lab6_mohanty_abhisek

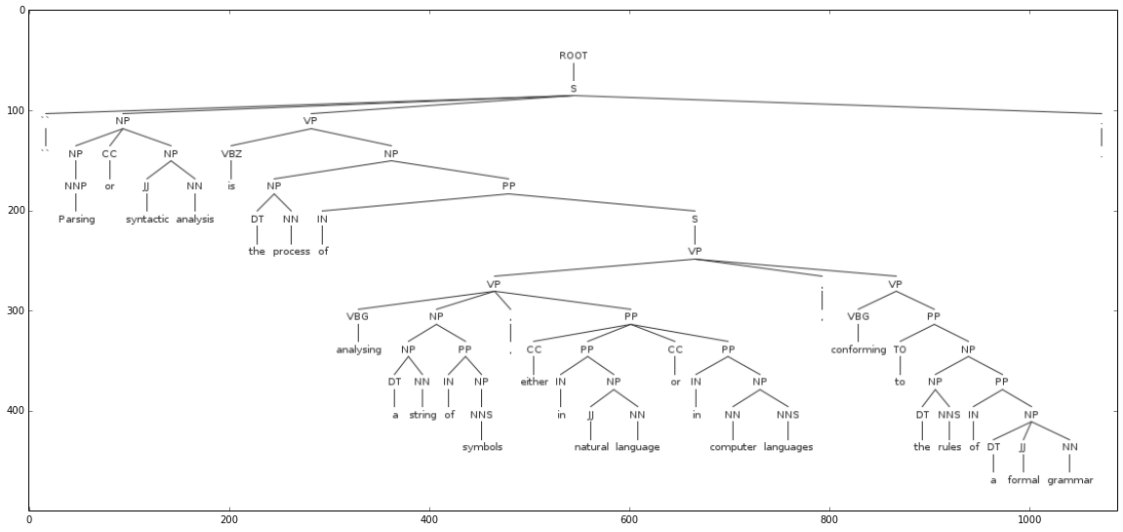
October 20, 2016

0.1 Question 1

```
In [19]: import requests
         title='cats'
         response = requests.get("http://en.wikipedia.org/w/api.php?format=json&action=query&titles="+s
         jsondata = response.json()
         content = jsondata['query']['pages'].values()[0]['revisions'][-1].get('*')
         # content
```

0.2 Question 2

```
In [104]: import matplotlib.pyplot as plt
          import matplotlib.image as mpimg
          %matplotlib inline
          image = mpimg.imread("parse.png")
          plt.figure(figsize=(20,10))
          plt.imshow(image)
          plt.show()
```



This looks correct overall, except for the right most tree. The rightmost S has a tree that only contains a VP which in turn is made up of 2 different VPs. This does not seem to be an appropriate statement structure.

0.3 Question 3

Modified this method to make sure it runs correctly for different verbs. Tried this on a list of 6 verbs. The output shows the verb and its corresponding statements.

```
In [80]: from lxml import etree
         parser = etree.XMLParser(recover=True)
         tree = etree.parse('/home/datascience/labs/lab6/cat.xml',parser)
         root=tree.getroot()

In [83]: 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)


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

         #are, have, were, can, may, use
         list_of_verbs = ['are', 'have', 'were', 'can', 'may', 'use']
         # were doesnt give anything because cats are not extinct !!

         for verb in list_of_verbs:
             print "\n" + verb + "\n"
             map(lambda (nn): agentact2(nn[0][0][0], title, verb), root)
         []
```

```
are

 Domestic cats are found in shorthair and longhair breeds .
 The big cats are well known : lions , tigers , leopards , jaguars , pumas , and chee
 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 extrem
 People who receive cats as gifts are recommended to get it examined for its health .
 All of these cats are called polydactyl cats .

have

 the cats would have something to eat in the afterlife
 Compared to other felines , domestic cats have narrowly spaced canine teeth
 Most cats have five claws on their front paws , and four on their rear paws .
 House cats have also been known to teach themselves to use lever-type doorknobs and toil
 Most cats have only four to five toes per paw , depending on whether it is the front
```

```
 These  mutated  cats  have  six  ,  seven  ,  and  in  rare  cases  even  more  .

were


can

 The  big  cats  and  wild  cats  are  not  tame  ,  and  can  be  very  dangerous  .
 cats  to  pass  their  body  through  any  space  into  which  they  can  fit  their  heads
 House  cats  can  become  overweight  through  lack  of  exercise  and  over-feeding  .

may

 many  cats  collect  extra  owners  ,  and  may  change  house  if  they  do  not  like  the  treatmen
 Some  cats  ,  depending  on  breed  ,  gender  ,  age  ,  and  general  health  ,  may  be  more  sus

use

 House  cats  have  also  been  known  to  teach  themselves  to  use  lever-type  doorknobs  and  toil
```

Out[83]: []

0.4 Question 4

```
In [2]: import json

        def pretty(jdata):
            str = json.dumps(jdata, sort_keys=True, indent=4).decode('string_escape')
            return str

        def saveas(sdata, fname):
            f = open(fname,'w')
            f.write(sdata)
            f.close()
```

Using en.wikipedia.org to get the full data. For parsing to SML, I am giving a clause of -maxLength = 200 in the parsetoxml file

parsetoxml.sh ->>> -outputFormat "xmlTree" -outputFormatOptions "xml" -maxLength 200

```
In [86]: import requests
         import mwparserfromhell as mwph

         title='Jim Parsons'
         response = requests.get("http://en.wikipedia.org/w/api.php?format=json&action=query&titles="+s
         jsondata = response.json()
         content = jsondata['query']['pages'].values()[0]['revisions'][-1].get('*')

         wikicode = mwph.parse(content)
         jim_text = wikicode.strip_code()
         saveas(pretty(jim_text), '/home/datascience/labs/lab6/'+title.replace (" ", "_")+'.txt')

In [87]: title='Barack Obama'
         response = requests.get("http://en.wikipedia.org/w/api.php?format=json&action=query&titles="+s
         jsondata = response.json()
         content = jsondata['query']['pages'].values()[0]['revisions'][-1].get('*')
```

```
        wikicode = mwph.parse(content)
        obama_text = wikicode.strip_code()
        saveas(pretty(obama_text), '/home/datascience/labs/lab6/'+title.replace (" ", "_")+'.txt')
```

In [93]:
```
from lxml import etree
parser = etree.XMLParser(recover=True)
```

* Keyword based Facts Extraction * Different keywords for Barack Obama and Jim Parsons * The first output is for Barack Obama, 2nd output is for Jim Parsons

In [100]:
```
## Keyword based information extraction

fact_keywords_obama = ['born', 'married', 'graduate', 'office', 'daughter', 'degree', 'father
fact_keywords_parsons = ['born','graduate','University','Sheldon', 'worked','award', 'film',

def makeStatement(node):
    line = ""
    for i in node.findall(".//leaf"):
        line = line + i.attrib['value'] + " "

    return line.strip()

def testnode(node, agent, action):
    if agent == 'Obama':
        keywords_list = fact_keywords_obama
    else:
        keywords_list = fact_keywords_parsons

    sentence = makeStatement(node)
    for word in keywords_list:
        if word in sentence and not ':' in sentence:
            print "\n" + sentence

title = 'Obama'
# title = 'Parsons'

if title == 'Obama':
    tree = etree.parse('/home/datascience/labs/lab6/Barack_Obama.xml',parser)
    root=tree.getroot()
else:
    tree = etree.parse('/home/datascience/labs/lab6/Jim_Parsons.xml',parser)
    root=tree.getroot()

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

print 'Facts about : ' + title + '\n\n'
map(lambda (nn): agentact(nn[0][0][0], title, 'check'), root)
[]
```

```
Facts about : Obama
```

‘‘ Barack Hussein Obama II -LRB- ; born August 4 , 1961 -RRB- is the 44th and current President of the U

‘‘ Barack Hussein Obama II -LRB- ; born August 4 , 1961 -RRB- is the 44th and current President of the U

‘‘ Barack Hussein Obama II -LRB- ; born August 4 , 1961 -RRB- is the 44th and current President of the U

to hold the office

Obama won the 2008 United States presidential election , on November 4 , 2008 .

As president , he slowly ended the wars in Afghanistan and Iraq , with intention to prepare the countri

He also signed the Affordable Care Act -LRB- often called ‘‘ Obamacare ’’ -RRB- which changed many healt

He became the first president to openly express support for gay marriage , proposed gun control as a res

Early life Obama was born on August 4 , 1961 in Kapi \ u02bbolani Medical Center for Women and Children

Early life Obama was born on August 4 , 1961 in Kapi \ u02bbolani Medical Center for Women and Children

to have been born in Hawaii

His father was a black exchange student from Kenya named Barack Obama Sr. .

Man of Destiny The Rosen Publishing Group 2010 page 7 His mother married Barack Obama Sr. in 1961 and di

His mother married Barack Obama Sr. in 1961 and divorced him in 1964

His stepfather was from Indonesia named Lolo Soetoro after his mother married him in 1965 .

His stepfather was from Indonesia named Lolo Soetoro after his mother married him in 1965 .

his mother married him in 1965

He spent most of his childhood in Hawaii and Chicago , Illinois , although he lived in Indonesia with hi

he lived in Indonesia with his mother and stepfather from age 6 to age 10

Education He started college at Occidental College in Los Angeles , and graduated from Columbia Universi

Before becoming president Obama worked for Alice Palmer , an Illinois state senator .

becoming president

Obama won the presidential election of 2008 .

Presidential campaigns 2008 presidential campaign Inaugural address by Barack Obama | left | thumb Barac

Presidential campaigns 2008 presidential campaign Inaugural address by Barack Obama | left | thumb Barac

2008 presidential campaign Inaugural address by Barack Obama | left | thumb Barack Obama ’s presidential

He raised the most amount of money ever for a presidential campaign .

Family thumb | left | Obama and the family presidential dog , Bo , running on the White House grounds .

Obama has been married to Michelle Obama since 1992 .

She has a Bachelor of Arts degree from Princeton University , and also a law degree from Harvard Law Sc

They have two daughters , Malia Ann , who was born in 1998Born on the 4th of July and Natasha -LRB- `` S

They have two daughters , Malia Ann , who was born in 1998Born on the 4th of July and Natasha -LRB- `` S

was born in 1998Born on the 4th of July and Natasha -LRB- `` Sasha '' -RRB- , born in 2001

Obama promised his daughters that the family would get a dog if he was elected President .

Obama promised his daughters that the family would get a dog if he was elected President .

the family would get a dog if he was elected President

he was elected President

In April 2009 , Senator Ted Kennedy , the brother of former President John F. Kennedy , gave Obama one

His father died from a car accident in Africa .

His maternal grandmother died just before Obama won the election to become President .

Obama won the election to become President

to become President

Presidency thumb | Obama talking in 2010 at the University of Minnesota Obama became President of the U

Obama signed the Patient Protection and Affordable Care Act which would bring health care reform to the

the Patient Protection and Affordable Care Act which would bring health care reform to the United State

would bring health care reform to the United States , which he said

Although his popularity was very high -LRB- around 70 % approval -RRB- when he entered office , his app

his popularity was very high -LRB- around 70 % approval -RRB- when he entered office

he entered office

On May 9 , 2012 , he became the first sitting US President to openly support legalizing same-sex marria

sitting US President to openly support legalizing same-sex marriage

US President to openly support legalizing same-sex marriage

Obama awarded several people , including former U.S. President Bill Clinton and media mogul Oprah Winfr

```
He has awarded the Presidential Medal of Freedom to many people , such as Stephen Hawking , Sandra Day (

He was officially nominated for his party 's Presidential choice on September 6 , 2012 .

Because of disagreements between Democrats and Republicans in Congress , neither side is getting anythi

Because of disagreements between Democrats and Republicans in Congress , neither side is getting anythi

Obama has resulted in using his Executive Order -LRB- his power as president -RRB- to help reform things

Obama has resulted in using his Executive Order -LRB- his power as president -RRB- to help reform things

using his Executive Order -LRB- his power as president -RRB- to help reform things like the immigration

using his Executive Order -LRB- his power as president -RRB- to help reform things like the immigration

to help reform things like the immigration system

Out[100]: []
```

Facts about Jim Parsons

```
In [101]: # title = 'Obama'
          title = 'Parsons'

          if title == 'Obama':
              tree = etree.parse('/home/datascience/labs/lab6/Barack_Obama.xml',parser)
              root=tree.getroot()
          else:
              tree = etree.parse('/home/datascience/labs/lab6/Jim_Parsons.xml',parser)
              root=tree.getroot()

          print 'Facts about : ' + title + '\n\n'
          map(lambda (nn): agentact(nn[0][0][0], title, 'check'), root)
          []

Facts about : Parsons


`` James Joseph `` Jim '' Parsons -LRB- born March 24 , 1973 -RRB- is an American actor .

He is known for playing Sheldon Cooper in the CBS sitcom The Big Bang Theory .

playing Sheldon Cooper in the CBS sitcom The Big Bang Theory

Sheldon Cooper in the CBS sitcom The Big Bang Theory

He has received several awards for his performance , including four Primetime Emmy Awards for Outstandi

He reprised the role in the film adaptation of the play , and received his seventh Emmy nomination , th

Early life Jim Parsons was born at St. Joseph Hospital in Houston , Texas , and was raised in one of its

After graduating from high school , Parsons received an undergraduate degree from the University of Hou
```

After graduating from high school , Parsons received an undergraduate degree from the University of Hous

Parsons enrolled in graduate school at the University of San Diego in 1999 .

Parsons enrolled in graduate school at the University of San Diego in 1999 .

He was one of seven students accepted into a special two-year course in classical theater , taught in pa

But we decided that he was so talented that we would give him a try and see how it worked out . ''

he was so talented that we would give him a try and see how it worked out

we would give him a try and see how it worked out

it worked out

Parsons graduated in 2001 and moved to New York .

in September 2013 and discovered French heritage from his father 's side .

Career Early career In New York , Parsons worked in Off-Broadway productions and made several televisio

After reading the pilot script , Parsons felt that the role of Sheldon Cooper would be a very good fit f

the role of Sheldon Cooper would be a very good fit for him

Parsons was cast as Sheldon Cooper , a physicist with social apathy who frequently belittles his friends

Parsons credits his University of San Diego training with giving him the tools to break down Sheldon 's

Parsons credits his University of San Diego training with giving him the tools to break down Sheldon 's

giving him the tools to break down Sheldon 's lines

to break down Sheldon 's lines

Television critic Andrew Dansby compares Parsons ' physical comedy to that of Buster Keaton and other si

In August 2009 , Parsons won the Television Critics Association award for individual achievement in come

In January 2011 , he won the Golden Globe award for Best Actor in a Television Series \ u2013 Comedy -LR

the award was presented by co-star Cuoco

Other works In 2011 , Parsons appeared with Jack Black , Owen Wilson , Steve Martin , and Rashida Jones

He voiced Oh , one of the lead roles in the DreamWorks Animation comedy film Home -LRB- 2015 -RRB- , al

Oh , one of the lead roles in the DreamWorks Animation comedy film Home -LRB- 2015 -RRB- , alongside Ril

His father died in a car crash on April 29 , 2001 .

His partner is art director Todd Spiewak .

Out[101]: []

0.5 Question 5

In [79]:
```
# Maintain a dictionary of place names and match the leaf values once we are done comparing ke
# like graduate for education, married/wife for spouse etc.

list_of_places = ['Honolulu', 'Hawaii', 'Washington', 'USA', 'New', 'York', 'City', 'Los', 'An
list_of_names = ['Ann','Barack','Malia','Sasha','Michelle','Obama']
dict_of_education = ['Occidental', 'University', 'College', 'Harvard', 'Stanford', 'Columbia']
special_chars = [',', '.']

def makeBagofWords(node):
    line = ""
    for i in node.findall(".//leaf"):
        line = line + i.attrib['value'] + " "

    return line.strip().split()

def testnode2(node, agent, action):
    if action == 'born':
        name = node.findall("./node[@value='NP']//node[@value='NNS']//leaf[@value='"+agent+"']
        aa = node.findall("./node[@value='VP']//node[@value='VP']//node[@value='VP']//node[@va
        place = node.findall("./node[@value='VP']//node[@value='VP']//node[@value='VP']//node[
        placeName = node.findall("./node[@value='VP']//node[@value='VP']//node[@value='VP']//n


        if len(name) > 0 and len(aa) > 0 and len(place) > 0 and len(placeName) > 0:
            for node in placeName:
                for leaf in node.findall(".//leaf"):
                    val = leaf.attrib['value']
                    if val in list_of_places or val in special_chars:
                        print leaf.attrib['value'] + ' ',

    elif action == 'spouse':
        name = node.findall("./node[@value='NP']//node[@value='NNP']//leaf[@value='"+agent+"']
        aa = node.findall("./node[@value='VP']//node[@value='VP']//node[@value='VP']//node[@va
        name = node.findall("./node[@value='VP']//node[@value='VP']//node[@value='VP']//node[@

        if len(name) > 0 and len(aa) > 0 and len(name) > 0:
            for node in name:
                for leaf in node.findall(".//leaf"):
                    val = leaf.attrib['value']
                    if val in list_of_names or val in special_chars:
                        print leaf.attrib['value'] + ' ',

    elif action == 'school':
        words = makeBagofWords(node)
#         print words
        if 'College' in words or 'University' in words:
            if (('He' in words and not 'She' in words) or ('Obama' in words)):
                if (not 'taught' in words and not 'talking' in words):
                    if 'started' in words or 'graduated' in words or 'went' in words:
#                         printnode(node)
```

```
                                nouns = node.findall(".//node[@value='NNP']")

                                for node in nouns:
                                    for leaf in node.findall(".//leaf"):
                                        val = leaf.attrib['value']
                                        if val in dict_of_education or val in special_chars or val in
                                            print leaf.attrib['value'] + ' ',
                    else:
                        return


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

        print title + " was born in : "
        map(lambda (nn): agentact2(nn[0][0][0], title, 'born'), root)

        print "\n\n" + title + " is married to : "
        map(lambda (nn): agentact2(nn[0][0][0], title, 'spouse'), root)

        print "\n\n" + title + " - Schools Attended : "
        map(lambda (nn): agentact2(nn[0][0][0], title, 'school'), root)
        []
```

```
Obama was born in :
Honolulu  ,  Hawaii

Obama is married to :
Michelle  Obama

Obama - Schools Attended :
Occidental  College  Los  Angeles  Columbia  University  New  York  City  Harvard  University
```

Out[79]: []

In []: