Homework 11

Andrew Lee ml45932

This homework is due on Dec 3 at 11:59pm. Please submit as a PDF or HTML file on Canvas. Before submission, please re-run all cells by clicking "Kernel" and selecting "Restart & Run All."

Problem 1 The following presents you with several real-world strings. Your task is to use regular expressions in python to match (and return) the requested parts in each.

NOTE: these examples are taken in part from a regex game http://play.inginf.units.it/#/level/1 (http://play.inginf.units.it/#/level/1)

I would strongly encourage you to try and play the game first; adapting it to python is as easy as putting your working regex in `re.findall()`

Problem 1a (2 pts): Using `re.findall()` , match all numbers and pull them from the following string (string1). Your results should look like this.

```
['12',
 '47',
 '48',
 '189',
 '2036',
 '314',
 '125',
 '789',
 '1450',
 '564',
 '90456',
 '7890']
```

In [1]:
```
import re

string1="We have to extract these numbers 12, 47, 48 The integers number
s are also interesting: 189 2036 314\',\' is a separator, so please extr
act these numbers 125,789,1450 and also these 564,90456 We like to offer
you 7890$ per month in order to complete this task... we are joking"
print(string1)

re.findall(r'[0-9]+', string1)
```

```
We have to extract these numbers 12, 47, 48 The integers numbers are al
so interesting: 189 2036 314',' is a separator, so please extract these
numbers 125,789,1450 and also these 564,90456 We like to offer you 7890
$ per month in order to complete this task... we are joking
```

Out[1]:
```
['12',
 '47',
 '48',
 '189',
 '2036',
 '314',
 '125',
 '789',
 '1450',
 '564',
 '90456',
 '7890']
```

Problem 1b (2 pts): Using `re.findall()` , match all IP addressses in this string (string2). Your results should look like this:

```
['213.92.153.167',
 '69.43.107.219',
 '69.43.112.233',
 '217.70.100.113',
 '74.125.186.208',
 '74.125.186.208',
 '140.105.63.158',
 '172.45.240.237',
 '217.70.177.60',
 '216.34.90.16',
 '69.43.85.253',
 '213.121.184.130',
 '213.121.184.130',
 '140.105.63.164']
```

In [2]:
```
string2="Jan 13 00:48:59: DROP service 68->67(udp) from 213.92.153.167 t
o 69.43.107.219, prefix: \"spoof iana-0/8\" \
(in: eth0 69.43.112.233(38:f8:b7:90:45:92):68 -> 217.70.100.113(00:21:8
7:79:9c:d9):67 UDP len:576 ttl:64) \
Jan 13 12:02:48: ACCEPT service dns from 74.125.186.208 to firewall(pub-
nic-dns), prefix: \"none\" \
(in: eth0 74.125.186.208(00:1a:e3:52:5d:8e):36008 -> 140.105.63.158(00:1
a:9a:86:2e:62):53 UDP len:82 ttl:38) \
Jan 13 17:44:52: DROP service 68->67(udp) from 172.45.240.237 to 217.70.
177.60, prefix: \"spoof iana-0/8\" \
(in: eth0 216.34.90.16(00:21:91:fe:a2:6f):68 -> 69.43.85.253(00:07:e1:7
c:53:db):67 UDP len:328 ttl:64) \
Jan 13 17:52:08: ACCEPT service http from 213.121.184.130 to firewall(pu
b-nic), prefix: \"none\" \
(in: eth0 213.121.184.130(00:05:2e:6a:a4:14):8504 -> 140.105.63.164(00:6
0:11:92:ed:1b):80 TCP flags: ****S* len:52 ttl:109)"

re.findall(r'[0-9]+(?:\.[0-9]+){3}', string2)
```

Out[2]:
```
['213.92.153.167',
 '69.43.107.219',
 '69.43.112.233',
 '217.70.100.113',
 '74.125.186.208',
 '74.125.186.208',
 '140.105.63.158',
 '172.45.240.237',
 '217.70.177.60',
 '216.34.90.16',
 '69.43.85.253',
 '213.121.184.130',
 '213.121.184.130',
 '140.105.63.164']
```

Problem 1c (2 pts): Using `re.findall()` , match all MAC addressses in the same string (string2). Your results should look like this:

```
['38:f8:b7:90:45:92',
 '00:21:87:79:9c:d9',
 '00:1a:e3:52:5d:8e',
 '00:1a:9a:86:2e:62',
 '00:21:91:fe:a2:6f',
 '00:07:e1:7c:53:db',
 '00:05:2e:6a:a4:14',
 '00:60:11:92:ed:1b']
```

In [3]:
```
re.findall(r'(?:[0-9a-fA-F]:?){12}', string2)
```

Out[3]:
```
['38:f8:b7:90:45:92',
 '00:21:87:79:9c:d9',
 '00:1a:e3:52:5d:8e',
 '00:1a:9a:86:2e:62',
 '00:21:91:fe:a2:6f',
 '00:07:e1:7c:53:db',
 '00:05:2e:6a:a4:14',
 '00:60:11:92:ed:1b']
```

Problem 1d (2 pts): Using `re.findall()` , match all ftp addresses in the string below (string3). Your results should look like this:

```
['ftp://ftp7.br.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp3.de.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp.is.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp4.jp.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp.no.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp3.no.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp.pt.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp1.ro.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp3.es.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp2.tw.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp6.uk.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp6.us.FreeBSD.org/pub/FreeBSD/']
```

In [4]:
```
string3=r"Fedora Core           ftp     \
Fedora Extras   http    ftp     rsync\
          ftp://ftp7.br.FreeBSD.org/pub/FreeBSD/ (ftp)\
          ftp://ftp3.de.FreeBSD.org/pub/FreeBSD/ (ftp)\
          ftp://ftp.is.FreeBSD.org/pub/FreeBSD/ (ftp / rsync)\
          ftp://ftp4.jp.FreeBSD.org/pub/FreeBSD/ (ftp)\
          ftp://ftp.no.FreeBSD.org/pub/FreeBSD/ (ftp / rsync)\
        *\
          ftp://ftp3.no.FreeBSD.org/pub/FreeBSD/ (ftp)\
          ftp://ftp.pt.FreeBSD.org/pub/FreeBSD/ (ftp)\
          ftp://ftp1.ro.FreeBSD.org/pub/FreeBSD/ (ftp / ftpv6)\
          ftp://ftp3.es.FreeBSD.org/pub/FreeBSD/ (ftp)\
          ftp://ftp2.tw.FreeBSD.org/pub/FreeBSD/ (ftp / ftpv6 / http / h
ttpv6 / rsync / rsyncv6)\
          ftp://ftp6.uk.FreeBSD.org/pub/FreeBSD/ (ftp)\
          ftp://ftp6.us.FreeBSD.org/pub/FreeBSD/ (ftp / http)"

re.findall(r'(ftp?://[^\s]+)', string3)
```

Out[4]:
```
['ftp://ftp7.br.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp3.de.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp.is.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp4.jp.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp.no.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp3.no.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp.pt.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp1.ro.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp3.es.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp2.tw.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp6.uk.FreeBSD.org/pub/FreeBSD/',
 'ftp://ftp6.us.FreeBSD.org/pub/FreeBSD/']
```

Problem 1e (2 pts): Using `re.findall()`, match all latex math-mode text (anything wrapped in `$`, including the `$`s) in the string below (string4). Your results should look like this:

```
['$\\mu_{A_T}$',
 '$\\sigma_{A_T}$',
 '$A_T$',
 '$\\rho_\\text{filler}$',
 '$\\rho_\\text{filler}=\\frac{\\sum_{T \\in \\mathcal{T} A_T}}{A}$',
 '$A$',
 '$\\mu_{C_T}$',
 '$C_T$',
 '$C_T = \\sigma_R+\\sigma_G+\\sigma_B$',
 '$\\sigma_R$',
 '$\\sigma_G$',
 '$\\sigma_B$',
 '$T$',
 '$P$',
 '$R$',
 '$C$']
```

In [5]:
```
string4=r"We try to quantitatively capture these characteristics by defi
ning a set of indexes,\
which can be computed using the mosaic image and the corresponding groun
d truth: \
\begin{itemize} \
    \item $\mu_{A_T}$ and $\sigma_{A_T}$, the mean and standard deviatio
n of the tiles area $A_T$, respectively; \
    \item $\rho_\text{filler}$, the ratio between the filler area and th
e overall mosaic are, computed as \
$\rho_\text{filler}=\frac{\sum_{T \in \mathcal{T} A_T}}{A}$, being $A$ t
he area of the mosaic; \
    \item \todo{does it worth?}; \
    \item \todo{does it worth?}; \
    \item $\mu_{C_T}$, the mean of the tiles \emph{color dispersion} $C_
T$, \
being $C_T = \sigma_R+\sigma_G+\sigma_B$, where $\sigma_R$, $\sigma_G$ a
nd $\sigma_B$ are the \
standard deviation of the red, green and blue channel values of the pixe
ls within the tile $T$.\
After applying a method to an image, we compare the segmented image (i.
e., the result) \
against the ground truth and assess the performance according to the fol
lowing three metrics: \
\begin{itemize} \
    \item average tile precision $P$ \
    \item average tile recall $R$ \
    \item tile count error $C$"


re.findall(r'\$.*?\$', string4)
```

Out[5]:
```
['$\\mu_{A_T}$',
 '$\\sigma_{A_T}$',
 '$A_T$',
 '$\\rho_\\text{filler}$',
 '$\\rho_\\text{filler}=\\frac{\\sum_{T \\in \\mathcal{T} A_T}}{A}$',
 '$A$',
 '$\\mu_{C_T}$',
 '$C_T$',
 '$C_T = \\sigma_R+\\sigma_G+\\sigma_B$',
 '$\\sigma_R$',
 '$\\sigma_G$',
 '$\\sigma_B$',
 '$T$',
 '$P$',
 '$R$',
 '$C$']
```

Problem 1f (2 pts): Using `re.findall()`, match all text of the form href="..." in the string below (string5). Your results should look like this:

```
['href="javascript:openurl(\'/Xplore/accessinfo.jsp\')"',
 'href="/iel5/4235/4079606/04079617.pdf?tp=&arnumber=4079617&isnumber=4079606"',
 "href='/xpl/RecentCon.jsp?punumber=10417'",
 'href="/xplorehelp/Help_start.html#Help_searchresults.html"',
 'href="/xpl/contactus.jsp"',
 'href="http://search.epnet.com/login.asp?profile=web&amp;defaultdb=geh"',
 'href="http://iimpft.chadwyck.com/"',
 'href="standartlar.html#tse"',
 'href="http://www.gutenberg.org/"',
 'href="http://proquestcombo.safaribooksonline.com/?
portal=proquestcombo&amp;uicode=istanbultek"',
 'href="http://www.scitation.org"',
 'href="/online/aip.html"',
 'href="http://www3.interscience.wiley.com/journalfinder.html"',
 'href="/xpl/periodicals.jsp"',
 'href="http://www.ieee.org/products/onlinepubs/resources/XploreTutorial.pdf"']
```

In [6]:
```
string5="<a href=\"javascript:openurl('/Xplore/accessinfo.jsp')\" class=
\"topUnderlineLinks\">\
                                                <A href=\"/iel5/4235/407960
6/04079617.pdf?tp=&arnumber=4079617&isnumber=4079606\" class=\"bodyCopy
\">PDF</A>(3141 KB) \
                          <A href='/xpl/RecentCon.jsp?punumber=10417'>Evol
utionary Computation, 2005. The 2005 IEEE Congress on</A><br>\
                 <td width=\"33%\" ><div align=\"right\"> <a href=\"/xplo
rehelp/Help_start.html#Help_searchresults.html\" class=\"subNavLinks\" t
arget=\"blank\">Help</a>   <a href=\"/xpl/contactus.jsp\"
class=\"subNavLinks\">Contact\
                 Kimya ile ilgili çeþitli temel referans\
<a href=\"http://search.epnet.com/login.asp?profile=web&amp;defaultdb=ge
h\"\
<a href=\"http://iimpft.chadwyck.com/\" target=\"_parent\">International
\
<a href=\"standartlar.html#tse\" target=\"_parent\">NFPA Standartlarý</a
>\
<a href=\"http://www.gutenberg.org/\" target=\"_parent\">Project Gutenbe
rg</a>\
<a href=\"http://proquestcombo.safaribooksonline.com/?portal=proquestcom
bo&amp;uicode=istanbultek\"\
<a href=\"http://www.scitation.org\" target=\"_parent\">Scitation</a>\
dergilerin listesini görmek için <a href=\"/online/aip.html\">bu yolu</a
>\
<a href=\"http://www3.interscience.wiley.com/journalfinder.html\"\
                <td width=\"46%\"><a href=\"/xpl/periodicals.jsp\" class=
\"dropDownNav\" accesskey=\"j\">Journals &amp; Magazines\
                <td><a href=\"http://www.ieee.org/products/onlinepubs/res
ources/XploreTutorial.pdf\" class=\"dropDownNav\">IEEE Xplore Demo</a></
td>"


re.findall(r'href=".*?"', string5)
```

Out[6]:
```
['href="javascript:openurl(\'/Xplore/accessinfo.jsp\')"',
 'href="/iel5/4235/4079606/04079617.pdf?tp=&arnumber=4079617&isnumber=4
079606"',
 'href="/xplorehelp/Help_start.html#Help_searchresults.html"',
 'href="/xpl/contactus.jsp"',
 'href="http://search.epnet.com/login.asp?profile=web&amp;defaultdb=ge
h"',
 'href="http://iimpft.chadwyck.com/"',
 'href="standartlar.html#tse"',
 'href="http://www.gutenberg.org/"',
 'href="http://proquestcombo.safaribooksonline.com/?portal=proquestcomb
o&amp;uicode=istanbultek"',
 'href="http://www.scitation.org"',
 'href="/online/aip.html"',
 'href="http://www3.interscience.wiley.com/journalfinder.html"',
 'href="/xpl/periodicals.jsp"',
 'href="http://www.ieee.org/products/onlinepubs/resources/XploreTutoria
l.pdf"']
```

Problem 1g (2 pts): Using `re.findall()`, match all urls in the string below (string6). Your results should look like this:

```
['http://www.classmates.com/go/e/200988231/CC123101BT/CM00',
 'http://graphics.classmates.com/graphics/spacer.gif',
 'http://graphics.classmates.com/graphics/sp',
 'http://itcapps.corp.enron.com/srrs/auth/emailLink.asp?
ID=000000000053239&Page=Approval',
 'http://www.enrononline.com',
 'http://www.classmates.com/go/e/200988231/CC122401BC/CM00',
 'http://graphics.classmates.com/graphics/spacer.gif',

'http://graphics.classmates.com/graphics/sphttp://www.btinternet.com/~pir8/arnie/n,'

 'http://zzz1.net/rd/rd.asp?ZXU=562&ZXD=1471085&UID=1471085',
 'http://www.egroups.com',
 'http://isc.enron.com/site/doclibrary/user/',
 'http://esource.enron.com/worldmarket.asp',
 'http://esource.enron.com/worldmarket_CountryAnalysis.asp',
 'http://ad.doubleclick.net/clk;3549492;6600300;c?
http://www.sportingbetusa.com/english/casino/casinonew-fr.asp?isLogged=notlogged',
 'http://ad.doubleclick.net/clk;3549492;6600300;c?http://www.sportingbetusa.c',
 'http://isc.enron.com/site/']
```

In [7]:
```
string6="<http://www.classmates.com/go/e/200988231/CC123101BT/CM00>  <ht
tp://graphics.classmates.com/graphics/spacer.gif>  <http://graphics.clas
smates.com/graphics/sp \
You have received this email because the requester specified you as thei
r Manager. Please click http://itcapps.corp.enron.com/srrs/auth/emailLin
k.asp?ID=000000000053239&Page=Approval to review and act upon this reque
st.     Request ID          : 000000000053239 Request Create Date\
ronOnline.   The following User ID and Password will give you access to
 live prices on the web-site http://www.enrononline.com.  User ID: ADM40
601 Password: WELCOME!   (note these are case sensitive)    Please keep
 your User I\
<http://www.classmates.com/go/e/200988231/CC122401BC/CM00>  <http://grap
hics.classmates.com/graphics/spacer.gif>  <http://graphics.classmates.co
m/graphics/sp\
http://www.btinternet.com/~pir8/arnie/\
n, just click on the following hyperlink and complete the order form by
 Tuesday February 12, 2002.  http://zzz1.net/rd/rd.asp?ZXU=562&ZXD=14710
85&UID=1471085  If you cannot link directly to the web site, simply cut
 and paste the address listed above into yo\
been successful getting in the group. To access the group should go to y
our web browser and type in http://www.egroups.com  The screen should sh
ow that you are a member of smu-betas group. When you replied to the ori
ginal \
mber and password. For more details on how to log-on to eHRonline, see s
tep-by-step instructions at http://isc.enron.com/site/doclibrary/user/
 2. Navigate to the pay advice using the following navigation menus: ? P
ay Information ? Paycheck I\
In addition to World Markets Energy information  <http://esource.enron.c
om/worldmarket.asp> and Country Analysis and Forecasting, <http://esourc
e.enron.com/worldmarket_CountryAnalysis.asp>  \
<http://ad.doubleclick.net/clk;3549492;6600300;c?http://www.sportingbetu
sa.com/english/casino/casinonew-fr.asp?isLogged=notlogged> A WEEKEND PAI
R-A-DICE <http://ad.doubleclick.net/clk;3549492;6600300;c?http://www.spo
rtingbetusa.c \
Mr. Skilling:  Your P number is P00500599.  For your convenience, you ca
n also go to http://isc.enron.com/site/ under"

re.findall(r'(http?://[^\s]+)', string6)
```

Out[7]:
```
['http://www.classmates.com/go/e/200988231/CC123101BT/CM00>',
 'http://graphics.classmates.com/graphics/spacer.gif>',
 'http://graphics.classmates.com/graphics/sp',
 'http://itcapps.corp.enron.com/srrs/auth/emailLink.asp?ID=000000000053
239&Page=Approval',
 'http://www.enrononline.com.',
 'http://www.classmates.com/go/e/200988231/CC122401BC/CM00>',
 'http://graphics.classmates.com/graphics/spacer.gif>',
 'http://graphics.classmates.com/graphics/sphttp://www.btinternet.com/~
pir8/arnie/n,',
 'http://zzz1.net/rd/rd.asp?ZXU=562&ZXD=1471085&UID=1471085',
 'http://www.egroups.com',
 'http://isc.enron.com/site/doclibrary/user/',
 'http://esource.enron.com/worldmarket.asp>',
 'http://esource.enron.com/worldmarket_CountryAnalysis.asp>',
 'http://ad.doubleclick.net/clk;3549492;6600300;c?http://www.sportingbe
tusa.com/english/casino/casinonew-fr.asp?isLogged=notlogged>',
 'http://ad.doubleclick.net/clk;3549492;6600300;c?http://www.sportingbe
tusa.c',
 'http://isc.enron.com/site/']
```

Question 2 (3 pts): In the following string (string7), using `re.findall()`, match restriction enzyme binding sites ANTAAT and GCRWTG.

Note that per the IUPAC nucleotide code, N is any base, R is A or G, W is A or T

How many cuts total in the sequence do you expect if you digest with both of these restriction enzymes? How many fragments do you expect?

In [8]:
```
string7="ATGGCAATAACCCCCCGTTTCTACTTCTAGAGGAGAAAAGTATTGACATGAGCGCTCCCGGCA
CAAGGGCCAAAGAAGTCTCCAATTTCTTATTTCCGAATGACATGCGTCTCCTTGCGGGTAAATCACCGACCG
CAATTCATAGAAGCCTGGGGGAACAGATAGGTCTAATTAGCTTAAGAGAGTAAATCCTGGGATCATTCAGTA
GTAACCATAAACTTACGCTGGGGCTTCTTCGGCGGATTTTTACAGTTACCAACCAGGAGATTTGAAGTAAAT
CAGTTGAGGATTTAGCCGCGCTATCCGGTAATCTCCAAATTAAAACATACCGTTCCATGAAGGCTAGAATTA
CTTACCGGCCTTTTCCATGCCTGCGCTATACCCCCCCACTCTCCCGCTTATCCGTCCGAGCGGAGGCAGTGC
GATCCTCCGTTAAGATATTCTTACGTGTGACGTAGCTATGTATTTTGCAGAGCTGGCGAACGCGTTGAACAC
TTCACAGATGGTAGGGATTCGGGTAAAGGGCGTATAATTGGGGACTAACATAGGCGTAGACTACGATGGCGC
CAACTCAATCGCAGCTCGAGCGCCCTGAATAACGTACTCATCTCAACTCATTCTCGGCAATCTACCGAGCGA
CTCGATTATCAACGGCTGTCTAGCAGTTCTAATCTTTTGCCAGCATCGTAATAGCCTCCAAGAGATTGATGA
TAGCTATCGGCACAGAACTGAGACGGCGCCGATGGATAGCGGACTTTCGGTCAACCACAATTCCCCACGGGA
CAGGTCCTGCGGTGCGCATCACTCTGAATGTACAAGCAACCCAAGTGGGCCGAGCCTGGACTCAGCTGGTTC
CTGCGTGAGCTCGAGACTCGGGATGACAGCTCTTTAAACATAGAGCGGGGGCGTCGAACGGTCGAGAAAGTC
ATAGTACCTCGGGTACCAACTTACTCAGGTTATTGCTTGAAGCTGTACTATTTTAGGGGGGGAGCGCTGAAG
GTCTCTTCTTCTCATGACTGAACTCGCGAGGGTCGTGAAGTCGGTTCCTTCAATGGTTAAAAAACAAAGGCT
TACTGTGCGCAGAGGAACGCCCATCTAGCGGCTGGCGTCTTGAATGCTCGGTCCCCTTTGTCATTCCGGATT
AATCCATTTCCCTCATTCACGAGCTTGCGAAGTCTACATTGGTATATGAATGCGACCTAGAAGAGGGCGCTT
AAAATTGGCAGTGGTTGATGCTCTAAACTCCATTTGGTTTACTCGTGCATCACCGCGATAGGCTGACAAAGG
TTTAACATTGAATAGCAAGGCACTTCCGGTCTCAATGAACGGCCGGGAAAGGTACGCGCGCGGTATGGGAGG
ATCAAGGGGCCAATAGAGAGGCTCCTCTCTCACTCGCTAGGAGGCAAATGTAAAACAATGGTTACTGCATCG
ATACATAAAACATGTCCATCGGTTGCCCAAAGTGTTAAGTGTCTATCACCCCTAGGGCCGTTTCCCGCATAT
AAACGCCAGGTTGTATCCGCATTTGATGCTACCGTGGATGAGTCTGCGTCGAGCGCGCCGCACGAATGTTGC
AATGTATTGCATGAGTAGGGTTGACTAAGAGCCGTTAGATGCGTCGCTGTACTAATAGTTGTCGACAGACCG
TCGAGATTAGAAAATGGTACCAGCATTTTCGGAGGTTCTCTAACTAGTATGGATTGCGGTGTCTTCACTGTG
CTGCGGCTACCCATCGCCTGAAATCCAGCTGGTGTCAAGCCATCCCCTCTCCGGGACGCCGCATGTAGTGAA
ACATATACGTTGCACGGGTTCACCGCGGTCCGTTCTGAGTCGACCAAGGACACAATCGAGCTCCGATCCGTA
CCCTCGACAAACTTGTACCCGACCCCCGGAGCTTGCCAGCTCCTCGGGTATCATGGAGCCTGTGGTTCATCG
CGTCCGATATCAAACTTCGTCATGATAAAGTCCCCCCCTCGGGAGTACCAGAGAAGATGACTACTGAGTTGT
GCGAT"

re.findall(r"A[ATGC]TAAT|GC[AG][AT]TG", string7)
#Four cuts in the sequence and five fragments.
```

Out[8]:
```
['GCGTTG', 'ATTAAT', 'GCAATG', 'ACTAAT']
```

Question 2-OPTIONAL BONUS (2 pts): This one will be difficult, so save it for the end! Note that I will not be giving help on the bonus: you are on your own if you attempt it!

Assume the restriction enzymes cut the sequence (string7, above) at the midpoint of the binding site \, so ANT\ATT and GCR\WTG. Using `re.split()`, cut the sequence at the cut points to digest the sequence, yielding the correct fragments. You might find that two separate `re.split()` commands is the easier way to go, but this my require a loop. Once you have the correct fragments, then use the `count_chars()` function from class to count the number of each base in each fragment. Modify it to additionally report the total length of each fragment.

In [9]:
```
# your code here
```

Question 3 (3 pts)

Take the following paragraph (string8) and remove punctuation marks `[.—,;']` with `re.sub()`. That is, replace them with `""`.

Split resulting string into a list of words with `re.split()` (split at any whitespace character)

Loop through the resulting list and make every word lower-case with `.lower()`. You can do this in several ways: create an empty list outside of your loop and then use `.append()` inside, for example.

Apply the `count_chars()` function from class to the resulting lower-case list: notice that it counts the words for you!

In [10]:
```
string8="Call me Ishmael. Some years ago — never mind how long precisely
— having little or no money in my purse, and nothing particular to inter
est me on shore, I thought I would sail about a little and see the water
y part of the world. It is a way I have of driving off the spleen, and r
egulating the circulation. Whenever I find myself growing grim about the
mouth; whenever it is a damp, drizzly November in my soul; whenever I fi
nd myself involuntarily pausing before coffin warehouses, and bringing u
p the rear of every funeral I meet; and especially whenever my hypos get
such an upper hand of me, that it requires a strong moral principle to p
revent me from deliberately stepping into the street, and methodically k
nocking people’s hats off — then, I account it high time to get to sea a
s soon as I can. This is my substitute for pistol and ball. With a philo
sophical flourish Cato throws himself upon his sword; I quietly take to
 the ship. There is nothing surprising in this. If they but knew it, alm
ost all men in their degree, some time or other, cherish very nearly the
same feelings towards the ocean with me."

string9=re.sub(r"[.\—\,;']",'\"',string8)

stringlist=re.split(r'\s',string9)
print(stringlist)
```

```
['Call', 'me', 'Ishmael"', 'Some', 'years', 'ago', '"', 'never', 'min
d', 'how', 'long', 'precisely', '"', 'having', 'little', 'or', 'no', 'm
oney', 'in', 'my', 'purse"', 'and', 'nothing', 'particular', 'to', 'int
erest', 'me', 'on', 'shore"', 'I', 'thought', 'I', 'would', 'sail', 'ab
out', 'a', 'little', 'and', 'see', 'the', 'watery', 'part', 'of', 'th
e', 'world"', 'It', 'is', 'a', 'way', 'I', 'have', 'of', 'driving', 'of
f', 'the', 'spleen"', 'and', 'regulating', 'the', 'circulation"', 'When
ever', 'I', 'find', 'myself', 'growing', 'grim', 'about', 'the', 'mout
h"', 'whenever', 'it', 'is', 'a', 'damp"', 'drizzly', 'November', 'in',
'my', 'soul"', 'whenever', 'I', 'find', 'myself', 'involuntarily', 'pau
sing', 'before', 'coffin', 'warehouses"', 'and', 'bringing', 'up', 'th
e', 'rear', 'of', 'every', 'funeral', 'I', 'meet"', 'and', 'especiall
y', 'whenever', 'my', 'hypos', 'get', 'such', 'an', 'upper', 'hand', 'o
f', 'me"', 'that', 'it', 'requires', 'a', 'strong', 'moral', 'principl
e', 'to', 'prevent', 'me', 'from', 'deliberately', 'stepping', 'into',
'the', 'street"', 'and', 'methodically', 'knocking', 'people’s', 'hat
s', 'off', '"', 'then"', 'I', 'account', 'it', 'high', 'time', 'to', 'g
et', 'to', 'sea', 'as', 'soon', 'as', 'I', 'can"', 'This', 'is', 'my',
'substitute', 'for', 'pistol', 'and', 'ball"', 'With', 'a', 'philosophi
cal', 'flourish', 'Cato', 'throws', 'himself', 'upon', 'his', 'sword"',
'I', 'quietly', 'take', 'to', 'the', 'ship"', 'There', 'is', 'nothing',
'surprising', 'in', 'this"', 'If', 'they', 'but', 'knew', 'it"', 'almos
t', 'all', 'men', 'in', 'their', 'degree"', 'some', 'time', 'or', 'othe
r"', 'cherish', 'very', 'nearly', 'the', 'same', 'feelings', 'towards',
'the', 'ocean', 'with', 'me"']
```

In [16]:
```
def count_chars(s):
    s=s.lower()
    L = [s.count(c) for c in s]
    return max(L)

print (count_chars(stringlist))
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-16-14b9460b4548> in <module>
      4     return max(L)
      5
----> 6 print (count_chars(stringlist))

<ipython-input-16-14b9460b4548> in count_chars(s)
      1 def count_chars(s):
----> 2     s=stringlist.lower()
      3     L = [s.count(c) for c in s]
      4     return max(L)
      5

AttributeError: 'list' object has no attribute 'lower'
```

In [17]:
```
len(stringlist)
```

Out[17]: 204

In [ ]:
```

```