CO Open in Colab

This notebook explores relation extraction by measuring common dependency paths between two entities that hold a given relation to each other -- here, the relation "born_in" between a PER entity and an GPE entity, using data from Wikipedia biographies. Feel free to run this notebook either on your on computer or on Google Colab. If using Google Colab, install these libraries first:

In [ ]:
```
#!pip install spacy==2.1.0
```

In [ ]:
```
#!python -m spacy download en_core_web_sm
```

In [ ]:
```
#!pip install neuralcoref --no-binary neuralcoref
```

Otherwise, use your local `anlp21-spacy2` environment we set up in class on Tuesday.

In [ ]:
```
import re
import spacy
import neuralcoref
from collections import Counter
```

In [ ]:
```
nlp = spacy.load('en_core_web_sm')
coref = neuralcoref.NeuralCoref(nlp.vocab)
nlp.add_pipe(coref, name='neuralcoref')
```

In [ ]:
```python
def get_path(one, two):

    """ Get dependency path between two tokens in a sentence; return None if

    one_heads=[x for x in one.ancestors]
    two_heads=[x for x in two.ancestors]

    up_path=[]
    down_path=[]
    up_path.append(one)
    down_path.append(two)

    lca=None
    for head in one_heads:
        if head in two_heads:
            lca=head
            break

        up_path.append(head)

    for head in two_heads:
        if head == lca:
            break

        down_path.append(head)

    if lca is None:
        return None

    path="%s->%s<-%s" % ('->'.join(["%s" % x.dep_ for x in up_path]), lca.tex
    return path
```

In [ ]:

```python
def get_closest_coref(entity1, clusters, target_entity):

    """ Given entities e1 and mention m2 of another entity, returns the menti

    targetCluster=None
    for chain in clusters:
        for mention in chain.mentions:
            if mention.start <= entity1.start and mention.end >= entity1.end:
                targetCluster=chain
                break

    if targetCluster is None:
        return None

    closestMention=None
    dist=100
    for mention in targetCluster:
            sentDist=abs(target_entity.sent.start-mention.sent.start)
            if sentDist < dist:
                dist=sentDist
                closestMention=mention
            if sentDist == dist and closestMention is not None:
                if abs(target_entity.start-mention.start) < abs(target_entity
                    closetMention=mention
    return closestMention
```

Q0. In class activity: here's a Google spreadsheet with the first 5 sentences from ~500 Wikipedia biographies. If you're present in class, pick 10 rows of this spreadsheet that other students haven't already claimed and put your student ID in the "Student ID" column; then go through those 10 rows and read the document. If you can infer that a person from the "Candidate people" column was born in a place in the "Candidate places" column, list that person in the "PER BORN" column and the place in the "PLACE BORN" column.

```
In [ ]:   def read_training(filename):

              """ Read in training data for <person, place> tuples that express the "bo

              -- Use coreference resolution to identity the person mention closest to t
              -- Use dependency parsing to extract the syntactic path from that person

              """

              data=[]
              with open(filename) as file:
                  for line in file:
                      cols=line.split("\t")
                      idd=cols[0]
                      doc=cols[1]
                      pers=cols[4]
                      place=cols[5].rstrip()

                      if pers != "" and place != "":
                          doc=nlp(doc)

                          target_person=None
                          target_place=None

                          # Annotations are at the type level, so let's anchor them to
                          for entity in doc.ents:
                              if entity.text == pers:
                                  target_person=entity
                              elif entity.text == place:
                                  target_place=entity

                          if target_person is not None and target_place is not None:

                              # Use coreference to get person mention that's closest to
                              closest_person_mention=get_closest_coref(target_person, d
                              if closest_person_mention is None:
                                  closest_person_mention=target_person

                              path=get_path(closest_person_mention.root, target_place.r

                              # if a path can be found between the two
                              if path is not None:
                                  data.append((pers, place, path, target_place.sent ))
              return data
```

At the end of class, save this Google sheet as a tsv in `born.tsv` and execute the
`read_training` function on it to read in the <person, place> tuples. If you're executing
this on Google Colab, download the born.tsv file here (after class). In both cases, adjust the
path to `born.tsv` for your environment.

In [ ]:
```
#!brew install wget
!wget https://raw.githubusercontent.com/dbamman/anlp21/master/data/born.tsv
```

```
--2021-11-21 14:37:33--  https://raw.githubusercontent.com/dbamman/anlp21/mast
er/data/born.tsv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111
.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.11
1.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 443267 (433K) [text/plain]
Saving to: 'born.tsv.2'

born.tsv.2          100%[===================>] 432.88K  --.-KB/s    in 0.04s

2021-11-21 14:37:33 (11.0 MB/s) - 'born.tsv.2' saved [443267/443267]
```

In [ ]:
```
trainingData=read_training("born.tsv")
for data in trainingData:
    print ('\t'.join([str(x) for x in data]))
```

```
Hutchinson      Alford  nsubjpass->born<-prep<-pobj     Hutchinson was born in
Alford , Lincolnshire , England , the daughter of Francis Marbury , an Anglica
n cleric and school teacher who gave her a far better education than most othe
r girls received .
William Henry Seward    Florida nsubjpass->born<-prep<-pobj      Seward was bor
n in Florida , Orange County , New York , where his father was a farmer and ow
ned slaves .
Konrad Henlein  Maffersdorf      nsubjpass->born<-prep<-pobj      Konrad Henlein
was born in Maffersdorf ( present-day Vratislavice nad Nisou ) near Reichenber
g ( Liberec ) , in what was then the Bohemian crown land of Austria-Hungary .
Robert Sylvester Kelly  Chicago nsubj->began<-advcl<-prep<-pobj A native of Ch
icago , Kelly began performing during the late 1980s and debuted in 1992 with
the group Public Announcement .
Alastair Nathan Cook    England nsubjpass->regarded<-prep<-pobj<-prep<-pobj<-r
elcl<-prep<-pobj        He is regarded as one of the greatest batsmen ever to
play for England , and is one of the most prolific batsmen of the modern era .
Michael Kaine   St. Paul        nsubj->grew<-advcl<-prep<-pobj  Born in St. Pa
ul , Minnesota , Kaine grew up in Overland Park , Kansas , graduated from the
University of Missouri , and earned a J.D .
Adam Mitchel Lambert    U.S.    nsubj->released<-prep<-pobj<-relcl<-prep<-pobj
Later that year , he released his debut album , `` For Your Entertainment '' ,
which debuted at number three on the U.S. `` Billboard '' 200 .
Sonia Maria Sotomayor   Bronx   nsubjpass->born<-prep<-pobj     Sotomayor was
born in The Bronx , New York City , to Puerto Rican-born parents .
Peter Gary Tatchell     Australia       nsubj->is<-attr<-prep<-pobj     Peter
Gary Tatchell ( born 25 January 1952 ) is a British human rights campaigner ,
originally from Australia , best known for his work with LGBT social movements
.
Maurice Yaméogo the Republic of Upper Volta     nsubj->was<-attr<-prep<-pobj
```

Maurice Yaméogo (31 December 1921 — 15 September 1993) was the first Preside
nt of the Republic of Upper Volta , now called Burkina Faso , from 1959 until
1966 .
Augustus Heinrich Petermann   Bleicherode   nsubjpass->born<-prep<-pobj
Petermann was born in Bleicherode , Germany .
Alfred Joseph Hitchcock London nsubj->entered<-advcl<-prep<-pobj<-prep<-pobj
Born on the outskirts of London , Hitchcock entered the film industry in 1919
as a title card designer after training as a technical clerk and copy writer f
or a telegraph-cable company .
Keith Ross Miller   Australia   nsubjpass->regarded<-prep<-pobj<-poss
Miller is widely regarded as Australia 's greatest ever all-rounder .
Condoleezza Rice   United States   nsubj->served<-prep<-pobj<-nmod She se
rved as the 66th United States Secretary of State , the second person to hold
that office in the administration of President George W. Bush .
Christopher James Christie   Newark nsubjpass->born<-prep<-pobj
Christie was born in Newark and raised in Livingston .
Benjamin Franklin   the United States   nsubj->was<-attr<-conj<-prep<-
pobj<-prep<-pobj   Benjamin Franklin (April 17 , 1790) was an American
polymath and one of the Founding Fathers of the United States .
Petermann   Bleicherode   nsubjpass->born<-prep<-pobj   Petermann was
born in Bleicherode , Germany .
Francisco Franco Bahamonde   Spain   nsubj->was<-attr<-relcl<-prep<-pobj
Francisco Franco Bahamonde (; ; 4 December 1892 — 20 November 1975) was a Sp
anish general who ruled over Spain as a military dictator from 1939 , after th
e nationalist victory in the Spanish Civil War , until his death in 1975 .
Jawaharlal Nehru   India   nsubj->emerged<-conj<-dobj   He emerged as
an eminent leader of the Indian independence movement under the tutelage of Ma
hatma Gandhi and served India as Prime Minister from its establishment as an i
ndependent nation in 1947 until his death in 1964 .
Ayaan Hirsi Ali Somali nsubj->csubj->is<-attr<-amod<-npadvmod Ayaan Hirsi Al
i (; ; born Ayaan Hirsi Magan , 13 November 1969) is a Somali-born Dutch-Ame
rican activist , feminist , author , scholar and former politician .
Timothy Francis Leary   the United States   nsubj->advcl->were<-prep<-pobj
As a clinical psychologist at Harvard University , Leary conducted experiments
under the Harvard Psilocybin Project in 1960-62 (LSD and psilocybin were stil
l legal in the United States at the time) , resulting in the Concord Prison E
xperiment and the Marsh Chapel Experiment .
Tycho Brahe   Scania nsubjpass->born<-prep<-pobj<-prep<-pobj He was born in
the then Danish peninsula of Scania .
Anne Hutchinson Puritan poss->nsubj->were<-prep<-pobj<-prep<-pobj<-compound
Her strong religious convictions were at odds with the established Puritan cle
rgy in the Boston area , and her popularity and charisma helped create a theol
ogical schism that threatened to destroy the Puritans ' religious community in
New England .
Warren Los Angeles   nsubjpass->born<-prep<-pobj   Warren was born in 189
1 in Los Angeles and was raised in Bakersfield , California .
Erich von Manstein   Germany nsubj->rose<-conj<-prep<-pobj<-acl<-ccomp<-nsu
bj   He rose to the rank of captain by the end of the war and was active in
the inter-war period helping Germany rebuild her armed forces .
Joseph Maurice Ravel   France nsubj->attended<-dobj<-poss   Born to a musi
c-loving family , Ravel attended France 's premier music college , the Paris C
onservatoire ; he was not well regarded by its conservative establishment , wh
ose biased treatment of him caused a scandal .

Jimmie Kenneth Johnson El Cajon   nsubjpass->born<-prep<-pobj
Johnson was born in El Cajon , California , and began racing motorcycles at th
e age of four .
John Winston Lennon   Liverpool   pobj->as<-pobj<-prep<-pobj   He was
born as John Winston Lennon in Liverpool , where he became involved in the ski
ffle craze as a teenager .
Donald John Trump   New York City   nsubjpass->born<-conj<-prep<-pobj<-com
pound   Trump was born and raised in the New York City borough of Queens .
Katharine Cornell   Berlin nsubjpass->born<-prep<-pobj   She was born i
n Berlin to American parents and raised in Buffalo , New York .
King   Atlanta nsubjpass->known<-advcl<-prep<-pobj   Born in Atlanta , King
is best known for advancing civil rights through nonviolence and civil disobed
ience , tactics his Christian beliefs and the nonviolent activism of Mahatma G
andhi helped inspire .
Brigid Lansbury Regents Park   nsubjpass->born<-prep<-pobj<-appos<-prep<-pobj
Lansbury was born to Irish actress Moyna Macgill and English politician Edgar
Lansbury , an upper-middle-class family in Regents Park , central London ; her
paternal grandfather was the British Labour Party leader George Lansbury .
Marshall   Fauquier County nsubjpass->born<-prep<-pobj   Marshall was b
orn in Fauquier County , Virginia in 1755 .
Nancy Davis Reagan   New York City   nsubjpass->born<-prep<-pobj   She wa
s born in New York City .
Jomo Kenyatta   Kenya   nsubj->was<-conj<-prep<-pobj<-prep<-pobj   He was
the country 's first indigenous head of government and played a significant ro
le in the transformation of Kenya from a colony of the British Empire into an
independent republic .
Robyn Rihanna Fenty   Saint Michael   nsubjpass->discovered<-advcl<-prep<-po
bj   Born in Saint Michael , Barbados , and raised in Bridgetown , she was
discovered by American record producer Evan Rogers in her home country of Barb
ados in 2003 .
Theodore Roosevelt Jr. New York   nsubj->served<-conj<-pobj<-prep<-pobj
He also served as the 25th Vice President of the United States from March to S
eptember 1901 and as the 33rd Governor of New York from 1899 to 1900 .
Mikhail Sergeyevich Gorbachev   the Soviet Union   nsubj->was<-advcl<-con
j<-pobj<-prep<-pobj   He was the country 's head of state from 1988 until 19
91 (titled as Chairman of the Presidium of the Supreme Soviet from 1988 to 19
89 , as Chairman of the Supreme Soviet from 1989 to 1990 , and as President of
the Soviet Union from 1990 to 1991) .
Katharine Cornell   Berlin nsubjpass->born<-prep<-pobj   She was born i
n Berlin to American parents and raised in Buffalo , New York .
Abbott London nsubjpass->born<-prep<-pobj   Abbott was born in London to a
n Australian mother and a British father , and moved to Sydney at the age of t
wo .
William Wilberforce   Kingston   nsubj->began<-advcl<-prep<-pobj A nati
ve of Kingston upon Hull , Yorkshire , he began his political career in 1780 ,
eventually becoming a Member of Parliament for Yorkshire (1784—1812) .
Karim Mostafa Benzema   Lyon   nsubjpass->born<-prep<-pobj<-prep<-pobj Benzem
a was born in the city of Lyon and began his football career with local club B
ron Terraillon .
Michael Collins Woodfield   nsubjpass->born<-prep<-pobj   Collins was bo
rn in Woodfield , County Cork , the youngest of eight children , and his famil
y had republican connections reaching back to the 1798 rebellion .
Hassan Sheikh Mohamud   Somalia nsubj->was<-attr<-prep<-pobj   He was the 8th

President of Somalia from 16 September 2012 until 16 February 2017 .
Robert Sylvester Kelly   Chicago nsubj->began<-advcl<-prep<-pobj A native of Ch
icago , Kelly began performing during the late 1980s and debuted in 1992 with
the group Public Announcement .
Aamir Khan   India   nsubj->has<-dobj<-prep<-pobj   He has a significant f
ollowing in India and China , and has been described by `` Newsweek '' as `` t
he biggest movie star '' in the world .
William Phips   the Province of Massachusetts Bay   nsubj->was<-attr<-acl<
-prep<-pobj<-conj<-prep<-pobj   Sir William Phips (or Phipps ; February 2 , 1
651 — February 18 , 1695) was a shepherd boy born in the Massachusetts Bay Co
lony , a shipwright , ship 's captain , treasure hunter , a major general , an
d the first royally appointed governor of the Province of Massachusetts Bay .
Gillard Barry   nsubj->migrated<-advcl<-prep<-pobj   Born in Barry , Wales
, Gillard migrated with her family to Adelaide , South Australia , in 1966 .
Anthony John Abbott   London nsubjpass->born<-prep<-pobj   Abbott was bor
n in London to an Australian mother and a British father , and moved to Sydney
at the age of two .
Paul Joseph Goebbels   Germany nsubj->was<-attr<-conj<-prep<-pobj<-prep<-pobj
Paul Joseph Goebbels (; 29 October 1897 — 1 May 1945) was a German Nazi poli
tician and Reich Minister of Propaganda of Nazi Germany from 1933 to 1945 .
William George Heirens Illinois   nsubj->spent<-dobj<-prep<-pobj<-prep<-
pobj<-prep<-pobj<-appos He spent the later years of his sentence at the Dixon
Correctional Center in Dixon , Illinois .
William Morris Walthamstow   nsubj->came<-advcl<-prep<-pobj Born in Waltha
mstow , Essex , to a wealthy middle-class family , Morris came under the stron
g influence of medievalism while studying Classics at Oxford University , ther
e joining the Birmingham Set .
Rodney Stephen Steiger Westhampton   nsubjpass->born<-prep<-pobj
Steiger was born in Westhampton , New York , the son of a vaudevillian .
Frederick Sewards Trueman   Stainton   nsubjpass->born<-prep<-pobj
He was born in Stainton , West Riding of Yorkshire , and died in Steeton with
Eastburn , West Yorkshire .
Francisco Franco Bahamonde   Spain   nsubj->was<-attr<-relcl<-prep<-pobj
Francisco Franco Bahamonde (; ; 4 December 1892 — 20 November 1975) was a Sp
anish general who ruled over Spain as a military dictator from 1939 , after th
e nationalist victory in the Spanish Civil War , until his death in 1975 .
Marx   Trier   npadvmod->amod->dep->Born<-prep<-pobj   Born in Trier , German
y , to a Jewish middle-class family , Marx studied law and philosophy at unive
rsity .
Earl Warren   Los Angeles   nsubjpass->born<-prep<-pobj   Warren was bor
n in 1891 in Los Angeles and was raised in Bakersfield , California .
Luis Alberto Suárez Díaz   Uruguay nsubj->won<-conj<-prep<-pobj   Often
regarded as one of the best players in the world , Suárez has won 17 trophies
in his career , including five league titles and a UEFA Champions League title
at club level , and a Copa América with Uruguay .
Julia Eileen Gillard   Wales   nsubj->migrated<-advcl<-prep<-pobj<-conj
Born in Barry , Wales , Gillard migrated with her family to Adelaide , South A
ustralia , in 1966 .
Frederick Sewards Trueman   Stainton   nsubjpass->born<-prep<-pobj
He was born in Stainton , West Riding of Yorkshire , and died in Steeton with
Eastburn , West Yorkshire .
Jeffrey Koons   Pennsylvania   nsubj->lives<-conj<-prep<-pobj<-conj<-prep<-po
bj<-appos   He lives and works in both New York City and his hometown of Y

ork , Pennsylvania .
Seward   Florida nsubjpass->born<-prep<-pobj   Seward was born in Florida , O
range County , New York , where his father was a farmer and owned slaves .
Joseph Michael Arpaio   America nsubj->styled<-prep<-pobj<-poss Arpaio styled
himself as `` America 's Toughest Sheriff '' .
Edward Scott Pruitt   Oklahoma   nsubj->is<-attr<-prep<-pobj<-prep<-pob
j   Edward Scott Pruitt (born May 9 , 1968) is an American lawyer and Re
publican politician from the state of Oklahoma .
Louis Pierre Althusser Algeria nsubjpass->born<-prep<-pobj   He was born in
Algeria and studied at the École Normale Supérieure in Paris , where he eventu
ally became Professor of Philosophy .
Irving Berlin   Imperial Russia nsubj->arrived<-advcl<-prep<-pobj   Born i
n Imperial Russia , Berlin arrived in the United States at the age of five .
Nash   Australia   nsubj->was<-attr<-conj<-dobj<-relcl<-prep<-pobj In cri
cket , Nash was a fast bowler and hard hitting lower order batsman who played
two Test matches for Australia , taking 10 wickets at 12.80 runs per wicket ,
and scoring 30 runs at a batting average of 15 .
John Richard Kasich Jr. Pittsburgh   nsubj->lived<-nsubj<-prep<-pobj A nati
ve of Pittsburgh , Pennsylvania , Kasich has lived much of his adulthood in Oh
io , specifically the state capital of Columbus .
Frederick Russell Burnham DSO   Minnesota   nsubj->relcl->reservation<-pre
p<-pobj Burnham was born on a Dakota Sioux Indian reservation in Minnesota whe
re he learned the ways of American Indians as a boy .
Edward Gough Whitlam   Australia   nsubj->won<-prep<-pcomp<-agent<-pobj<-
prep<-pobj   He won the 1974 election before being controversially dismisse
d by the Governor-General of Australia , Sir John Kerr , at the climax of the
1975 Australian constitutional crisis .
Ian Douglas Smith   Rhodesia   dobj->ccomp->remains<-nsubj<-relcl<-pr
ep<-pcomp<-dobj Smith , who has been described as personifying white Rhodesia
, remains a highly controversial figure—supporters venerate him as a man of in
tegrity and vision `` who understood the uncomfortable truths of Africa '' , w
hile critics describe an unrepentant racist whose policies and actions caused
the deaths of thousands and contributed to Zimbabwe 's later crises .
Aafia Siddiqui Pakistan   nsubjpass->born<-prep<-pobj   Siddiqui was b
orn in Pakistan to a Deobandi Muslim family .
Johann Mickl   Austro nsubjpass->commissioned<-conj<-prep<-pobj<-amod<-compo
und   He was commissioned shortly before the outbreak of World War I , and s
erved with Austro-Hungarian forces on the Eastern and Italian Fronts as compan
y commander in the Imperial-Royal Mountain Troops .
Josephine Sarah New York   nsubjpass->born<-prep<-pobj   Josephine was
born in New York to a Prussian Jewish family ; her father was a baker .
Rodney Stephen Steiger New York   nsubjpass->born<-prep<-pobj<-appos
Steiger was born in Westhampton , New York , the son of a vaudevillian .
Burnham Minnesota   nsubjpass->born<-prep<-pobj<-prep<-pobj Burnham was bo
rn on a Dakota Sioux Indian reservation in Minnesota where he learned the ways
of American Indians as a boy .
Joseph Rudyard Kipling India   nsubjpass->born<-prep<-pobj   He was born in
India , which inspired much of his work .
Joseph Rudyard Kipling India   nsubjpass->born<-prep<-pobj   He was born in
India , which inspired much of his work .
Ahmet Davutoğlu Turkey nsubj->is<-attr<-relcl<-attr<-prep<-pobj   Ahmet
Davutoğlu (; born 26 February 1959) is a Turkish academic , politician and f
ormer diplomat who was the Prime Minister of Turkey and leader of the Justice

and Development Party (AKP) from August 2014 to May 2016 .
Roderick David Stewart London nsubj->is<-advcl<-conj<-prep<-pobj   Born a
nd raised in London , he is of Scottish and English ancestry .
Suharto Indonesia   nsubj->csubj->was<-attr<-relcl<-prep<-pobj<-prep<-pobj
Suharto (also written Soeharto , ; 8 June 1921 — 27 January 2008) was an Ind
onesian military leader and politician who served as the second President of I
ndonesia , holding the office for 31 years , from the ousting of Sukarno in 19
67 until his resignation in 1998 .
Sonia Maria Sotomayor   Bronx   nsubjpass->born<-prep<-pobj   Sotomayor was
born in The Bronx , New York City , to Puerto Rican-born parents .
Charles Spencer Chaplin London poss->childhood<-prep<-pobj   Chaplin 's chi
ldhood in London was one of poverty and hardship .
Roderick David Stewart London nsubj->is<-advcl<-conj<-prep<-pobj   Born a
nd raised in London , he is of Scottish and English ancestry .
Beyoncé Houston appos->pobj->in<-pobj   Born and raised in Houston , Texas , B
eyoncé performed in various singing and dancing competitions as a child .
Archibald Campbell MacLaren   England nsubj->played<-prep<-pobj   A righ
t-handed batsman , he played 35 Test matches for England , as captain in 22 of
those games , and led the team to defeat in four Ashes series against Australi
a .
Omar Ahmed Sayid Khadr Canada nsubjpass->taken<-advcl<-prep<-pobj   Born i
n Canada , Khadr was taken to Afghanistan by his father , who was affiliated w
ith Al-Qaeda and other terrorist organizations .
Donald John Trump   New York City   nsubjpass->born<-conj<-prep<-pobj<-com
pound   Trump was born and raised in the New York City borough of Queens .
George Stanley McGovern South Dakota   nsubj->grew<-prep<-pobj<-appos
McGovern grew up in Mitchell , South Dakota , where he was a renowned debater
.
Josephine Sarah New York   nsubjpass->born<-prep<-pobj   Josephine was
born in New York to a Prussian Jewish family ; her father was a baker .
Margaret Alice Murray   Calcutta   nsubj->divided<-advcl<-prep<-pobj<-pre
p<-pobj Born to a wealthy middle-class English family in Calcutta , British In
dia , Murray divided her youth between India , Britain , and Germany , trainin
g as both a nurse and a social worker .
William Edward Burghardt Du Bois   Great Barrington   nsubj->grew<-a
dvcl<-prep<-pobj   Born in Great Barrington , Massachusetts , Du Bois gre
w up in a relatively tolerant and integrated community .
Roberto Luongo Montreal   nsubj->is<-advcl<-prep<-pobj   Born in Montre
al , Quebec , he is of Italian and Irish ancestry .
John Sherman   U.S.   nsubj->was<-attr<-prep<-pobj<-compound John Sherman (
May 10 , 1823October 22 , 1900) was a politician from the U.S. state of Ohio
during the American Civil War and into the late nineteenth century .
Franz Liszt   Raiding nsubjpass->born<-prep<-pobj<-prep<-pobj Franz Liszt wa
s born on October 22 , 1811 , in the village of Raiding ( ) in the Kingdom of
Hungary , then part of the Habsburg Empire (and since 1920 also part of Austr
ia) , in Sopron County ( ) .
John Lawrence McKenzie Brazil nsubjpass->born<-prep<-pobj<-prep<-pobj John L
awrence McKenzie (October 9 , 1910 — March 2 , 1991) was born on October 9 ,
1910 , in Brazil , Indiana , the first of the six children of Myra (Daly) an
d Harry McKenzie .
Julius Kambarage Nyerere   Tanzania   nsubj->governed<-conj<-prep<-p
obj<-relcl<-dobj<-appos He governed Tanganyika as its Prime Minister from 1961
to 1963 and then as its President from 1963 to 1964 , after which he led its s

uccessor state , Tanzania , as its President from 1964 until 1985 .
Elizabeth Canning   the City of London   nsubj->disappeared<-prep<-pcom
p<-prep<-pobj<-prep<-pobj   She disappeared on 1 January 1753 , before ret
urning almost a month later to her mother 's home in Aldermanbury in the City
of London , emaciated and in a `` deplorable condition '' .
Charles Albert Sardinia   nsubj->was<-attr<-prep<-pobj   Charles Albert
(; 2 October 1798 — 28 July 1849) was the King of Sardinia from 27 April 183
1 to 23 March 1849 .
Michael Collins Woodfield   nsubjpass->born<-prep<-pobj   Collins was bo
rn in Woodfield , County Cork , the youngest of eight children , and his famil
y had republican connections reaching back to the 1798 rebellion .
Copernicus   the Kingdom of Poland   nsubjpass->born<-conj<-prep<-pobj<-app
os<-relcl<-attr<-prep<-pobj   Copernicus was born and died in Royal Prussia
, a region that had been part of the Kingdom of Poland since 1466 .

Q1: Count the syntactic paths identified in the training data. What are the two that are most frequently attested?

In [ ]:
```python
trainingData=read_training("born.tsv")
synt_dict = {}
for data in trainingData:
    synt = data[2]
    try:
        synt_dict[synt] += 1
    except KeyError:
        synt_dict[synt] = 1
```

In [ ]:
```python
sort_orders = sorted(synt_dict.items(), key=lambda x: x[1], reverse=True)
```

In [ ]:
```python
sort_orders[0:2]
# most frequent: nsubjpass->born<-prep<-pobj
# second most frequent: nsubjpass->born<-prep<-pobj<-prep<-pobj
```

Out[ ]:
```
[('nsubjpass->born<-prep<-pobj', 29),
 ('nsubjpass->born<-prep<-pobj<-prep<-pobj', 5)]
```

Q2: Write a function to read in a target file (containing one document per line) and a syntactic path and identify all people/places that are joined by that path. Hint: you can use the get_path functon defined above to retrieve the syntactic path between two tokens.

In [ ]:
```python
#helper fcn
def has_numbers(inputString):
    return any(char.isdigit() for char in inputString)
```

In [ ]:
```python
def extract_relations(filename, target_path):

    """ Extract new relations from a file.
    Input:
        - filename containing one document per line
        - target_path: the syntactic dependency path connecting the person en
    Output:
        - a list of (person, place, path, sentence) tuples in the same format

    """
    data=[]

    with open(filename) as file:
        for line in file:

            doc=line
            #print(doc)

            doc=nlp(doc)

            # Annotations are at the type level, so let's anchor them to spec
            for entity in doc.ents:
                for entity_2 in doc.ents:
                    if (entity != entity_2) and (not has_numbers(str(entity_2

                        if entity is not None and entity_2 is not None:

                            closest_person_mention=get_closest_coref(entity,
                            if closest_person_mention is not None:

                                path=get_path(closest_person_mention.root, en

                                if path == target_path:
                                    data.append((closest_person_mention, enti

    return list(set(data))
```

If you're executing this on Google Colab, download the wiki.bio.born.test.txt file here.

In [ ]:
```python
!wget https://raw.githubusercontent.com/dbamman/anlp21/master/data/wiki.bio.b
```

```
--2021-11-21 14:38:21--  https://raw.githubusercontent.com/dbamman/anlp21/mast
er/data/wiki.bio.born.test.txt
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111
.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.11
1.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 144574 (141K) [text/plain]
Saving to: 'wiki.bio.born.test.txt.8'

wiki.bio.born.test. 100%[===================>] 141.19K  --.-KB/s    in 0.02s

2021-11-21 14:38:22 (8.55 MB/s) - 'wiki.bio.born.test.txt.8' saved [144574/144
574]
```

In [ ]:
```python
#new_examples=extract_relations("wiki.bio.born.test.txt", "nsubjpass->born<-p
#for data in new_examples:
    #print ('\t'.join([str(x) for x in data]))
```

Q3: Execute `extract_relations` on `wiki.bio.born.test.txt` and the two most frequent paths identified in the training data above.

In [ ]:
```python
# most frequent path

new_examples=extract_relations("wiki.bio.born.test.txt", "nsubjpass->born<-pr
for data in new_examples:
    print ('\t'.join([str(x) for x in data]))
```

```
Gladstone       Liverpool        nsubjpass->born<-prep<-pobj     Gladstone was
born in Liverpool to Scottish parents .

Kerry   Aurora  nsubjpass->born<-prep<-pobj     Kerry was born in Aurora , Col
orado and attended boarding school in Massachusetts and New Hampshire .

He      Uetersen , Holstein      nsubjpass->born<-prep<-pobj     He was born in
Uetersen , Holstein , in present-day Germany .
He      Busseto nsubjpass->born<-prep<-pobj     He was born near Busseto to a
provincial family of moderate means , and developed a musical education with t
he help of a local patron .
Joel    Bronx   nsubjpass->born<-prep<-pobj     Joel was born in 1949 in The B
ronx , New York , and grew up on Long Island , New York , both places that inf
luenced his music .
Bonnet  Bassillac        nsubjpass->born<-prep<-pobj     Bonnet was born in Bas
sillac , Dordogne , the son of a lawyer .
Stengel Kansas City      nsubjpass->born<-prep<-pobj     Stengel was born in Ka
nsas City , Missouri .
Julia Jackson   Calcutta         nsubjpass->born<-prep<-pobj     Julia Jackson
was born in Calcutta to an Anglo-Indian family , and when she was two her moth
er and her two sisters moved back to England .
Rudd    Nambour nsubjpass->born<-prep<-pobj     Rudd was born in Nambour , Que
```

```
ensland .
Agnew    Baltimore       nsubjpass->born<-prep<-pobj      Agnew was born in Balt
imore , to an American-born mother and a Greek immigrant father .
He       Louisville      nsubjpass->born<-prep<-pobj      He was born in Louisvi
lle , Kentucky , to Jewish immigrant parents from Bohemia ( now in the Czech R
epublic ) , who raised him in a secular home .
Biden    Scranton        nsubjpass->born<-prep<-pobj      Biden was born in Scra
nton , Pennsylvania , in 1942 , and lived there for ten years before moving wi
th his family to Delaware .
Bush     New Haven       nsubjpass->born<-prep<-pobj      Bush was born in New H
aven , Connecticut , and grew up in Texas .
He       Germany nsubjpass->born<-prep<-pobj      He was born in Uetersen , Hols
tein , in present-day Germany .
Adams    Boston  nsubjpass->born<-prep<-pobj      Adams was born in Boston , bro
ught up in a religious and politically active family .

He       Dallas  nsubjpass->born<-prep<-pobj      He was born in Dallas , Texas
, but raised in Chattanooga , Tennessee until moving to Atlanta , Georgia .
Chopin   Duchy   nsubjpass->born<-prep<-pobj      Chopin was born Fryderyk Franc
iszek Chopin in the Duchy of Warsaw and grew up in Warsaw , which in 1815 beca
me part of Congress Poland .
Tryon    Bulwick Park     nsubjpass->born<-prep<-pobj      Tryon was born at Bulw
ick Park , Northamptonshire , England , the third son of Thomas Tryon and his
wife Anne Trollope .
Hanna    New Lisbon       nsubjpass->born<-prep<-pobj      Hanna was born in New
Lisbon ( today Lisbon ) , Ohio , in 1837 .
John Whitfield Bunn     Hunterdon County        nsubjpass->born<-prep<-pobj
John Whitfield Bunn was born June 21 , 1831 , in Hunterdon County , New Jersey
.
Tutu     Xhosa   nsubjpass->born<-prep<-pobj      Tutu was born of mixed Xhosa a
nd Motswana heritage to a poor family in Klerksdorp , British Imperial South A
frica .

Taft     Cincinnati      nsubjpass->born<-prep<-pobj      Taft was born in Cinci
nnati in 1857 .

Foraker Ohio    nsubjpass->born<-prep<-pobj      Foraker was born in rural Ohio
in 1846 , and enlisted at age 16 in the Union Army during the American Civil W
ar .
```

In [ ]:
```python
# second most frequent path

new_examples=extract_relations("wiki.bio.born.test.txt", "nsubjpass->born<-pre
for data in new_examples:
    print ('\t'.join([str(x) for x in data]))
```

```
Chopin  Warsaw  nsubjpass->born<-prep<-pobj<-prep<-pobj Chopin was born Fryder
yk Franciszek Chopin in the Duchy of Warsaw and grew up in Warsaw , which in 1
815 became part of Congress Poland .
He      Bohemia nsubjpass->born<-prep<-pobj<-prep<-pobj He was born in Louisvi
lle , Kentucky , to Jewish immigrant parents from Bohemia ( now in the Czech R
epublic ) , who raised him in a secular home .
Joel Stephen Kovel       Brooklyn         nsubjpass->born<-prep<-pobj<-prep<-pob
j        Joel Stephen Kovel was born on August 27 , 1936 , in Brooklyn , New Yo
rk .
Mundy   Hammersmith      nsubjpass->born<-prep<-pobj<-prep<-pobj Mundy was born
to a conservative middle-class family in Hammersmith , London .
```

In [ ]: