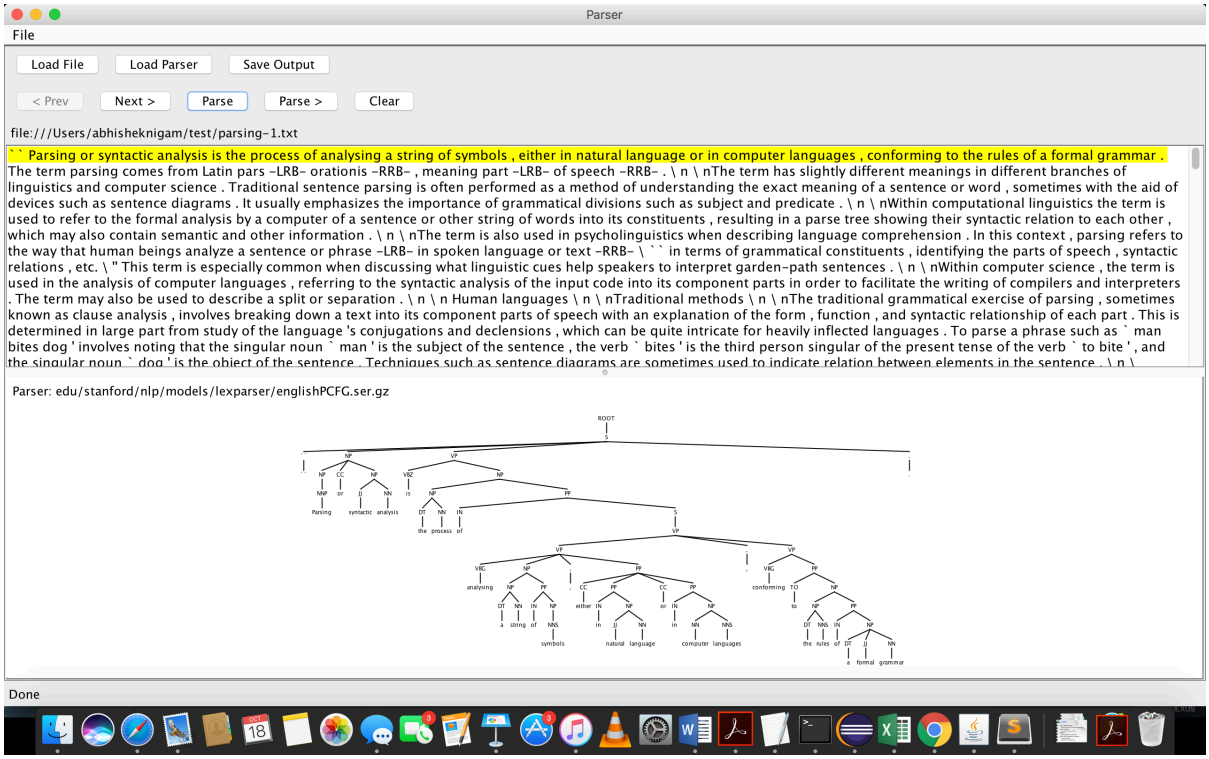
Question 1 :- Write code to extract the actual content of the current version of a Wikipedia page in the missing code.

```
import requests

title='parsing'

response = requests.get("http://en.wikipedia.org/w/api.php?format=json&
action=query&titles="+str(title)+"&prop=revisions&rvprop=content")


jsondata = response.json()


content = jsondata['query']['pages'].values()[0]['revisions'][-1]

The above line gets the current version.
```

Question 2 :- Load the first sentence of the “parsing” wikipedia article using the stanford parser GUI. Did it parse correctly? Explain.

Yes, the first statement of the “parsing” page from wikipedia parsed successfully. Please find the attached screenshot of the gui for the same.



Question 3 :- Modify the given testnode function such that other facts about cats can be extracted, use the aganetact2 function below to test it.

```
In [14]: saveas(pretty(text), title+'-1.txt')

In [15]: from lxml import etree
         parser = etree.XMLParser(recover=True)
         tree = etree.parse('cat.xml',parser)
```

```
In [20]: s=root[6][0][0][0]
         s.attrib['value']

Out[20]: 'S'

In [21]: s[:]

Out[21]: [<Element node at 0x7febd8262ea8>,
          <Element node at 0x7febd8262ab8>,
          <Element node at 0x7febd8262dd0>]

In [22]: map(lambda x: x.attrib['value'], s[:])

Out[22]: ['NP', 'VP', '.']
```

```
In [54]:
def printnode(node):
    for i in node.findall(".//leaf"):
        print(" " + i.attrib['value']),
    print('')

def testnode2(node, agent, action):
    aa = node.findall("./node[@value='NP']//node[@value='NNS']//leaf[@value='"+agent+"']")
    bb = node.findall("./node[@value='VP']//leaf[@value='"+action+"']")
    if (len(aa) > 0 and len(bb) > 0):
        printnode(node)

def agentact2(node, agent, action):
    testnode2(node, agent, action)
    snodes = node.findall(".//node[@value='S']")
    for snode in snodes:
        testnode2(snode, agent, action)
```

```
In [162]: title = 'cats'
          list(map(lambda nn: agentact2(nn[0][0][0], title, 'are'), root))
          []

nDomestic cats are found in shorthair and longhair breeds
The big cats are well known : lions , tigers , leopards , jaguars , pumas , and cheetahs .
The big cats and wild cats are not tame , and can be very dangerous .
These kinds of cats are called \ `` feral cats \ '' .
While dogs have great stamina and will chase prey over long distances , cats are extremely fast , but
only over short distances .
People who receive cats as gifts are recommended to get it examined for its health .
All of these cats are called polydactyl cats .

Out[162]: []

In [161]: title = 'cat'
          list(map(lambda nn: agentact2(nn[0][0][0], title, 'are'), root))
          []

Unlike human arms , cat forelimbs are attached to the shoulder by free-floating clavicle bones .
a male cat 's penis has a band of about 120-150 backwards-pointing spines , which are about one millim
eter long ; upon withdrawal of the penis , the spines rake the walls of the female 's vagina , which
is a triggerTrigger : in the sense of an event which starts other events

Out[161]: []

In [160]: list(map(lambda nn: agentact2(nn[0][0][0], title, 'can'), root))
          []

The cat creeps towards a chosen victim , keeping its body flat and near to the ground so that it can
not be seen easily , until it is close enough for a rapid dash or pounce .
The cat 's tongue can act as a hairbrush and can clean and untangle a cat 's fur .
a cat gets fleas because fleas can make cats uncomfortable

Out[160]: []
```

As shown above, we have modified the testnode2 function finds other relevant facts about the cat using the verbs “are” and “can”. We have later tested this code using the agentact2 code.

Question 4 :- Q4. Extract facts about this people's wikipedia pages

-- Jim Parsons

-- Barack Obama

JIM PARSONS :-

```
In [37]: import requests

         title='Jim_Parsons'

         response = requests.get("http://simple.wikipedia.org/w/api.php?format=json&action=query&titles="+str(title)+"&prop=revisio

         response

Out[37]: <Response [200]>

In [38]: jsondata = response.json()

In [39]: saveas(pretty(jsondata), title+'.json')

In [40]: type(jsondata)

         jsondata.keys()

Out[40]: [u'batchcomplete', u'query']

In [355]: jsondata['query']

Out[355]: {'normalized': [{'from': 'Jim_Parsons', 'to': 'Jim Parsons'}],
           'pages': {'291858': {'ns': 0,
             'pageid': 291858,
             'revisions': [{'*': '{{Infobox person\n| name        = Jim Parsons\n| image       = Jim Parsons.png\n| imagesize   = 2
          30px\n| caption     = Jim Parsons, 2008\n| birthname   = James Joseph Parsons\n| birth_date   = {{birth date and age|1973
          |3|24}}\n| birth_place  = [[Houston]], [[Texas]], [[United States]]\n| occupation  = [[Actor]]\n| yearsactive = 1994-pres
          ent\n}}\n\'\'\'James Joseph "Jim" Parsons\'\'\' (born March 24, 1973 in [[Houston]], [[Texas]]) is an [[Americans|America
          n]] [[actor]]. Parsons is best known for playing the role of [[Sheldon Cooper]] in the [[television series]], \'\'[[The B
          ig Bang Theory]]\'\'. In 2010, he won an [[Emmy Award]] for \'\'Outstanding Lead Actor In a Comedy Series\'\' and in 2011
          , he won a [[Golden Globe Award]] for \'\'Best Actor - Television Series - Musical or Comedy\'\'. Both of these awards we
          re given to him for his role in The Big Bang Theory.\n\nAccording to [[The New York Times]], Parsons is [[gay]].<ref>[htt
          p://www.nytimes.com/2012/05/27/theater/jim-parsons-prepares-for-his-lead-role-in-harvey.html?_r=0&pagewanted=2 Stalked by
```

```
In [42]: content = jsondata['query']['pages'].values()[0]['revisions'][-1]

In [43]: content = content.get('*')

In [44]: import mwparserfromhell as mwph
         wikicode = mwph.parse(content)

In [45]: wikicode.filter_comments()

         wikicode.filter_headings()

         wikicode.filter_wikilinks()

Out[45]: [u'[[Houston]]',
          u'[[Texas]]',
          u'[[United States]]',
          u'[[Actor]]',
          u'[[Houston]]',
          u'[[Texas]]',
          u'[[Americans|American]]',
          u'[[actor]]',
          u'[[Sheldon Cooper]]',
          u'[[television series]]',
          u'[[The Big Bang Theory]]',

In [84]: text = wikicode.strip_code()
         text

Out[84]: u'James Joseph "Jim" Parsons (born March 24, 1973 in Houston, Texas) is an American actor. Parsons is best known for play
         ing the role of Sheldon Cooper in the television series, The Big Bang Theory. In 2010, he won an Emmy Award for Outstandi
         ng Lead Actor In a Comedy Series and in 2011, he won a Golden Globe Award for Best Actor - Television Series - Musical or
          Comedy. Both of these awards were given to him for his role in The Big Bang Theory.\n\nAccording to The New York Times,
         Parsons is gay.Stalked by shadows\n\nFilmography\nMovies\nNowhere to Go But Up (2003)\nGarden State (2004)\nThe King\'s I
         nn (2005)\nThe Great New Wonderful (2005)\nHeights (2005)\n10 Items or Less (2006)\nSchool for Scoundrels (2006)\nOn the
         Road with Judas (2007)\nGardener of Eden (2007)\nThe Big Year (2011)\nWish I Was Here (2014)\n"Home" (2015)\n\nTelevision
         \nEd as Chet (2002)\nWhy Blitt? as Mike (2003)\nTaste as Kris (2004)\nJudging Amy as Rob Holbrook (2004-05)\nThe Big Bang
          Theory as Sheldon Cooper (2007-present)\nFamily Guy as Sheldon Cooper (2009)\nGlenn Martin, DDS as Draven (2010)\nThe Su
         per Hero Squad Show as Nightmare (2011)\n\n References \n\nOther websites\n\nCategory:1973 births\nCategory:Living people
         \nCategory:Actors from Houston, Texas\nCategory:LGBT people from Texas\nCategory:American movie actors\nCategory:American
          stage actors\nCategory:American television actors\nCategory:Emmy Award winning actors\nCategory:Gay men\nCategory:Golden
          Globe Award winning actors\nCategory:LGBT actors'

In [90]: saveas(pretty(text), 'Jim_Prasons.txt')

In [96]: from lxml import etree
         parser = etree.XMLParser(recover=True)
         tree = etree.parse('Jim_Parsons.xml',parser)
         root = tree.getroot()

In [97]: s=root[0][0][0][0]
         map(lambda (x): x.attrib['value'], s[:])

Out[97]: ['``', 'NP', 'VP', '.']
```

Below be have created a testnode3 function which is specialized to find the facts about a person. In this, we can pass multiple nouns and verbs which are first split and then searched across the XML. We also manually append pronouns like “he”,”she” because the sentences don’t necessarily address a person by his/her name.

```
In [120]: def testnode3(node, agent, action):
              names = agent.split(" ")
              names.append("he")
              names.append("He")
              names.append("she")
              names.append("She")

              verbes = action.split(" ")
              bb = []
              aa = []
              for i in names:
                  aa += node.findall("./node[@value='NP']//leaf[@value='"+i+"']")
              for i in verbes:
                  bb += node.findall("./node[@value='VP']//leaf[@value='"+i+"']")

              if (len(aa) > 0 and len(bb) > 0):
                  printnode(node)

          def agentact3(node, agent, action):
              testnode3(node, agent, action)
              snodes = node.findall(".//node[@value='S']")
              for snode in snodes:
                  testnode2(snode, agent, action)
```

Below are the facts outputted by the language processing that we have used for Jim Parsons.

```
In [126]: map(lambda (nn): agentact3(nn[0][0][0], 'Jim Parsons', 'is winning'), root)
          []

           `` James Joseph \ `` Jim \ '' Parsons -LRB- born March 24 , 1973 in Houston , Texas -RRB- is an Am
          erican actor .
           Parsons is best known for playing the role of Sheldon Cooper in the television series , The Big Bang
          Theory .

Out[126]: []
```

BARACK OBAMA :-

Similarly for Barack Obama after creating his XML tree using the lexparser.sh, we pass it to the already created testnode3 function which specializes in finding facts about a person.

```
In [12]: from lxml import etree
         parser = etree.XMLParser(recover=True)
         tree = etree.parse('Barack.xml',parser)
         root = tree.getroot()

In [13]: s=root[0][0][0][0]
         map(lambda (x): x.attrib['value'], s[:])
Out[13]: ['``', 'NP', 'VP', '.']
```

```
In [18]: def printnode(node):
             for i in node.findall(".//leaf"):
                 print(" " + i.attrib['value']),
             print('')

         def testnode3(node, agent, action):
             names = agent.split(" ")
             names.append("he")
             names.append("He")
             names.append("she")
             names.append("She")

             verbes = action.split(" ")
             bb = []
             aa = []
             for i in names:
                 aa += node.findall("./node[@value='NP']//leaf[@value='"+i+"']")
             for i in verbes:
                 bb += node.findall("./node[@value='VP']//leaf[@value='"+i+"']")

             if (len(aa) > 0 and len(bb) > 0):
                 printnode(node)

         def agentact3(node, agent, action):
             testnode3(node, agent, action)
             snodes = node.findall(".//node[@value='S']")
             for snode in snodes:
                 testnode3(snode, agent, action)
```

Below are the facts outputted by the language processing that we have used for Barack Obama.

```
In [22]: map(lambda (nn): agentact3(nn[0][0][0], 'Barack Obama', 'president USA war War'), root)
         []
          He became the first president to openly express support for gay marriage , proposed gun control as a
         result of the Sandy Hook school shooting and has called for improving relations with Cuba .
          He was also against the war in Iraq .
          \ n \ nIn foreign policy , Obama made a plan to slowly withdraw troops from Iraq , ending the War
         in Iraq by the end of 2011 , while adding more troops to Afghanistan to help the United States win
         the War in Afghanistan .
          He also decided that the USA should help in the war against Libya .
          Obama has started a no-fly-zone policy on Libya which ended the civil war in October 2011 with the
         killing of Muammar al-Gaddafi .
          Obama has resulted in using his Executive Order -LRB- his power as president -RRB- to help reform
         things like the immigration system
Out[22]: []
```

Question 5 :-

Can you write code to automatically extract the following type facts about a given person's wikipedia page? Test your code using Barack Obama's wikipedia page

-- Place of birth
-- Spouse
-- Schools attended

We have written two generic functions testnode2 and testnode3 which finds the appropriate information using the grammar provided. They take in the parameter name and parameters like "born" and "in" for Place of birth and "University", "College" to find the appropriate sentences for any input.

```
def testnode2(node, agent,action, prep):
    names = agent.split(" ")
    aa = []
    for i in names:
        aa += node.findall("./node[@value='NP']//leaf[@value='"+i+"']")
    bb = node.findall("./node[@value='VP']//leaf[@value='"+action+"']")
    cc = node.findall("./node[@value='VP']//leaf[@value='"+prep+"']")
    if (len(aa) > 0 and len(bb) > 0 and len(cc) > 0):
        printnode(node)

def agentact2(node, agent,action,prep):
    testnode2(node, agent, action,prep)
    snodes = node.findall(".//node[@value='S']")
    for snode in snodes:
        testnode2(snode, agent,action, prep)
```

-- Place of birth :-

```
In [92]: map(lambda (nn): agentact2(nn[0][0][0], title, 'born','in'), root)
         []

          Early life Obama was born on August 4 , 1961 in Kapi \ u02bbolani Medical Center for Women and Childr
         en -LRB- called Kapi \ u02bbolani Maternity & Gynecological Hospital in 1961 -RRB- in Honolulu , Hawaii
         and is the first President to have been born in Hawaii .

Out[92]: []
```

-- Schools attended :-

```
In [117]: map(lambda (nn): agentact3(nn[0][0][0],title, 'University', 'College'), root)
          []

           Education He started college at Occidental College in Los Angeles , and graduated from Columbia
          University in New York City .
           After taking time off to community organize , Obama went to law school at Harvard University .
           He taught law part-time at the University of Chicago Law School .
           She has a Bachelor of Arts degree from Princeton University , and also a law degree from Harvard Law
          School .

Out[117]: []
```

For Spouse we have written another set of functions i.e testnode3 and agentact3. These are generic functions and when we provide them with POS like “wife” and “married” they return the correct output.

-- Spouse :-

```
def testnode3(node, agent,action, prep):
    names = agent.split(" ")
    names.append("He")
    names.append("She")
    aa = []
    for i in names:
        aa += node.findall("./node[@value='NP']//leaf[@value='"+i+"']")
    bb = node.findall("./node[@value='VP']//leaf[@value='"+action+"']")
    cc = node.findall("./node[@value='VP']//leaf[@value='"+prep+"']")
    if (len(aa) > 0 and (len(bb) > 0 or len(cc) > 0)):
        printnode(node)

def agentact3(node, agent,action,prep):
    testnode3(node, agent, action,prep)
    snodes = node.findall(".//node[@value='S']")
    for snode in snodes:
        testnode3(snode, agent,action, prep)
```

```
In [118]: map(lambda (nn): agentact3(nn[0][0][0],title, 'wife', 'married'), root)
          []

           Obama has been married to Michelle Obama since 1992 .
```