zneezpr" >\n
<b>read the full story</b></a>\n \n <br><img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">\n <a href="http://clickthru.online.com/click?q=cb-cjriqjsr-
vmtnihyxyhrf_uahkrr" >\n <b>david berlind\'s rx for mobile happiness</b></a>\n
\n <p>-->\n \n \n <!-- write to me -->\n \n <img
src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8" height="10"
border="0">write me at <a
href="mailto:david.berlind@cnet.com"><b>david.berlind@cnet.com</b></a><br><br>\n
\n         <img src="http://www.zdnet.com/techupdate/i/arrow_blue.gif" width="8"
height="10" border="0"><a href="#top"><b>back to top</b></a><br><br>\n \n \n
<p>\n         </font></td></tr>\n         </table>\n \t\t\n <!---editors
choice---> \n \n \n \n \n </td>\n <td width=1 bgcolor="#000000"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=10
bgcolor="#1e5c99"><img src="http://www.zdnet.com/b.gif" width=10
height=1><br></td>\n <td width=1 bgcolor="#83a3cb"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n </tr>\n </table>\n
<!-- /main -->\n \n \n <!--bottom -->\n <table cellpadding=0 cellspacing=0
border=0 width=612>\n <tr valign=top>\n <td width=1 bgcolor="#83a3cb"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=10
bgcolor="#1e5c99"><img src="http://www.zdnet.com/b.gif" width=10
height=1><br></td>\n <td width=1 bgcolor="#000000"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=12
bgcolor="#ffffff"><img src="http://www.zdnet.com/b.gif" width="12"
height=1></td>\n <td width=564 bgcolor="#ffffff">\n <br><br>\n          \n
<!-------- also on zdnet ----------->\n         <font face="arial, helvetica"
size="+1" color="#990000"><b>also on tech update today</b></font><br>\n
<table width="100%" cellpadding="0" cellspacing="0" border="0">\n
<tr><td bgcolor="#cccccc"><img src="http://www.zdnet.com/b.gif" width="1"
height="1"></td></tr>\n         <tr><td><img src="http://www.zdnet.com/b.gif"
width="1" height="4"></td></tr>\n         </table>\n         \n         \n  <!--
quote block -->\n  <!-- \n         <font face="verdana, geneva" color="#666666"
size="-2"><b>you said it</b></font><sub> </sub><br>\n         <table
width="100%" cellpadding="0" cellspacing="0" border="0"><tr>\n         <td
width="20"><img src="http://www.zdnet.com/b.gif" width="20" height="1"></td>\n
<td width="100%"><font face="arial, helvetica" size="-1">\n <font size="+0"><a
href="http://clickthru.online.com/click?q=e0-cisrqfdhklzggtykphb7mxut9air"
><b>microsoft: the real remedy</b></a></font><p>\n \n <img
src="http://techupdate.zdnet.com/techupdate/i/quote_blue_left.gif"
align="left">\n \n web services will require application-level firewalls\n \n
firewalls were built to plug network holes and shield\n application data. but
gartner says the integration-heavy\n demands of web services require securing
information at\n the application level. \n
http://techupdate.zdnet.com/techupdate/stories/main/0,14179,2864540,00.html\n \n
<img src="http://techupdate.zdnet.com/techupdate/i/quote_blue_right.gif"
```

```
align="right">\n <p align="right">\n \n <font size="-1" color="#990000"><b>--
richard spellman in response to <a
href="http://clickthru.online.com/click?q=f6-qktiqmwrvvc8laxyyrgot9_jmp4r"
>"embedded xp isn\'t the remedy."</a></b></a></font><br>\n     \n \n \n \n
<br><br>\n \n </font>  </td>\n         </tr></table>\n -->\n <!--end quote
block-->\n \n \n \t\t\n \n         <!-- review -->\n \n <font face="verdana,
geneva" color="#666666"
size="-2"><b>commentary</b></font><sub> </sub><br>\n \n <table width="100%"
cellpadding="0" cellspacing="0" border="0"><tr>\n <td width="20"><img
src="http://www.zdnet.com/b.gif" width="20" height="1"></td>\n <td
width="100%"><font face="arial, helvetica" size="-1">\n \n <font size="+0">\n <a
href="http://clickthru.online.com/click?q=0b-0ni-inq4zccfd6_gftmzctvxyhrr" >\n
<b>next ms office? here\'s what to add--and delete</b></a></font><br>\n \n the
next version of microsoft office is due in the next year or so. if you were
product manager for the industry-standard office suite, what would you add? what
would you get rid of? what would you fix? here\'s <b>david coursey\'s</b> wish-
list.\n \n \n  <br><br>\n  \n         </font></td>\n         </tr></table>\n \n
<!-- feature -->\n \t\t\n \n \n <font face="verdana, geneva" color="#666666"
size="-2"><b>q&a</b></font><sub> </sub><br>\n \n <table width="100%"
cellpadding="0" cellspacing="0" border="0"><tr>\n <td width="20"><img
src="http://www.zdnet.com/b.gif" width="20" height="1"></td>\n <td
width="100%"><font face="arial, helvetica" size="-1">\n \n <font size="+0">\n <a
href="http://clickthru.online.com/click?q=20-poifirzzey9aiybtr-pp7us2qfir" >\n
<b>we can handle ceo cheats</b></a></font><br>\n \n \n <b>phil bond,</b> the
u.s. undersecretary of commerce for technology, explains why the financial
industry can ride out the current \n storm without major government
intervention.\n \n \n \n \n  <br><br>\n  \n         </font></td>\n
</tr></table>\n \t\t\n \n \t\t        <!-- download -->\n \n \n \n \n
<!-- /coming tomorrow -->\n \t\t\n \t\t\t\t        <!-- coming tomorrow -->\n \n
\n         <font face="verdana, geneva" color="#666666" size="-2"><b>previously
on tech update today</b></font><sub> </sub><br>\n         <table
width="100%" cellpadding="0" cellspacing="0" border="0"><tr>\n         <td
width="20"><img src="http://www.zdnet.com/b.gif" width="20" height="1"></td>\n
<td width="100%"><font face="arial, helvetica" size="-1">\n \n <font
size="+0">\n <a
href="http://clickthru.online.com/click?q=35-d6jtibiziyemlvlknxltsypwc3pr" >\n
<b>oracle issues a new challenge to microsoft</b></a></font><br>\n \n \n
oracle\'s new <b>collaboration suite</b> includes e-mail, calendar and messaging
software and is less expensive than microsoft\'s offerings, says oracle ceo
<i>larry ellison</i>. \n \n \n <br>\n <br>\n \n \n         </font></td>\n
</tr></table>\n \n         <!-- /coming tomorrow -->\n \t\t\n \n <!-- ### -->\n
<table width="100%"  cellpadding="0" cellspacing="0" border="0"
bgcolor="#ffffff">\n <tr bgcolor="#999999"><td colspan="7" height="15"><font
face="ms sans serif, geneva" color="#ffffff" size="-2"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=3><b>crucial clicks--products worth looking at</b></font></td></tr>\n
<tr>\n <td bgcolor="#999999" rowspan="3"><img
```

```
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=1></td>\n <td rowspan="3"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=5></td>\n <td colspan="3"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=5
width=1></td>\n <td rowspan="3"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=5></td>\n <td bgcolor="#999999" rowspan="3"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=1></td>\n </tr>\n <tr valign="top">\n <td width="100%">\n \t<table
width="100%"  cellpadding="0" cellspacing="0" border="0">\n \t<tr><td
colspan="3"><font face="verdana, geneva" color="#333333"
size="-2"><b>networking</b></font><br /><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=3
width=1><br /></td></tr>\n \t<tr valign="top">\n \t<td width="85"><a
href="http://clickthru.online.com/click?q=4a-hcguippcjuklco9sxolkktkhn-dr" ><img
src="http://www.zdnet.com/products/graphics/052302gateway.jpg" width="85"
height="110" border="0" alt="" /></a></td>\n \t<td width="5"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=5></td>\n \t<td width="100%"><font face="arial, helvetica" size="-1"><a
href="http://clickthru.online.com/click?q=5f-wk5biikx_z7cra46fez6pa6cijur"
style="text-decoration:none"><b>find the right home network</b></a> <br
/>choosing the right technology to match your networking needs is vital. here\'s
help in making the choice.</font></td>\n \t</tr>\n \t<tr><td colspan="3"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=3
width=1><br /><img src="http://techupdate.zdnet.com/techupdate/i/arrow_blue.gif"
width="8" height="10" border="0" alt="" /><font face="arial, helvetica"
size="-1"><a
href="http://clickthru.online.com/click?q=74-vav8izzsxvx6enqryg8y0gspzxsr"
style="text-decoration:none">read reviews</a></font></td></tr>\n \t</table>\n
</td>\n <td><img src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif"
height=1 width=5></td>\n <td width="180">\n \t<table cellpadding="0"
cellspacing="0" width="180" border="0" bgcolor="#eeeeee">\n \t<tr
bgcolor="#999999"><td colspan="3" height="15"><font face="ms sans serif, geneva"
color="#ffffff" size="-2"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=3>most popular products</font></td></tr>\n \t<tr>\n \t<td><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="10"
height="1" alt="" border="0" /></td>\n \t<td><font face="ms sans serif, geneva"
size="-2"><img src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif"
width="1" height="5" alt="" border="0" /><br />\n \t<b>networking</b><br /><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="1"
height="7" alt="" border="0" /><br />\n \t<b>1</b>. <a
href="http://clickthru.online.com/click?q=89-h28bq-j1j6gwv5lkumbxb6w5wwlr"
style="text-decoration:none">linksys etherfast wireless ap</a><br /><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="1"
height="3" alt="" border="0" /><br />\n \t<b>2</b>. <a
```

```
href="http://clickthru.online.com/click?q=9f-o7xmq1jdphss86jylcghvilcu_er"
style="text-decoration:none">linksys etherfast router</a><br /><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="1"
height="3" alt="" border="0" /><br />\n \t<b>3</b>. <a
href="http://clickthru.online.com/click?q=b4-ajlbq6z1ugcyfxkqjzdxdxozlcyr"
style="text-decoration:none">siemens speedstream router</a><br /><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="1"
height="3" alt="" border="0" /><br />\n \t<b>4</b>. <a
href="http://clickthru.online.com/click?q=c9-3ozlqpgkbbzo4wyjg0sxipzvh6lr"
style="text-decoration:none">wireless 802.11b router</a><br /><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="1"
height="3" alt="" border="0" /><br />\n \t<b>5</b>. <a
href="http://clickthru.online.com/click?q=de-7tpkqmwawxfqteattbds-bjjho4r"
style="text-decoration:none">netgear he102 802.11a wireless ap</a><br /><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="1"
height="6" alt="" border="0" /><br />\n \t<img
src="http://techupdate.zdnet.com/techupdate/i/arrow_blue.gif" width="8"
height="10" border="0" alt="" /><b><a
href="http://clickthru.online.com/click?q=f3-ljwgqjowqdnmmspcs3cyjqsjeorr" >more
popular networking products</a></b><br />\n \t<img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="1"
height="10" alt="" border="0" /><br />\n \t</font></td>\n \t<td><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" width="10"
height="1" alt="" border="0" /></td>\n \t</tr>\n \t</table>\n </td>\n </tr>\n
<tr valign="top"><td colspan="3"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=5
width=1></td></tr>\n <tr><td colspan="7" bgcolor="#999999"><img
src="http://www.zdnet.com/graphics/anchordesk/frontpage/b.gif" height=1
width=1></td></tr>\n </table>\n <!-- /### -->\n \n <br>\n <br>\n \n
<!-------- elsewhere ----------->\n         <font face="arial, helvetica"
size="+1" color="#990000"><b>elsewhere on zdnet</b></font><br>\n         <table
width="100%" cellpadding="0" cellspacing="0" border="0">\n         <tr><td
bgcolor="#cccccc"><img src="http://www.zdnet.com/b.gif" width="1"
height="1"></td></tr>\n         <tr><td><img src="http://www.zdnet.com/b.gif"
width="1" height="4"></td></tr>\n         </table>   \n \t\t<font face="arial,
helvetica" size="-1">\n <br>\n \n <b>\n \n <a
href="http://clickthru.online.com/click?q=08-ifoqirjfbppvkmbg3f-ercpgy9ir" >need
a memory upgrade? find out with cnet\'s memory configurator.</a>\n <p>\n <a
href="http://clickthru.online.com/click?q=1d-6ktjigqfrznejkhrl28nrqsnqqpr"
>clearance center: get discounts on pcs, pdas, mp3 players and more!</a>\n <p>\n
<a href="http://clickthru.online.com/click?q=38-blkmirm2py08nj_nhjisuxud-dir"
>find out the top 10 web services security requirements at tech update.</a>\n
<p>\n <a href="http://clickthru.online.com/click?q=4d--
ursikzdjpsn0htdab0s22bbkazr" >builder.com shows you how to bring java to the
masses with cold fusion mx.</a>\n <p>\n <a
href="http://clickthru.online.com/click?q=62-4obiiepgpihmgiyg9tmikj15brdr"
>check out thousands of it job listings in zdnet\'s career center.</a>\n </b>\n
```

\n \n
\n
   \n   \n
</td>\n <td width=12
bgcolor="#ffffff"><img src="http://www.zdnet.com/b.gif" width=12
height=1></td>\n <td width=1 bgcolor="#000000"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=10
bgcolor="#1e5c99"><img src="http://www.zdnet.com/b.gif" width=10
height=1></td>\n <td width=1 bgcolor="#83a3cb"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n </tr>\n </table>\n
<!-- /bottom -->\n <!-- ### footer ### -->\n <table cellpadding=0 cellspacing=0
border=0 width=612>\n <tr valign=top>\n <td width=1 bgcolor="#83a3cb"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=10
bgcolor="#1e5c99"><img src="http://www.zdnet.com/b.gif" width=10
height=1>
</td>\n <td width=1 bgcolor="#000000"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=588
bgcolor="#cccccc"><img src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n
<td width=1 bgcolor="#000000"><img src="http://www.zdnet.com/b.gif" width=1
height=1></td>\n <td width=10 bgcolor="#1e5c99"><img
src="http://www.zdnet.com/b.gif" width=10 height=1>
</td>\n <td width=1
bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n
</tr>\n </table>\n \n <!-- ### subscription management ### -->\n <table
width="612" bgcolor="#ffffef" cellpadding="0" cellspacing="0" border="0">\n
<tr><td width="1" bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif"
width="1" height="1"></td>\n <td width="10" bgcolor="#1e5c99"><img
src="http://www.zdnet.com/b.gif" width="10" height="1"></td>\n <td width="1"
bgcolor="#000000"><img src="http://www.zdnet.com/b.gif" width="1"
height="1"></td>\n <td width="12"><img src="http://www.zdnet.com/b.gif"
width="12" height="1"></td>\n <td width="564"><font face="arial, helvetica"
size="-1"><img src="http://www.zdnet.com/b.gif" width="1" height="5">
\n
sign up for more <a
href=\'http://nl.com.com/servlet/url_login?email=qqqqqqqqqq-
zdnet@example.com&brand=zdnet\'>free newsletters from zdnet
\n
<img src="http://www.zdnet.com/b.gif" width="1" height="5">
\n <!--
subscription management -->\n \n \n
the e-mail address for your subscription is qqqqqqqqqq-
zdnet@example.com\n   <p><font face="arial, helvetica"
size="-1"> \n   <a notrack
href=\'http://clickthru.online.com/click?q=77-u6t6zqpp7m-je7ugalav-
lzjplmsr9rr\'>unsubscribe | \n   <a
href=\'http://nl.com.com/servlet/url_login?email=qqqqqqqqqq-
zdnet@example.com&brand=zdnet\'>manage \n   my
subscriptions | <a
href="http://clickthru.online.com/click?q=8c-ld7aq6jbhwf2vjbvfm8il40awnyr"
>faq | \n   <a
href="http://clickthru.online.com/click?q=a1-afvtq-zkijbdzflgkehaxed3gblr"
>advertise
\n   \n \n \n <!-- /subscription management-->\n
<img src="http://www.zdnet.com/b.gif" width="1" height="8"></td>\n <td
width="12"><img src="http://www.zdnet.com/b.gif" width="12" height="1"></td>\n
<td width="1" bgcolor="#000000"><img src="http://www.zdnet.com/b.gif" width="1"

```
height="1"></td>\n <td width="10" bgcolor="#1e5c99"><img
src="http://www.zdnet.com/b.gif" width="10" height="1"></td>\n <td width="1"
bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif" width="1"
height="1"></td></tr>\n </table>\n <!-- /### subscription management ### -->\n
\n <table width=612 cellpadding=0 cellspacing=0 border=0><tr>\n <td width=1
bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n
<td width=10 bgcolor="#1e5c99"><img src="http://www.zdnet.com/b.gif" width=10
height=1></td>\n <td width=1 bgcolor="#000000"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=588
bgcolor="#000000"><img src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n
<td width=1 bgcolor="#000000"><img src="http://www.zdnet.com/b.gif" width=1
height=1></td>\n <td width=10 bgcolor="#1e5c99"><img
src="http://www.zdnet.com/b.gif" width=10 height=1></td>\n <td width=1
bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n
</tr></table>\n <table width=612 cellpadding=0 cellspacing=0 border=0><tr
bgcolor="#1e5c99">\n <td width=1 bgcolor="#83a3cb"><img
src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n <td width=10><img
src="http://www.zdnet.com/b.gif" width=10 height=1></td>\n <td width=75><a
href="http://clickthru.online.com/click?q=b7-fogpq8jxs54d2ik0npbnptbdl3er" ><img
src="http://www.zdnet.com/graphics/newsletters/gl_ft.gif" width=75 height=43
border=0 alt="zdnet"></a></td>\n <td width=525><font face="arial, helvetica"
size="-2" color="#ffffff"><a
href="http://clickthru.online.com/click?q=cc-h4fqq0oq5ajrrit7qezdozn6f5yr"
style="color: #fff">home</a> |<a
href="http://clickthru.online.com/click?q=e1-6rvnqxk2egkosanphsyk3awi9qlr"
style="color: #fff">ebusiness</a> | <a
href="http://clickthru.online.com/click?q=f6-qktiqmwrvvcrwa6yyrgot9_jmp4r"
style="color: #fff">security</a> | <a
href="http://clickthru.online.com/click?q=0b-0ni-inq4zccsy6xgftmzctvxyhrr"
style="color: #fff">networking</a> | <a
href="http://clickthru.online.com/click?q=20-poifirzzey93bydtr-pp7us2qfir"
style="color: #fff">applications</a> | <a
href="http://clickthru.online.com/click?q=35-d6jtibiziyes2vxknxltsypwc3pr"
style="color: #fff">platforms</a> | <a
href="http://clickthru.online.com/click?q=4a-hcguippcjukdnolsxolkktkhn-dr"
style="color: #fff">hardware</a> | <a
href="http://clickthru.online.com/click?q=5f-wk5biikx_z7yqas6fez6pa6cijur"
style="color: #fff">contact us</a></font></td>\n <td width=1
bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif" width=1 height=1></td>\n
</tr></table>\n <table width=612 cellpadding=0 cellspacing=0 border=0>\n <tr><td
bgcolor="#83a3cb"><img src="http://www.zdnet.com/b.gif" width=1
height=1></td></tr>\n <tr><td><font face="arial, helvetica" size="-2"
color="#ffffff"><table width=100% border=0 cellspacing=2 cellpadding=1> <tr
valign=bottom> <td width=75% height=31> <p></font> <br><b><font face=arial,
helvetica, sans-serif size=2>    copyright 2002 cnet networks, inc. all rights
reserved. zdnet is a registered service mark of cnet networks, inc.
</font></b> </p></td><td height=31 valign=top> <div align=right> <img src="http:
```

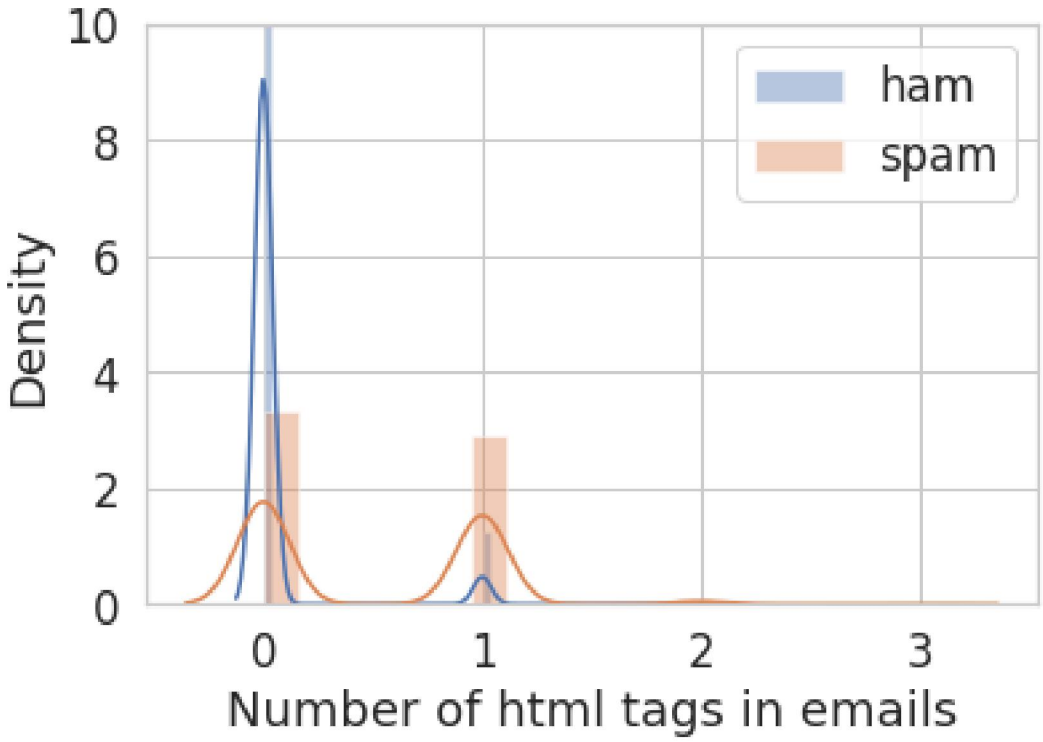
```
//gserv.zdnet.com/clear/outbound.gif?appid=2&emid=25136482&nl=e539&issue=2002-07
-15" height=1 width=1>    </div></td></tr><tr>  <td colspan=2><font face=arial,
helvetica, sans-serif size=2>    </font></td></tr></table></font></td></tr>\n
</table>\n <!-- /footer -->\n \n \n </body>\n </html>\n \n <img height=1 width=1
src="http://clickthru.online.com/click?q=75-whvtngu1zf-4ro34qezuwvjzbdrr">\n \n'
```

[30]:
```
pattern_html = r"<html>"
```

[31]:
```
ham_html_count_list = find_the_count_in_email(ham, pattern_html)
spam_html_count_list = find_the_count_in_email(spam, pattern_html)
```

[32]:
```
sns.distplot(ham_html_count_list, label = 'ham')
sns.distplot(spam_html_count_list, label = 'spam')
plt.xlabel('Number of html tags in emails')
plt.ylim(0, 10)
plt.legend()
```

[32]:
```
<matplotlib.legend.Legend at 0x7f5e783b5be0>
```



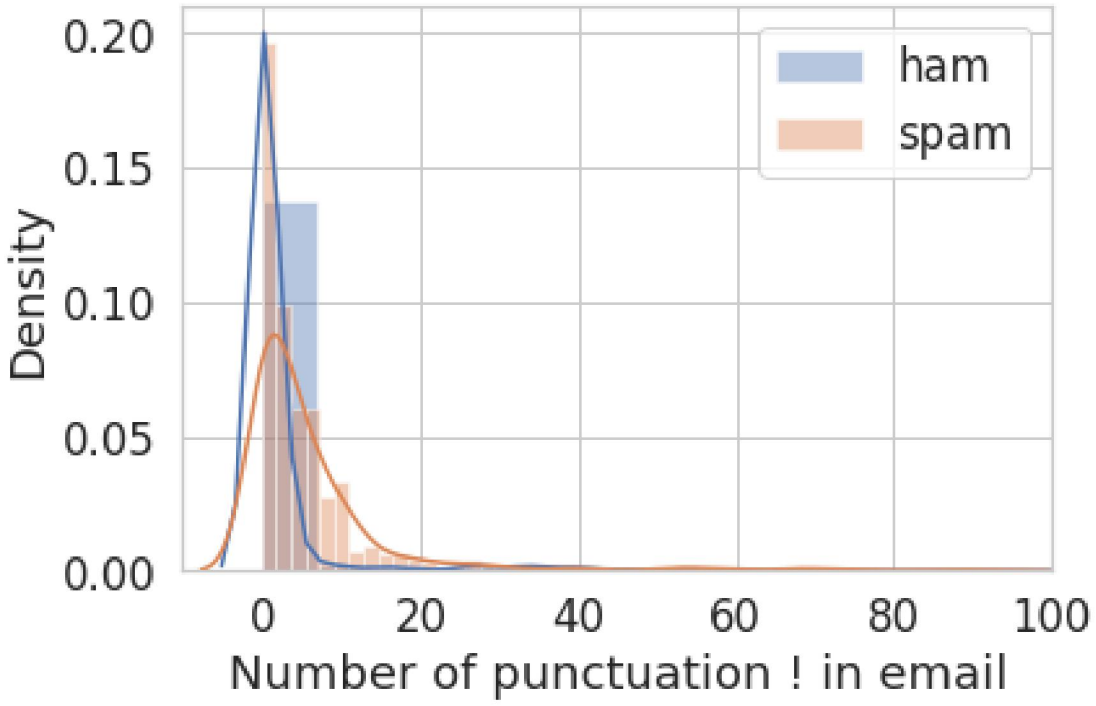
- the use of html tags is significantly less in ham than in spam! Which could be used in the feature selection.

Use of punctuation

[33]:
```
# Use of punctuation in email ('!' )
pattern_punc = r'!'
ham_punc_count_list = find_the_count_in_email(ham, pattern_punc)
spam_punc_count_list = find_the_count_in_email(spam, pattern_punc)
```

[34]:
```
sns.distplot(ham_punc_count_list, label = 'ham')
sns.distplot(spam_punc_count_list, label = 'spam')
plt.xlabel('Number of punctuation ! in email')
plt.xlim(-10, 100)
plt.legend()
```

[34]:
```
<matplotlib.legend.Legend at 0x7f5e782e0070>
```



[35]:
```
sum(spam_punc_count_list) / len(spam['email'])
```

[35]:
```
6.470281543274244
```

[36]:
```
sum(ham_punc_count_list) / len(ham['email'])
```

[36]:
```
2.0875781948168006
```

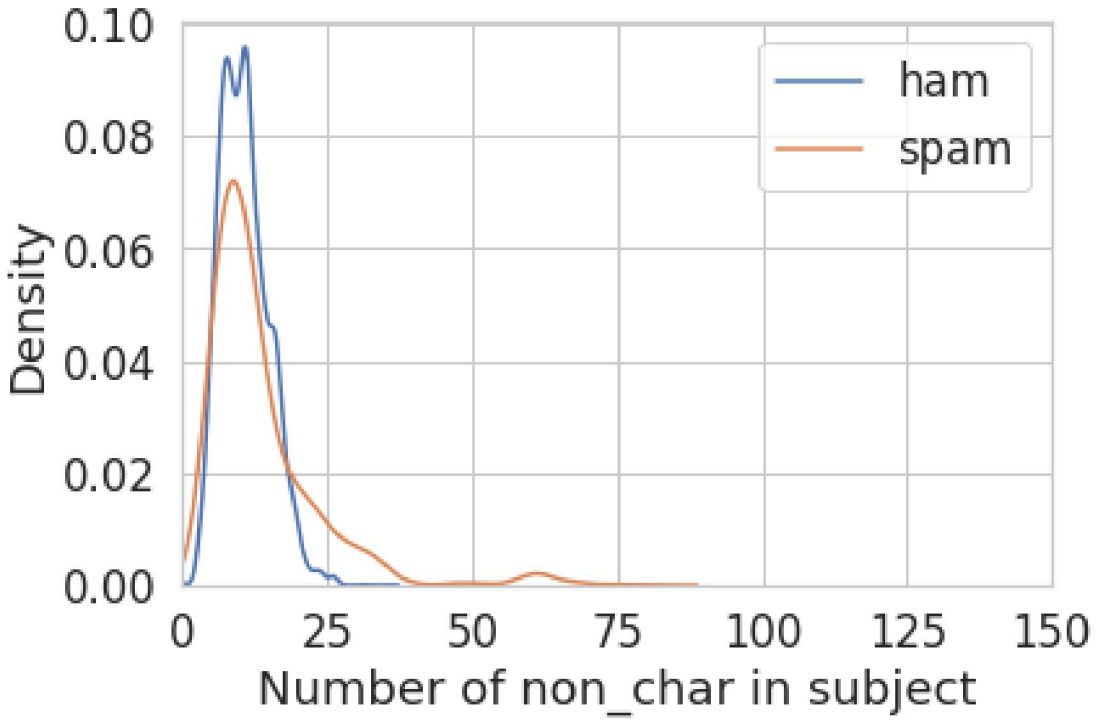
- More(almost 3 folds) ! is used in spam than ham! Which could be used in the final model.

Non character use in subject

[37]:
```
non_char_pattern = r'[^A-Za-z0-9]'
ham_non_char_count_list = find_the_count_in_subject(ham, non_char_pattern)
spam_non_char_count_list = find_the_count_in_subject(spam, non_char_pattern)
```

[38]:
```
sns.distplot(ham_non_char_count_list, label = 'ham', hist=False)
sns.distplot(spam_non_char_count_list, label = 'spam', hist=False)
plt.xlabel('Number of non_char in subject')
plt.xlim(0, 150)
plt.legend()
```

[38]:
```
<matplotlib.legend.Legend at 0x7f5e84f6a970>
```



- noticable difference in spam and ham in using non-characters in subject line, could be useful feature

Number / percentage of capital letters

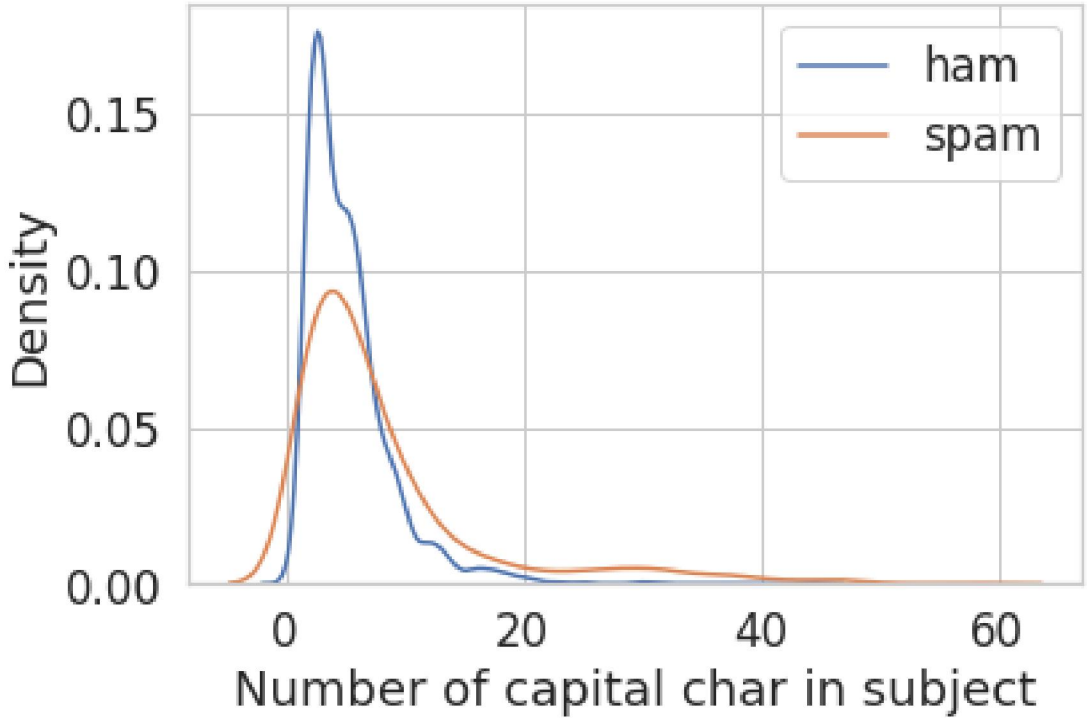
[39]:
```
pattern_cap = r'[A-Z]'
```

[40]:
```
ham_cap_in_subject_count_list = find_the_count_in_subject(ham, pattern_cap)
spam_cap_in_subject_count_list = find_the_count_in_subject(spam, pattern_cap)
```

[41]:
```
sns.distplot(ham_cap_in_subject_count_list, label = 'ham', hist=False)
sns.distplot(spam_cap_in_subject_count_list, label = 'spam', hist=False)
```

```
plt.xlabel('Number of capital char in subject')
#plt.xlim(0, 200)
plt.legend()
```

[41]: <matplotlib.legend.Legend at 0x7f5e780ae430>

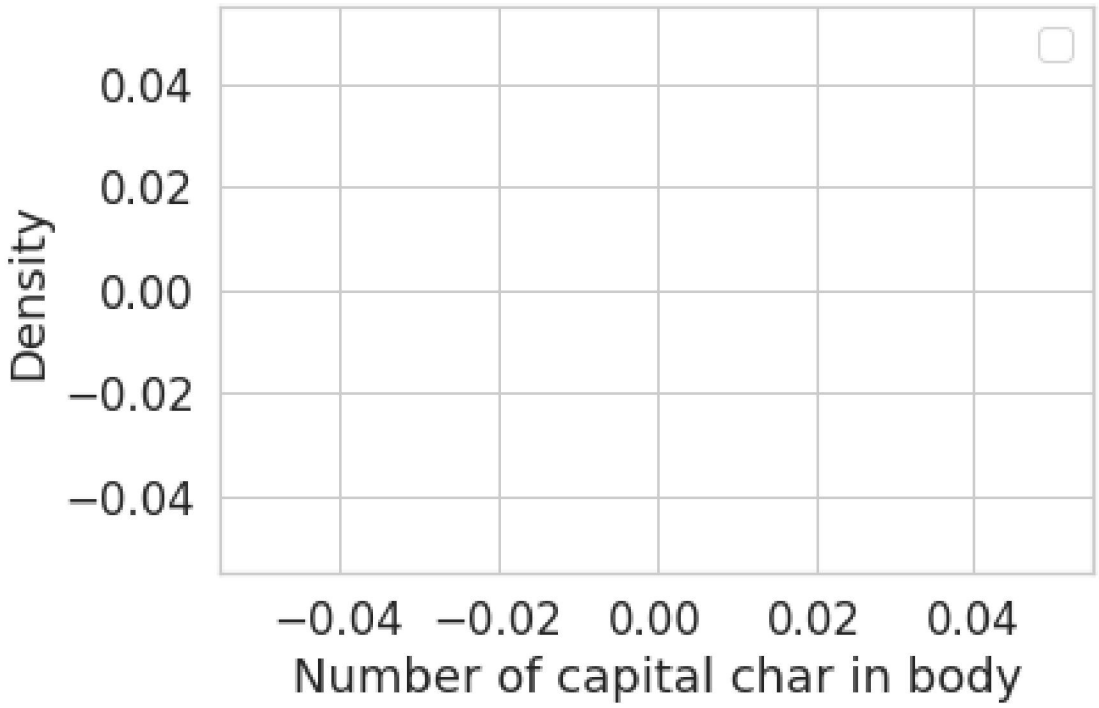


[ ]:
```
ham_cap_in_body_count_list = find_the_count_in_email(ham, pattern_cap)
spam_cap_in_body_count_list = find_the_count_in_email(spam, pattern_cap)
spam_cap_in_body_count_list
```

[43]:
```
sns.distplot(ham_cap_in_body_count_list, label = 'ham', hist=False)
sns.distplot(spam_cap_in_body_count_list, label = 'spam', hist=False)
plt.xlabel('Number of capital char in body')
#plt.xlim(0, 200)
plt.legend()
```

```
/opt/conda/lib/python3.8/site-packages/seaborn/distributions.py:306:
UserWarning: Dataset has 0 variance; skipping density estimate.
  warnings.warn(msg, UserWarning)
/opt/conda/lib/python3.8/site-packages/seaborn/distributions.py:306:
UserWarning: Dataset has 0 variance; skipping density estimate.
  warnings.warn(msg, UserWarning)
No handles with labels found to put in legend.
```

[43]: <matplotlib.legend.Legend at 0x7f5e780df820>

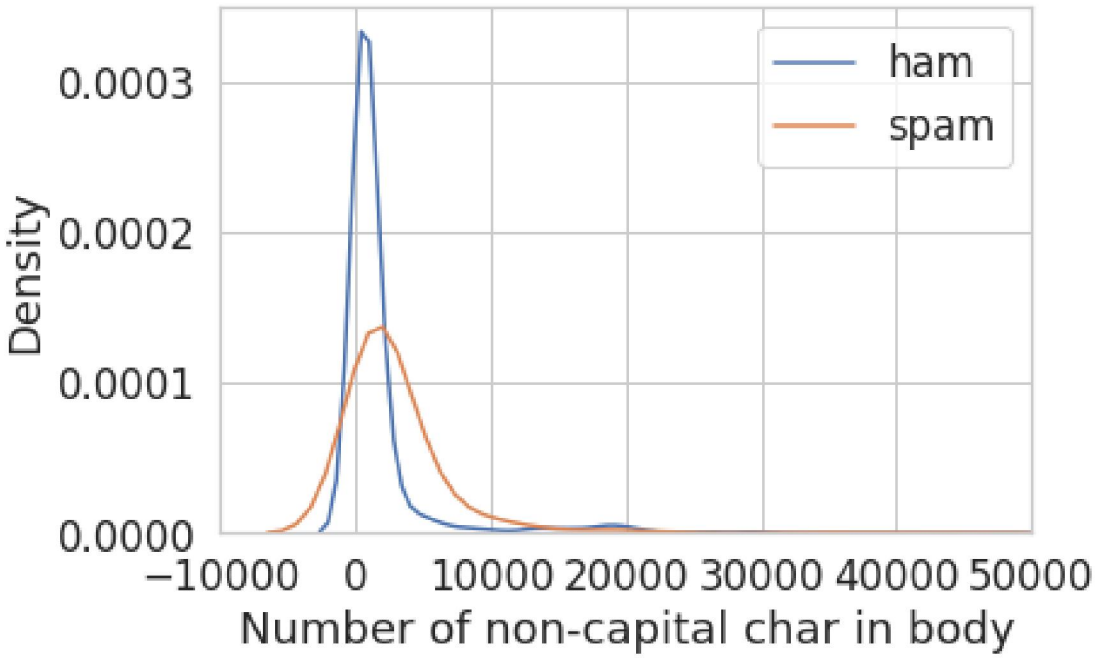


- no capital letter find because we used .lower in the preprocessing

[44]:
```
#non-cap words
pattern_non_cap = r'[a-z]'
ham_no_cap_in_body_count_list = find_the_count_in_email(ham, pattern_non_cap)
spam_no_cap_in_body_count_list = find_the_count_in_email(spam, pattern_non_cap)
```

[45]:
```
sns.distplot(ham_no_cap_in_body_count_list, label = 'ham', hist=False)
sns.distplot(spam_no_cap_in_body_count_list, label = 'spam', hist=False)
plt.xlabel('Number of non-capital char in body')
plt.xlim(-10000, 50000)
plt.legend()
```

[45]: <matplotlib.legend.Legend at 0x7f5e6f789af0>



- the use of capital letters in subject line has a bigger difference in distibution between ham and spam, could be a useful feature

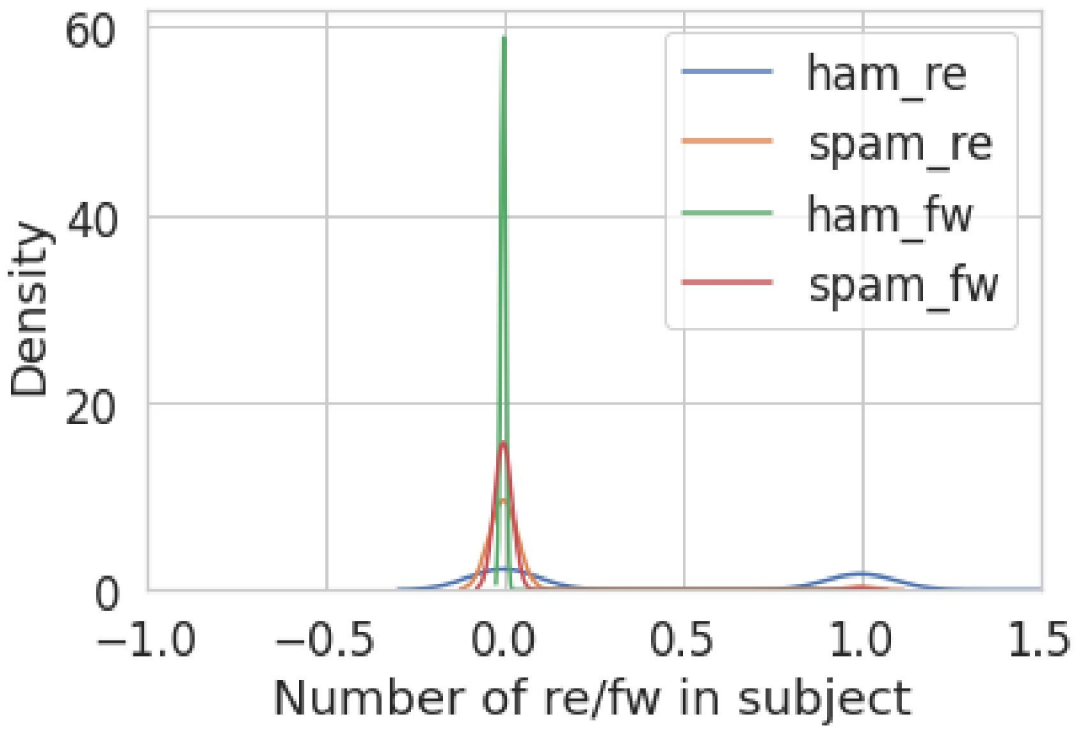
Whether the email is a reply to an earlier email or a forwarded email

[46]:
```
pattern_re = r"Re:"
pattern_forward = r"Fw:"
```

[47]:
```
ham_re_count_list = find_the_count_in_subject(ham, pattern_re)
spam_re_count_list = find_the_count_in_subject(spam, pattern_re)
ham_fw_count_list = find_the_count_in_subject(ham, pattern_forward)
spam_fw_count_list = find_the_count_in_subject(spam, pattern_forward)
```

[48]:
```
sns.distplot(ham_re_count_list, label = 'ham_re', hist=False)
sns.distplot(spam_re_count_list, label = 'spam_re', hist=False)
sns.distplot(ham_fw_count_list, label = 'ham_fw', hist=False)
sns.distplot(spam_fw_count_list, label = 'spam_fw', hist=False)
plt.xlabel('Number of re/fw in subject')
plt.xlim(-1, 1.5)
plt.legend()
```

[48]:
```
<matplotlib.legend.Legend at 0x7f5e78183790>
```



[49]:
```
#proportions in ham emails
print('#ham forward: ', sum(ham_fw_count_list))
print('prop ham forward: ', sum(ham_fw_count_list) / len(ham['subject']))
print('---')
print('#ham reply: ', sum(ham_re_count_list))
print('prop ham reply: ', sum(ham_re_count_list) / len(ham['subject']))
```

```
#ham forward:  8
prop ham forward:  0.0014298480786416443
---
#ham reply:  2663
prop ham reply:  0.47596067917783735
```

[50]:
```
#proportions in spam emails
print('#spam forward: ',sum(spam_fw_count_list))
print('prop spam forward: ', sum(spam_fw_count_list) / len(spam['subject']))
print('---')

print('#spam reply: ',sum(spam_re_count_list))
print('prop spam reply: ', sum(spam_re_count_list) / len(spam['subject']))
```

```
#spam forward:  25
prop spam forward:  0.013034410844629822
---
```

```
#spam reply:  66
prop spam reply:  0.03441084462982273
```

- a noticeable difference between spam and ham eamils, lots of reply in ham, which could be a useful feature in the model.

Dig into the Email words

[51]:
```
def word_bags(df):
    word_count_list = {}

    for email in df['email']:
        word_list = re.findall('\w+', email)
        print(word_list)

        for word in word_list:
            if word in word_count_list:
                word_count_list[word] = (word_count_list[word] + 1)
            else:
                word_count_list[word] =1
    return word_count_list
```

[ ]:
```
spam_word = word_bags(spam)
```

[ ]:
```
#top spammy words
pd.Series(spam_word).sort_values(ascending=False)
```

[ ]:
```
# top training words
train_word_bag = (pd.Series(word_bags(train)) / train.shape[0]).
 ↪sort_values(ascending=False)[:200]
training_words = train_word_bag.index.tolist()
training_words
```

Final Model Feature Selection

[56]:
```
#Final regression model with cv
from sklearn.linear_model import LogisticRegressionCV

def process_data_set(df):
    #get the bag of words
    some_words = ['drug', 'bank', '$', 'free', 'money', '!',
                  'offer', 'business', 'body', 'html', 'please',
                  'account', 'buy', 'gift','good'] + training_words

    #matrix
    X_train = np.array(words_in_texts(some_words, df['email'])).astype(int)

    #feature selections based on the EDA
    feature = pd.concat([
```

```
        #count of characters in email
        find_the_count_in_email(df, pattern_ch),

        #count the html tag using
        find_the_count_in_email(df, pattern_html),

        #count of forward and reply in subject
        find_the_count_in_subject(df, pattern_forward),
        find_the_count_in_subject(df, pattern_re),

        #count of capital letters in subject and body
        find_the_count_in_subject(df, pattern_cap),
        find_the_count_in_email(df, pattern_cap),

        #count of non capital letters in body
        #find_the_count_in_email(df, pattern_non_cap),

        #count the ! usage
        find_the_count_in_email(df, pattern_punc),

        #cound of non characters in subject
        find_the_count_in_subject(df, non_char_pattern)], axis=1).values


    #added features
    X_train = np.concatenate((X_train, feature), axis=1)
    return X_train
```

[57]:
```
import warnings
warnings.filterwarnings('ignore')

#generating X/Y train
X_train = process_data_set(train)
Y_train = train['spam']

#create final logistic regress model with cross validation
model_final = LogisticRegressionCV(Cs=4, fit_intercept=True, verbose =1,␣
 ↪cv=10,random_state=42)
model_final.fit(X_train, Y_train)

#training accuracy
training_accuracy = model_final.score(X_train, Y_train)
print("Training Accuracy: ", training_accuracy)
```

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
```

```
[Parallel(n_jobs=1)]: Done  10 out of  10 | elapsed:   18.0s finished

Training Accuracy:  0.9629974710501797
```

1. How did you find better features for your model?
   - Visulization, compare the changes in distribution base on features in hams and spam emails.
   - And then within the ones that are visually different, I tested it in the final model, compare the training accuracy adding and removing them.
2. What did you try that worked or didn't work?
   - The capital characters in emails doesn't work because the .lower() at beginning.
   - The non_cap characters count doesn't improve model prediction power, so I didn't include it in the final model
   - Most of my intuition about spam emails worked well, for example, the ! mark, the capital letters in subject line, the amount of words in subject line, the non_character in subject and body.
3. What was surprising in your search for good features?
   - looking at the bad of words, some of the top words doesn't appear to be spammy, like the numbers: 3rd, 45399 and such, might be some identification or zip code?

Generate your visualization in the cell below and provide your description in a comment.

[58]:
```
# Write your description (2-3 sentences) as a comment here:
# body and html might co-occur relatively frequently
#bigger variance in for tag html, spams have more '<body>', which might␣
 ↪indicates spam emails tend to be wordier than ham.
# overall ham has smaller distribution in both of these words, or tags␣
 ↪comparing to spam

# Write the code to generate your visualization here:

#spam distribution
x_spam = find_the_count_in_email(spam, r'html')
y_spam = find_the_count_in_email(spam, r'body')

#ham distribution
x_ham = find_the_count_in_email(ham, r'html')
y_ham = find_the_count_in_email(ham, r'body')
```

[59]:
```
plt.rcParams["figure.figsize"] = [10, 5]
```

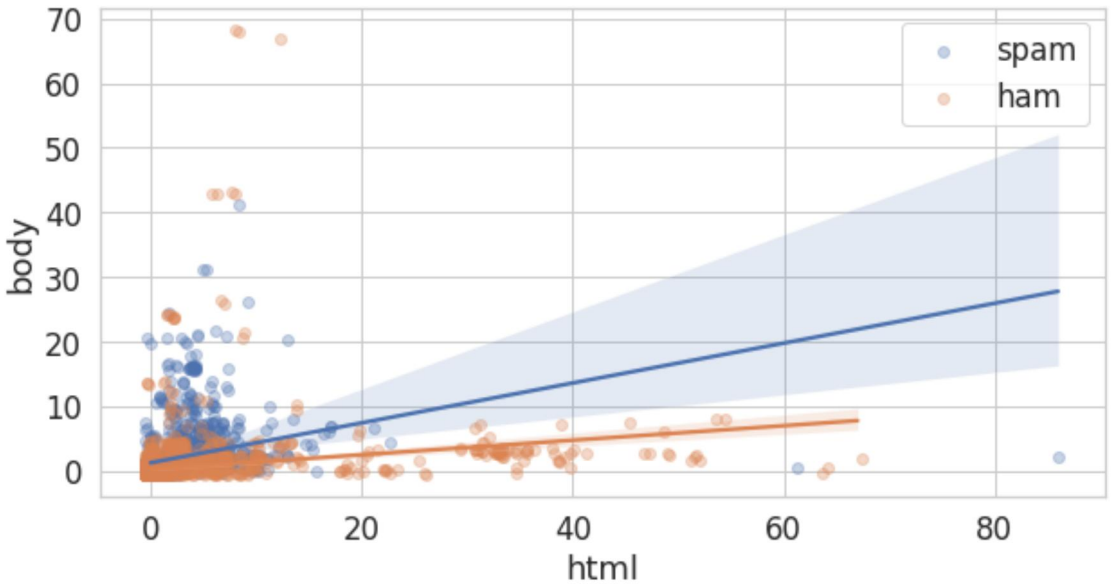
[60]:
```
sns.regplot(x = x_spam,
                 y= y_spam,
                 x_jitter=0.5, y_jitter=0.6, label = "spam", scatter_kws =␣
 ↪{"alpha":0.3})

sns.regplot(x = x_ham,
                 y= y_ham,
```

```
                x_jitter=0.5, y_jitter=0.6, label = "ham", scatter_kws =␣
↪{"alpha":0.3})


plt.legend()
plt.xlabel('html')
plt.ylabel('body')
```

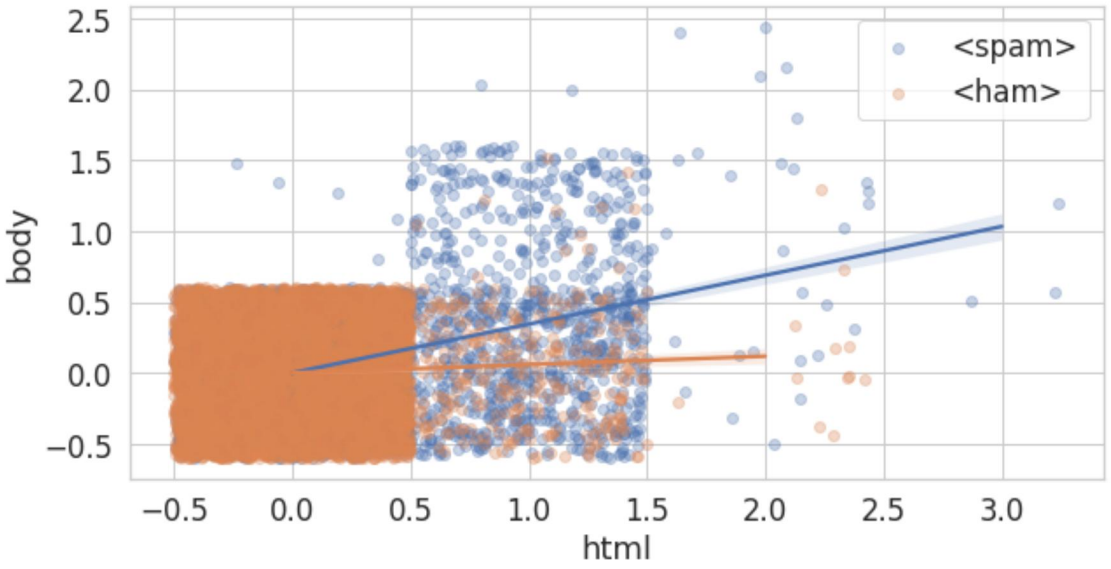
[60]: Text(0, 0.5, 'body')



[61]:
```
sns.regplot(x = find_the_count_in_email(spam, r'<html>'),
                y= find_the_count_in_email(spam, r'<body>'),
                x_jitter=0.5, y_jitter=0.6, label = "<spam>", scatter_kws =␣
↪{"alpha":0.3})

sns.regplot(x = find_the_count_in_email(ham, r'<html>'),
                y= find_the_count_in_email(ham, r'<body>'),
                x_jitter=0.5, y_jitter=0.6, label = "<ham>", scatter_kws =␣
↪{"alpha":0.3})
plt.legend()
plt.xlabel('html')
plt.ylabel('body')
```

[61]: Text(0, 0.5, 'body')



0.0.2 Question 3: ROC Curve

In most cases we won't be able to get 0 false positives and 0 false negatives, so we have to compromise. For example, in the case of cancer screenings, false negatives are comparatively worse than false positives — a false negative means that a patient might not discover that they have cancer until it's too late, whereas a patient can just receive another screening for a false positive.

Recall that logistic regression calculates the probability that an example belongs to a certain class. Then, to classify an example we say that an email is spam if our classifier gives it ≥ 0.5 probability of being spam. However, *we can adjust that cutoff*: we can say that an email is spam only if our classifier gives it ≥ 0.7 probability of being spam, for example. This is how we can trade off false positives and false negatives.

The ROC curve shows this trade off for each possible cutoff probability. In the cell below, plot a ROC curve for your final classifier (the one you use to make predictions for Gradescope) on the training data. Refer to Lecture 20 or Section 23.7 of the course text to see how to plot an ROC curve.

[62]:
```
from sklearn.metrics import roc_curve


# Note that you'll want to use the .predict_proba(...) method for your␣
 ↪classifier
# instead of .predict(...) so you get probabilities, not classes

scores = []

for i in model_final.predict_proba(X_train):
    scores.append(i[1])
```

```
fpr, tpr, thresholds = roc_curve(Y_train, scores, pos_label=1)
```

[63]:
```
plt.step(fpr, tpr, color='b', alpha=0.5,
         where='post')
plt.xlabel('False Positive Rate (1 - Specificity)')
plt.ylabel('Sensitivity')
plt.title('ROC Curve')
```

[63]: Text(0.5, 1.0, 'ROC Curve')