Technology Review

Xiaolong Yang

Topic Selection:

Recommendation combined Topic Models & Neural Language Models on Wikipedia data.

1.What is the Purpose of the project?

The tools aimed to build recommendation system for Wikipedia.

Wikipedia is a free online encyclopedia. It is one of the famous Big Text data source and used in many text mining related project, such as Google Knowledge Graph, Wikipedia Search Engine. I am building the tool to automatedly recommend relevant Wiki page on user's interested page, such as recommend President Trump on President Obama's page.

Each Wikipedia page is a topic, you can easily find related information within the topic. Take President Obama as example, you can find his personal details, early life and career, presidential campaigns, presidency and so on in his wiki page. Wikipedia use hyperlink in the page to indicate other topics related to current topic. In the President Obama example, you can find "United States presidential election, 2008" within the wiki page, which is highly related events. However, there are also meaningless topics such as "English", "Seattle", which is the language he knew or where he stayed for a while.

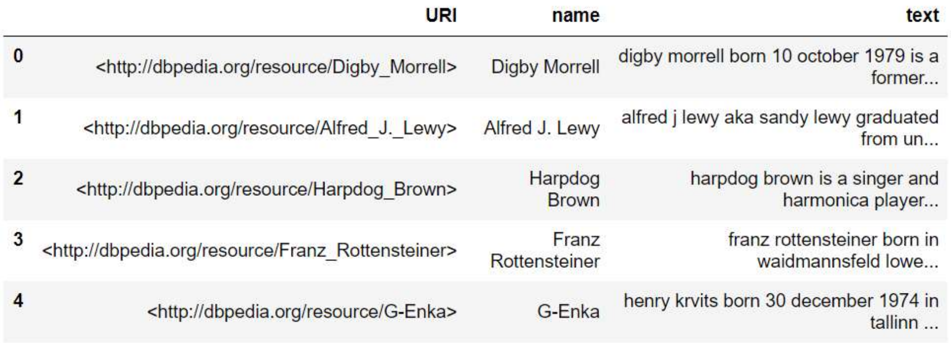
In this case, if I can develop a tool to show merely the highly related topics of user interested, it would save user's time to explore the document.

2. What is the data source?

In the beginning of the project, I tried to use the whole dataset from Wikipedia. https://dumps.wikimedia.org/enwiki/20170901/. However, the dataset is as large as 14 GB and I can merely finish my project based on the raw data. After then, I found there are 3 categories in the dataset: page-meta (All pages), page-article (Articles, templates, media/file descriptions, and primary meta-pages), abstract (extracted page abstracts for Yahoo). I tried to directly analyze the abstract data (4.9 GB in total), but it was hard to find took too much time for each iteration.

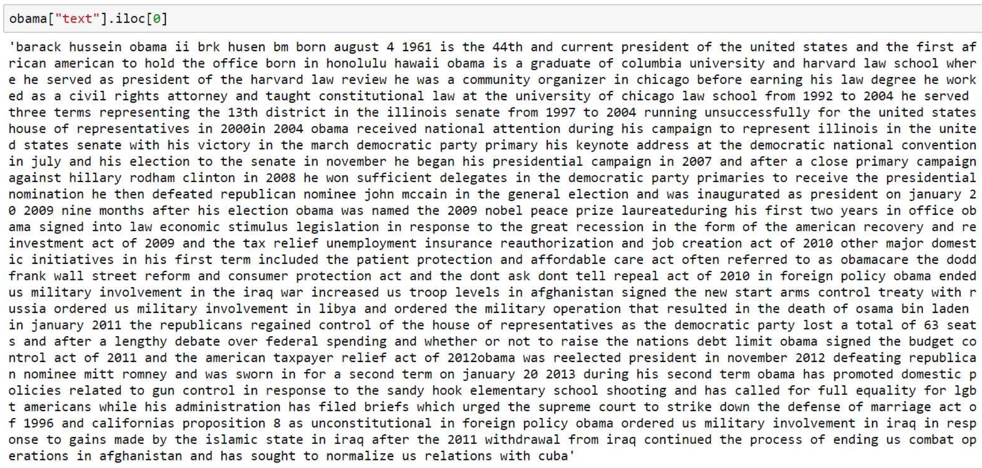
After then, I narrow my search to a manageable data source of Wikipedia, but still capable to recommend "Clinton" to "Obama". In other words, I tried to find dataset contained only "people" Wiki page. There is a data called "people_wiki.gl" from University of Washington https://d396qusza40orc.cloudfront.net/phoenixassets/people_wiki.csv, which met the requirement and only 150 MB.

There are totally 59071 people in the dataset, each people had 3 columns: URI, name, and text. The "Name" is the name of people, the "Text" is the wiki page related to the people. Here is an example how the data looks like:



	URI	name	text
0	<http://dbpedia.org/resource/Digby_Morrell>	Digby Morrell	digby morrell born 10 october 1979 is a former...
1	<http://dbpedia.org/resource/Alfred_J._Lewy>	Alfred J. Lewy	alfred j lewy aka sandy lewy graduated from un...
2	<http://dbpedia.org/resource/Harpdog_Brown>	Harpdog Brown	harpdog brown is a singer and harmonica player...
3	<http://dbpedia.org/resource/Franz_Rottensteiner>	Franz Rottensteiner	franz rottensteiner born in waidmannsfeld lowe...
4	<http://dbpedia.org/resource/G-Enka>	G-Enka	henry krvits born 30 december 1974 in tallinn ...

Here is an example of Barack Obama’s text:



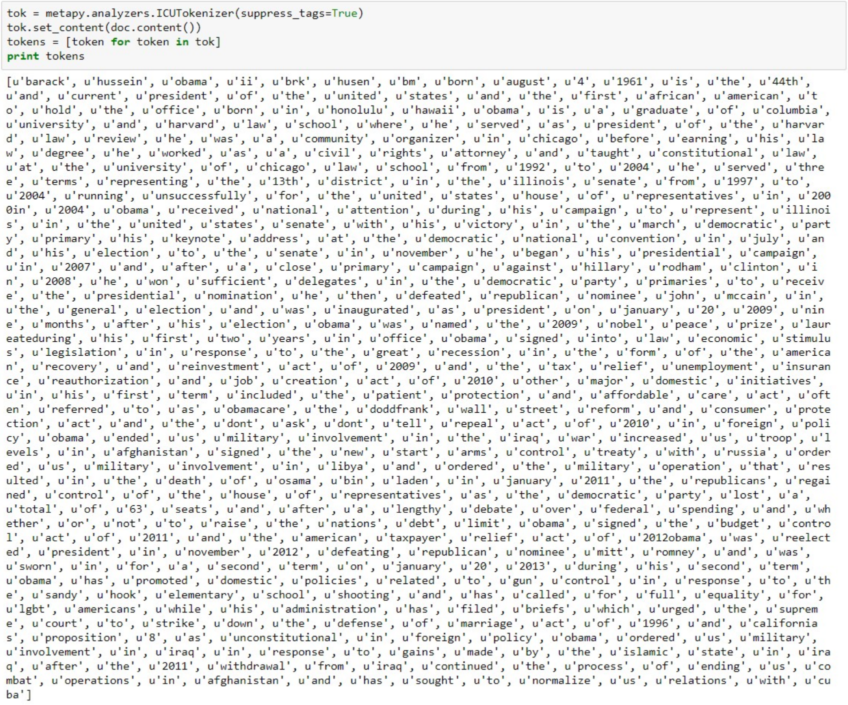
```
obama["text"].iloc[0]
```

```
'barack hussein obama ii brk husen bm born august 4 1961 is the 44th and current president of the united states and the first af
rican american to hold the office born in honolulu hawaii obama is a graduate of columbia university and harvard law school wher
e he served as president of the harvard law review he was a community organizer in chicago before earning his law degree he work
ed as a civil rights attorney and taught constitutional law at the university of chicago law school from 1992 to 2004 he served 
three terms representing the 13th district in the illinois senate from 1997 to 2004 running unsuccessfully for the united states
house of representatives in 2000in 2004 obama received national attention during his campaign to represent illinois in the unite
d states senate with his victory in the march democratic party primary his keynote address at the democratic national convention
in july and his election to the senate in november he began his presidential campaign in 2007 and after a close primary campaign
against hillary rodham clinton in 2008 he won sufficient delegates in the democratic party primaries to receive the presidential
nomination he then defeated republican nominee john mccain in the general election and was inaugurated as president on january 2
0 2009 nine months after his election obama was named the 2009 nobel peace prize laureateduring his first two years in office ob
ama signed into law economic stimulus legislation in response to the great recession in the form of the american recovery and re
investment act of 2009 and the tax relief unemployment insurance reauthorization and job creation act of 2010 other major domest
ic initiatives in his first term included the patient protection and affordable care act often referred to as obamacare the dodd
frank wall street reform and consumer protection act and the dont ask dont tell repeal act of 2010 in foreign policy obama ended
us military involvement in the iraq war increased us troop levels in afghanistan signed the new start arms control treaty with r
ussia ordered us military involvement in libya and ordered the military operation that resulted in the death of osama bin laden 
in january 2011 the republicans regained control of the house of representatives as the democratic party lost a total of 63 seat
s and after a lengthy debate over federal spending and whether or not to raise the nations debt limit obama signed the budget co
ntrol act of 2011 and the american taxpayer relief act of 2012obama was reelected president in november 2012 defeating republica
n nominee mitt romney and was sworn in for a second term on january 20 2013 during his second term obama has promoted domestic p
olicies related to gun control in response to the sandy hook elementary school shooting and has called for full equality for lgb
t americans while his administration has filed briefs which urged the supreme court to strike down the defense of marriage act o
f 1996 and californias proposition 8 as unconstitutional in foreign policy obama ordered us military involvement in iraq in resp
onse to gains made by the islamic state in iraq after the 2011 withdrawal from iraq continued the process of ending us combat op
erations in afghanistan and has sought to normalize us relations with cuba'
```

3. What tech used to text retrieval?

I followed the same steps as MP1, tokenize, remove stop words and stem, and uni-gram.

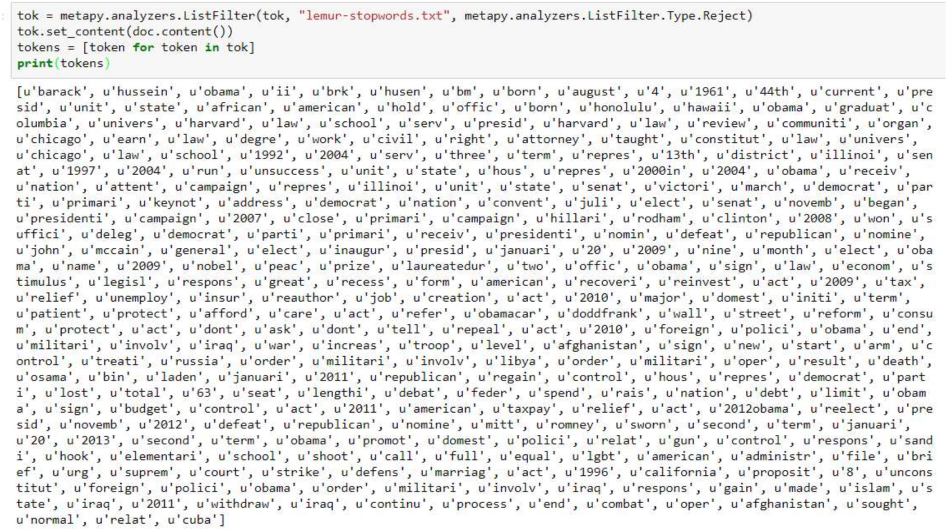
Here is an example of tokenize on Barack Obama’s text:



```
tok = metapy.analyzers.ICUTokenizer(suppress_tags=True)
tok.set_content(doc.content())
tokens = [token for token in tok]
print tokens
```

```
[u'barack', u'hussein', u'obama', u'ii', u'brk', u'husen', u'bm', u'born', u'august', u'4', u'1961', u'is', u'the', u'44th',
u'and', u'current', u'president', u'of', u'the', u'united', u'states', u'and', u'the', u'first', u'african', u'american', u't
o', u'hold', u'the', u'office', u'born', u'in', u'honolulu', u'hawaii', u'obama', u'is', u'a', u'graduate', u'of', u'columbia',
u'university', u'and', u'harvard', u'law', u'school', u'where', u'he', u'served', u'as', u'president', u'of', u'the', u'harvar
d', u'law', u'review', u'he', u'was', u'a', u'community', u'organizer', u'in', u'chicago', u'before', u'earning', u'his', u'la
w', u'degree', u'he', u'worked', u'as', u'a', u'civil', u'rights', u'attorney', u'and', u'taught', u'constitutional', u'law',
u'at', u'the', u'university', u'of', u'chicago', u'law', u'school', u'from', u'1992', u'to', u'2004', u'he', u'served', u'thre
e', u'terms', u'representing', u'the', u'13th', u'district', u'in', u'the', u'illinois', u'senate', u'from', u'1997', u'to',
u'2004', u'running', u'unsuccessfully', u'for', u'the', u'united', u'states', u'house', u'of', u'representatives', u'in', u'200
0in', u'2004', u'obama', u'received', u'national', u'attention', u'during', u'his', u'campaign', u'to', u'represent', u'illinoi
s', u'in', u'the', u'united', u'states', u'senate', u'with', u'his', u'victory', u'in', u'the', u'march', u'democratic', u'part
y', u'primary', u'his', u'keynote', u'address', u'at', u'the', u'democratic', u'national', u'convention', u'in', u'july', u'an
d', u'his', u'election', u'to', u'the', u'senate', u'in', u'november', u'he', u'began', u'his', u'presidential', u'campaign',
u'in', u'2007', u'and', u'after', u'a', u'close', u'primary', u'campaign', u'against', u'hillary', u'rodham', u'clinton', u'i
n', u'2008', u'he', u'won', u'sufficient', u'delegates', u'in', u'the', u'democratic', u'party', u'primaries', u'to', u'receiv
e', u'the', u'presidential', u'nomination', u'he', u'then', u'defeated', u'republican', u'nominee', u'john', u'mccain', u'in',
u'the', u'general', u'election', u'and', u'was', u'inaugurated', u'as', u'president', u'on', u'january', u'20', u'2009', u'nin
e', u'months', u'after', u'his', u'election', u'obama', u'was', u'named', u'the', u'2009', u'nobel', u'peace', u'prize', u'laur
eateduring', u'his', u'first', u'two', u'years', u'in', u'office', u'obama', u'signed', u'into', u'law', u'economic', u'stimulu
s', u'legislation', u'in', u'response', u'to', u'the', u'great', u'recession', u'in', u'the', u'form', u'of', u'the', u'america
n', u'recovery', u'and', u'reinvestment', u'act', u'of', u'2009', u'and', u'the', u'tax', u'relief', u'unemployment', u'insuran
ce', u'reauthorization', u'and', u'job', u'creation', u'act', u'of', u'2010', u'other', u'major', u'domestic', u'initiatives',
u'in', u'his', u'first', u'term', u'included', u'the', u'patient', u'protection', u'and', u'affordable', u'care', u'act', u'oft
en', u'referred', u'to', u'as', u'obamacare', u'the', u'doddfrank', u'wall', u'street', u'reform', u'and', u'consumer', u'prote
ction', u'act', u'and', u'the', u'dont', u'ask', u'dont', u'tell', u'repeal', u'act', u'of', u'2010', u'in', u'foreign', u'poli
cy', u'obama', u'ended', u'us', u'military', u'involvement', u'in', u'the', u'iraq', u'war', u'increased', u'us', u'troop', u'l
evels', u'in', u'afghanistan', u'signed', u'the', u'new', u'start', u'arms', u'control', u'treaty', u'with', u'russia', u'order
ed', u'us', u'military', u'involvement', u'in', u'libya', u'and', u'ordered', u'the', u'military', u'operation', u'that', u'res
ulted', u'in', u'the', u'death', u'of', u'osama', u'bin', u'laden', u'in', u'january', u'2011', u'the', u'republicans', u'regai
ned', u'control', u'of', u'the', u'house', u'of', u'representatives', u'as', u'the', u'democratic', u'party', u'lost', u'a',
u'total', u'of', u'63', u'seats', u'and', u'after', u'a', u'lengthy', u'debate', u'over', u'federal', u'spending', u'and', u'wh
ether', u'or', u'not', u'to', u'raise', u'the', u'nations', u'debt', u'limit', u'obama', u'signed', u'the', u'budget', u'contro
l', u'act', u'of', u'2011', u'and', u'the', u'american', u'taxpayer', u'relief', u'act', u'of', u'2012obama', u'was', u'reelect
ed', u'president', u'in', u'november', u'2012', u'defeating', u'republican', u'nominee', u'mitt', u'romney', u'and', u'was',
u'sworn', u'in', u'for', u'a', u'second', u'term', u'on', u'january', u'20', u'2013', u'during', u'his', u'second', u'term',
u'obama', u'has', u'promoted', u'domestic', u'policies', u'related', u'to', u'gun', u'control', u'in', u'response', u'to', u'th
e', u'sandy', u'hook', u'elementary', u'school', u'shooting', u'and', u'has', u'called', u'for', u'full', u'equality', u'for',
u'lgbt', u'americans', u'while', u'his', u'administration', u'has', u'filed', u'briefs', u'which', u'urged', u'the', u'suprem
e', u'court', u'to', u'strike', u'down', u'the', u'defense', u'of', u'marriage', u'act', u'of', u'1996', u'and', u'california
s', u'proposition', u'8', u'as', u'unconstitutional', u'in', u'foreign', u'policy', u'obama', u'ordered', u'us', u'military',
u'involvement', u'in', u'iraq', u'in', u'response', u'to', u'gains', u'made', u'by', u'the', u'islamic', u'state', u'in', u'ira
q', u'after', u'the', u'2011', u'withdrawal', u'from', u'iraq', u'continued', u'the', u'process', u'of', u'ending', u'us', u'co
mbat', u'operations', u'in', u'afghanistan', u'and', u'has', u'sought', u'to', u'normalize', u'us', u'relations', u'with', u'cu
ba']
```

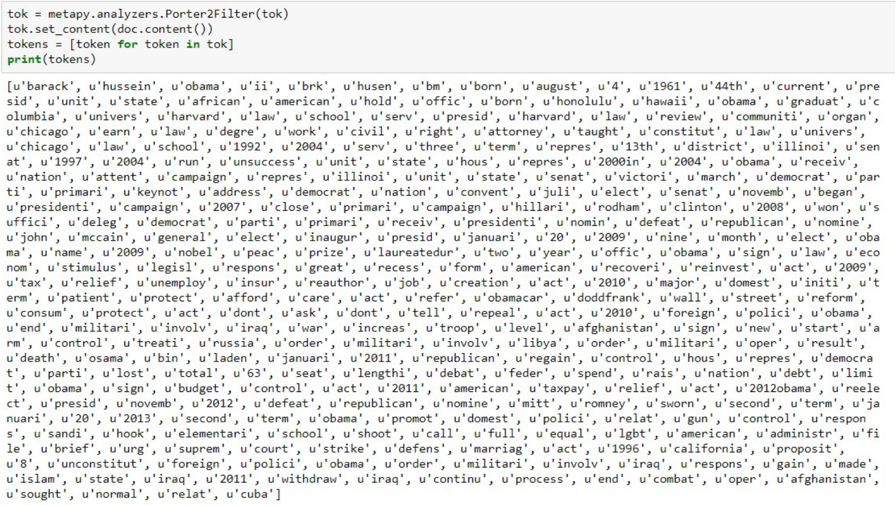
Here is an example of stop words removal on Barack Obama’s text:



```
tok = metapy.analyzers.ListFilter(tok, "lemur-stopwords.txt", metapy.analyzers.ListFilter.Type.Reject)
tok.set_content(doc.content())
tokens = [token for token in tok]
print(tokens)
```

```
[u'barack', u'hussein', u'obama', u'ii', u'brk', u'husen', u'bm', u'born', u'august', u'4', u'1961', u'44th', u'current', u'pre
sid', u'unit', u'state', u'african', u'american', u'hold', u'offic', u'born', u'honolulu', u'hawaii', u'obama', u'graduat', u'c
olumbia', u'univers', u'harvard', u'law', u'school', u'serv', u'presid', u'harvard', u'law', u'review', u'communiti', u'organ',
u'chicago', u'earn', u'law', u'degre', u'work', u'civil', u'right', u'attorney', u'taught', u'constitut', u'law', u'univers',
u'chicago', u'law', u'school', u'1992', u'2004', u'serv', u'three', u'term', u'repres', u'13th', u'district', u'illinoi', u'sen
at', u'1997', u'2004', u'run', u'unsuccess', u'unit', u'state', u'hous', u'repres', u'2000in', u'2004', u'obama', u'receiv',
u'nation', u'attent', u'campaign', u'repres', u'illinoi', u'unit', u'state', u'senat', u'victori', u'march', u'democrat', u'par
ti', u'primari', u'keynot', u'address', u'democrat', u'nation', u'convent', u'juli', u'elect', u'senat', u'novemb', u'began',
u'presidenti', u'campaign', u'2007', u'close', u'primari', u'campaign', u'hillari', u'rodham', u'clinton', u'2008', u'won', u's
uffici', u'deleg', u'democrat', u'parti', u'primari', u'receiv', u'presidenti', u'nomin', u'defeat', u'republican', u'nomine',
u'john', u'mccain', u'general', u'elect', u'inaugur', u'presid', u'januari', u'20', u'2009', u'nine', u'month', u'elect', u'oba
ma', u'name', u'2009', u'nobel', u'peac', u'prize', u'laureatedur', u'two', u'offic', u'obama', u'sign', u'law', u'econom', u's
timulus', u'legisl', u'respons', u'great', u'recess', u'form', u'american', u'recoveri', u'reinvest', u'act', u'2009', u'tax',
u'relief', u'unemploy', u'insur', u'reauthor', u'job', u'creation', u'act', u'2010', u'major', u'domest', u'initi', u'term',
u'patient', u'protect', u'afford', u'care', u'act', u'refer', u'obamacar', u'doddfrank', u'wall', u'street', u'reform', u'consu
m', u'protect', u'act', u'dont', u'ask', u'dont', u'tell', u'repeal', u'act', u'2010', u'foreign', u'polici', u'obama', u'end',
u'militari', u'involv', u'iraq', u'war', u'increas', u'troop', u'level', u'afghanistan', u'sign', u'new', u'start', u'arm', u'c
ontrol', u'treati', u'russia', u'order', u'militari', u'involv', u'libya', u'order', u'militari', u'oper', u'result', u'death',
u'osama', u'bin', u'laden', u'januari', u'2011', u'republican', u'regain', u'control', u'hous', u'repres', u'democrat', u'part
i', u'lost', u'total', u'63', u'seat', u'lengthi', u'debat', u'feder', u'spend', u'rais', u'nation', u'debt', u'limit', u'obam
a', u'sign', u'budget', u'control', u'act', u'2011', u'american', u'taxpay', u'relief', u'act', u'2012obama', u'reelect', u'pre
sid', u'novemb', u'2012', u'defeat', u'republican', u'nomine', u'mitt', u'romney', u'sworn', u'second', u'term', u'januari',
u'20', u'2013', u'second', u'term', u'obama', u'promot', u'domest', u'polici', u'relat', u'gun', u'control', u'respons', u'sand
i', u'hook', u'elementari', u'school', u'shoot', u'call', u'full', u'equal', u'lgbt', u'american', u'administr', u'file', u'bri
ef', u'urg', u'suprem', u'court', u'strike', u'defens', u'marriag', u'act', u'1996', u'california', u'proposit', u'8', u'uncons
titut', u'foreign', u'polici', u'obama', u'order', u'militari', u'involv', u'iraq', u'respons', u'gain', u'made', u'islam', u's
tate', u'iraq', u'2011', u'withdraw', u'iraq', u'continu', u'process', u'end', u'combat', u'oper', u'afghanistan', u'sought',
u'normal', u'relat', u'cuba']
```

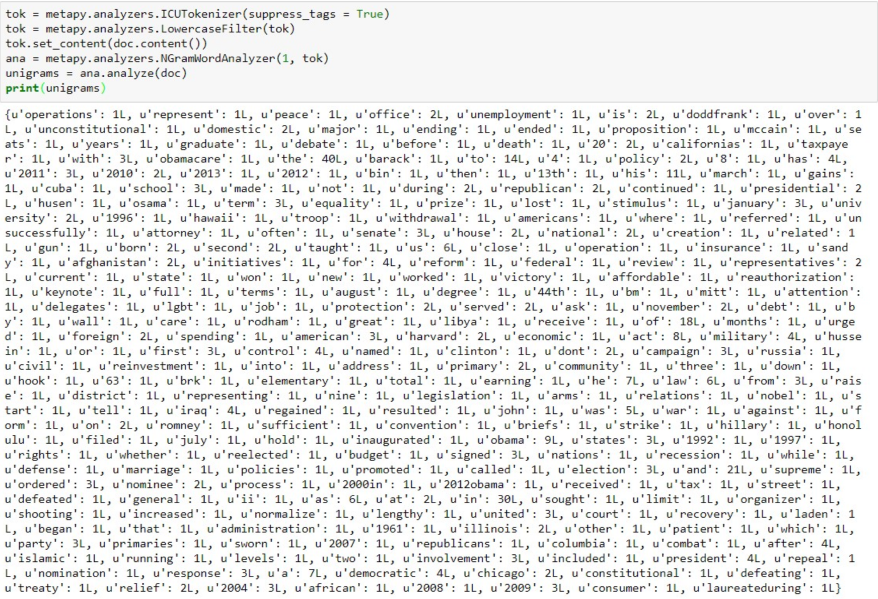
Here is an example of stemming on Barack Obama’s text:



```
tok = metapy.analyzers.Porter2Filter(tok)
tok.set_content(doc.content())
tokens = [token for token in tok]
print(tokens)
```

```
[u'barack', u'hussein', u'obama', u'ii', u'brk', u'husen', u'bm', u'born', u'august', u'4', u'1961', u'44th', u'current', u'pre
sid', u'unit', u'state', u'african', u'american', u'hold', u'offic', u'born', u'honolulu', u'hawaii', u'obama', u'graduat', u'c
olumbia', u'univers', u'harvard', u'law', u'school', u'serv', u'presid', u'harvard', u'law', u'review', u'communiti', u'organ',
u'chicago', u'earn', u'law', u'degre', u'work', u'civil', u'right', u'attorney', u'taught', u'constitut', u'law', u'univers',
u'chicago', u'law', u'school', u'1992', u'2004', u'serv', u'three', u'term', u'repres', u'13th', u'district', u'illinoi', u'sen
at', u'1997', u'2004', u'run', u'unsuccess', u'unit', u'state', u'hous', u'repres', u'2000in', u'2004', u'obama', u'receiv',
u'nation', u'attent', u'campaign', u'repres', u'illinoi', u'unit', u'state', u'senat', u'victori', u'march', u'democrat', u'par
ti', u'primari', u'keynot', u'address', u'democrat', u'nation', u'convent', u'juli', u'elect', u'senat', u'novemb', u'began',
u'presidenti', u'campaign', u'2007', u'close', u'primari', u'campaign', u'hillari', u'rodham', u'clinton', u'2008', u'won', u's
uffici', u'deleg', u'democrat', u'parti', u'primari', u'receiv', u'presidenti', u'nomin', u'defeat', u'republican', u'nomine',
u'john', u'mccain', u'general', u'elect', u'inaugur', u'presid', u'januari', u'20', u'2009', u'nine', u'month', u'elect', u'oba
ma', u'name', u'2009', u'nobel', u'peac', u'prize', u'laureatedur', u'two', u'year', u'offic', u'obama', u'sign', u'law', u'eco
nom', u'stimulus', u'legisl', u'respons', u'great', u'recess', u'form', u'american', u'recoveri', u'reinvest', u'act', u'2009',
u'tax', u'relief', u'unemploy', u'insur', u'reauthor', u'job', u'creation', u'act', u'2010', u'major', u'domest', u'initi', u't
erm', u'patient', u'protect', u'afford', u'care', u'act', u'refer', u'obamacar', u'doddfrank', u'wall', u'street', u'reform',
u'consum', u'protect', u'act', u'dont', u'ask', u'dont', u'tell', u'repeal', u'act', u'2010', u'foreign', u'polici', u'obama',
u'end', u'militari', u'involv', u'iraq', u'war', u'increas', u'troop', u'level', u'afghanistan', u'sign', u'new', u'start', u'a
rm', u'control', u'treati', u'russia', u'order', u'militari', u'involv', u'libya', u'order', u'militari', u'oper', u'result',
u'death', u'osama', u'bin', u'laden', u'januari', u'2011', u'republican', u'regain', u'control', u'hous', u'repres', u'democra
t', u'parti', u'lost', u'total', u'63', u'seat', u'lengthi', u'debat', u'feder', u'spend', u'rais', u'nation', u'debt', u'limi
t', u'obama', u'sign', u'budget', u'control', u'act', u'2011', u'american', u'taxpay', u'relief', u'act', u'2012obama', u'reele
ct', u'presid', u'novemb', u'2012', u'defeat', u'republican', u'nomine', u'mitt', u'romney', u'sworn', u'second', u'term', u'ja
nuari', u'20', u'2013', u'second', u'term', u'obama', u'promot', u'domest', u'polici', u'relat', u'gun', u'control', u'respon
s', u'sandi', u'hook', u'elementari', u'school', u'shoot', u'call', u'full', u'equal', u'lgbt', u'american', u'administr', u'fi
le', u'brief', u'urg', u'suprem', u'court', u'strike', u'defens', u'marriag', u'act', u'1996', u'california', u'proposit',
u'8', u'unconstitut', u'foreign', u'polici', u'obama', u'order', u'militari', u'involv', u'iraq', u'respons', u'gain', u'made',
u'islam', u'state', u'iraq', u'2011', u'withdraw', u'iraq', u'continu', u'process', u'end', u'combat', u'oper', u'afghanistan',
u'sought', u'normal', u'relat', u'cuba']
```

Here is an example of uni-gram on Barack Obama's text:



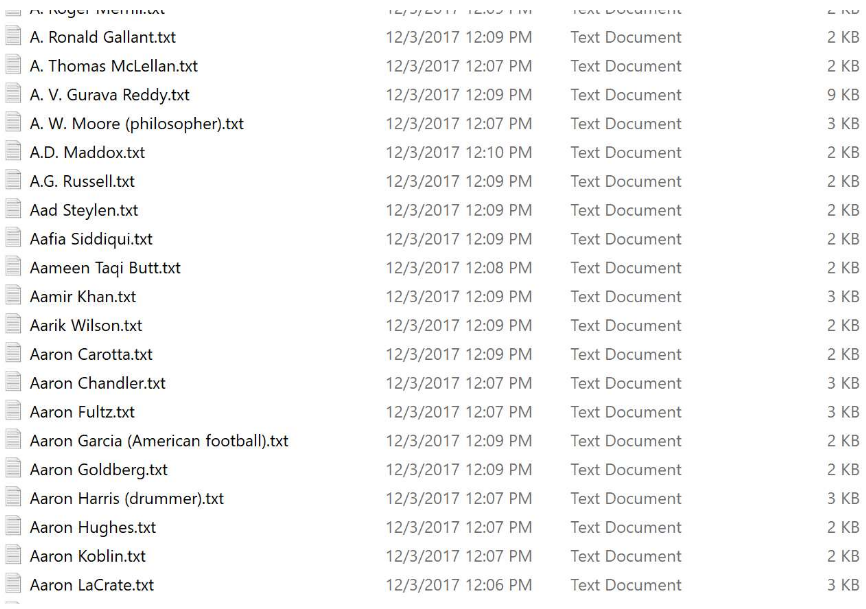
```
tok = metapy.analyzers.ICUTokenizer(suppress_tags = True)
tok = metapy.analyzers.LowercaseFilter(tok)
tok.set_content(doc.content())
ana = metapy.analyzers.NGramWordAnalyzer(1, tok)
unigrams = ana.analyze(doc)
print(unigrams)
```

```
{u'operations': 1L, u'represent': 1L, u'peace': 1L, u'office': 2L, u'unemployment': 1L, u'is': 2L, u'doddfrank': 1L, u'over': 1
L, u'unconstitutional': 1L, u'domestic': 2L, u'major': 1L, u'ending': 1L, u'ended': 1L, u'proposition': 1L, u'mccain': 1L, u'se
ats': 1L, u'years': 1L, u'graduate': 1L, u'debate': 1L, u'before': 1L, u'death': 1L, u'20': 2L, u'californias': 1L, u'taxpaye
r': 1L, u'with': 3L, u'obamacare': 1L, u'the': 40L, u'barack': 1L, u'to': 14L, u'4': 1L, u'policy': 2L, u'8': 1L, u'has': 4L,
u'2011': 3L, u'2010': 2L, u'2013': 1L, u'2012': 1L, u'bin': 1L, u'then': 1L, u'13th': 1L, u'his': 11L, u'march': 1L, u'gains':
1L, u'cuba': 1L, u'school': 3L, u'made': 1L, u'not': 1L, u'during': 2L, u'republican': 2L, u'continued': 1L, u'presidential': 2
L, u'husen': 1L, u'osama': 1L, u'term': 3L, u'equality': 1L, u'prize': 1L, u'lost': 1L, u'stimulus': 1L, u'january': 3L, u'univ
ersity': 2L, u'1996': 1L, u'hawaii': 1L, u'troop': 1L, u'withdrawal': 1L, u'americans': 1L, u'where': 1L, u'referred': 1L, u'un
successfully': 1L, u'attorney': 1L, u'often': 1L, u'senate': 3L, u'house': 2L, u'national': 2L, u'creation': 1L, u'related': 1
L, u'gun': 1L, u'born': 2L, u'second': 2L, u'taught': 1L, u'us': 6L, u'close': 1L, u'operation': 1L, u'insurance': 1L, u'sand
y': 1L, u'afghanistan': 2L, u'initiatives': 1L, u'for': 4L, u'reform': 1L, u'federal': 1L, u'review': 1L, u'representatives': 2
L, u'current': 1L, u'state': 1L, u'won': 1L, u'new': 1L, u'worked': 1L, u'victory': 1L, u'affordable': 1L, u'reauthorization':
1L, u'keynote': 1L, u'full': 1L, u'terms': 1L, u'august': 1L, u'degree': 1L, u'44th': 1L, u'bm': 1L, u'mitt': 1L, u'attention':
1L, u'delegates': 1L, u'lgbt': 1L, u'job': 1L, u'protection': 2L, u'served': 2L, u'ask': 1L, u'november': 2L, u'debt': 1L, u'b
y': 1L, u'wall': 1L, u'care': 1L, u'rodham': 1L, u'great': 1L, u'libya': 1L, u'receive': 1L, u'of': 18L, u'months': 1L, u'urge
d': 1L, u'foreign': 2L, u'spending': 1L, u'american': 3L, u'harvard': 2L, u'economic': 1L, u'act': 8L, u'military': 4L, u'husse
in': 1L, u'or': 1L, u'first': 3L, u'control': 4L, u'named': 1L, u'clinton': 1L, u'dont': 2L, u'campaign': 3L, u'russia': 1L,
u'civil': 1L, u'reinvestment': 1L, u'into': 1L, u'address': 1L, u'primary': 2L, u'community': 1L, u'three': 1L, u'down': 1L,
u'hook': 1L, u'63': 1L, u'brk': 1L, u'elementary': 1L, u'total': 1L, u'earning': 1L, u'he': 7L, u'law': 6L, u'from': 3L, u'rais
e': 1L, u'district': 1L, u'representing': 1L, u'nine': 1L, u'legislation': 1L, u'arms': 1L, u'relations': 1L, u'nobel': 1L, u's
tart': 1L, u'tell': 1L, u'iraq': 4L, u'regained': 1L, u'resulted': 1L, u'john': 1L, u'was': 5L, u'war': 1L, u'against': 1L, u'f
orm': 1L, u'on': 2L, u'romney': 1L, u'sufficient': 1L, u'convention': 1L, u'briefs': 1L, u'strike': 1L, u'hillary': 1L, u'honol
ulu': 1L, u'filed': 1L, u'july': 1L, u'hold': 1L, u'inaugurated': 1L, u'obama': 9L, u'states': 3L, u'1992': 1L, u'1997': 1L,
u'rights': 1L, u'whether': 1L, u'reelected': 1L, u'budget': 1L, u'signed': 3L, u'nations': 1L, u'recession': 1L, u'while': 1L,
u'defense': 1L, u'marriage': 1L, u'policies': 1L, u'promoted': 1L, u'called': 1L, u'election': 3L, u'and': 21L, u'supreme': 1L,
u'ordered': 3L, u'nominee': 2L, u'process': 1L, u'2000in': 1L, u'2012obama': 1L, u'received': 1L, u'tax': 1L, u'street': 1L,
u'defeated': 1L, u'general': 1L, u'ii': 1L, u'as': 6L, u'at': 2L, u'in': 30L, u'sought': 1L, u'limit': 1L, u'organizer': 1L,
u'shooting': 1L, u'increased': 1L, u'normalize': 1L, u'lengthy': 1L, u'united': 3L, u'court': 1L, u'recovery': 1L, u'laden': 1
L, u'began': 1L, u'that': 1L, u'administration': 1L, u'1961': 1L, u'illinois': 2L, u'other': 1L, u'patient': 1L, u'which': 1L,
u'party': 3L, u'primaries': 1L, u'sworn': 1L, u'2007': 1L, u'republicans': 1L, u'columbia': 1L, u'combat': 1L, u'after': 4L,
u'islamic': 1L, u'running': 1L, u'levels': 1L, u'two': 1L, u'involvement': 3L, u'included': 1L, u'president': 4L, u'repeal': 1
L, u'nomination': 1L, u'response': 3L, u'a': 7L, u'democratic': 4L, u'chicago': 2L, u'constitutional': 1L, u'defeating': 1L,
u'treaty': 1L, u'relief': 2L, u'2004': 3L, u'african': 1L, u'2008': 1L, u'2009': 3L, u'consumer': 1L, u'laureateduring': 1L}
```

3. How to transfer data to meet Metapy requirement?

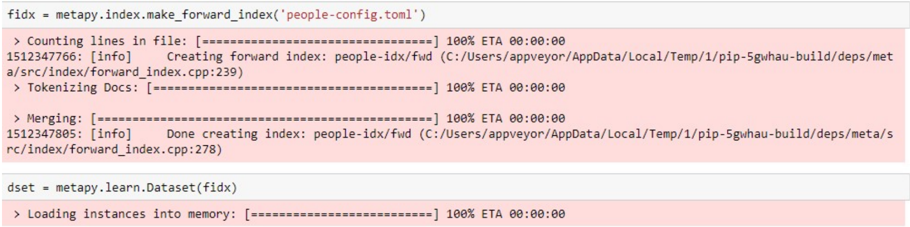
Metapy stored all process data in disk index. Before we started modeling, we need first index the data. However, since the data I got was stored as a big data frame, the first step is to divide the data into different text file.

Take a glimpse how it looks like:



A. Ronald Gallant.txt	12/3/2017 12:09 PM	Text Document	2 KB
A. Thomas McLellan.txt	12/3/2017 12:07 PM	Text Document	2 KB
A. V. Gurava Reddy.txt	12/3/2017 12:09 PM	Text Document	9 KB
A. W. Moore (philosopher).txt	12/3/2017 12:07 PM	Text Document	3 KB
A.D. Maddox.txt	12/3/2017 12:10 PM	Text Document	2 KB
A.G. Russell.txt	12/3/2017 12:09 PM	Text Document	2 KB
Aad Steylen.txt	12/3/2017 12:09 PM	Text Document	2 KB
Aafia Siddiqui.txt	12/3/2017 12:09 PM	Text Document	2 KB
Aameen Taqi Butt.txt	12/3/2017 12:08 PM	Text Document	2 KB
Aamir Khan.txt	12/3/2017 12:09 PM	Text Document	3 KB
Aarik Wilson.txt	12/3/2017 12:09 PM	Text Document	2 KB
Aaron Carotta.txt	12/3/2017 12:09 PM	Text Document	2 KB
Aaron Chandler.txt	12/3/2017 12:07 PM	Text Document	3 KB
Aaron Fultz.txt	12/3/2017 12:07 PM	Text Document	3 KB
Aaron Garcia (American football).txt	12/3/2017 12:09 PM	Text Document	2 KB
Aaron Goldberg.txt	12/3/2017 12:09 PM	Text Document	2 KB
Aaron Harris (drummer).txt	12/3/2017 12:07 PM	Text Document	3 KB
Aaron Hughes.txt	12/3/2017 12:07 PM	Text Document	2 KB
Aaron Koblin.txt	12/3/2017 12:07 PM	Text Document	2 KB
Aaron LaCrate.txt	12/3/2017 12:06 PM	Text Document	3 KB

And then combined into a data called data.dat. I also defined config files in the people.toml and transfer the data into index format:



```
fidx = metapy.index.make_forward_index('people-config.toml')
```

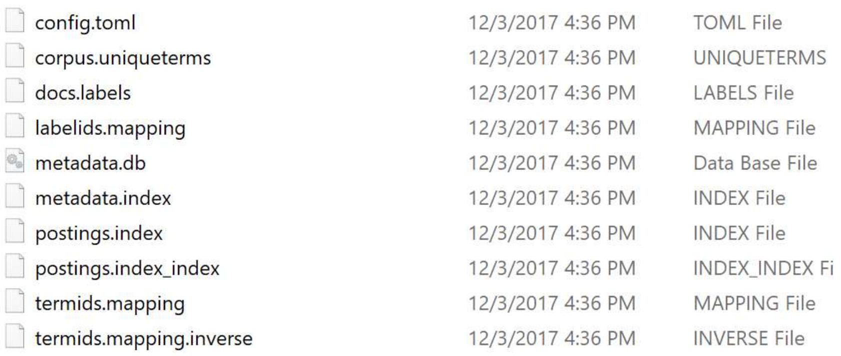
```
 > Counting lines in file: [=================================] 100% ETA 00:00:00
1512347766: [info]     Creating forward index: people-idx/fwd (C:/Users/appveyor/AppData/Local/Temp/1/pip-5gwhau-build/deps/met
a/src/index/forward_index.cpp:239)
 > Tokenizing Docs: [========================================] 100% ETA 00:00:00

 > Merging: [================================================] 100% ETA 00:00:00
1512347805: [info]     Done creating index: people-idx/fwd (C:/Users/appveyor/AppData/Local/Temp/1/pip-5gwhau-build/deps/meta/s
rc/index/forward_index.cpp:278)
```

```
dset = metapy.learn.Dataset(fidx)
```

```
 > Loading instances into memory: [==========================] 100% ETA 00:00:00
```

Here is how it looks like in file system:



config.toml	12/3/2017 4:36 PM	TOML File
corpus.uniqueterms	12/3/2017 4:36 PM	UNIQUETERMS
docs.labels	12/3/2017 4:36 PM	LABELS File
labelids.mapping	12/3/2017 4:36 PM	MAPPING File
metadata.db	12/3/2017 4:36 PM	Data Base File
metadata.index	12/3/2017 4:36 PM	INDEX File
postings.index	12/3/2017 4:36 PM	INDEX File
postings.index_index	12/3/2017 4:36 PM	INDEX_INDEX Fi
termids.mapping	12/3/2017 4:36 PM	MAPPING File
termids.mapping.inverse	12/3/2017 4:36 PM	INVERSE File

4. What tech used for recommendation?

I built my recommender with two techs, Topic Models & Neural Language Models. The main difference between these models is the contextual information they use: LSA and topic models use *documents* as contexts, and Neural language models and distributional semantic models instead use *words* as contexts. These different contextual representations capture different types of semantic similarity; the document-based models capture semantic relatedness (e.g. “boat” – “water”) while the word-based models capture semantic similarity (e.g. “boat” – “ship”).

For President Obama example, I tried to find “Politics” topic using topic models, and “Trump”, “Clinton” words using neural language models.

I tried to do the same thing as the following documents, and it took so much time for training. Right now, I was still working on pruning the parameters for better result.

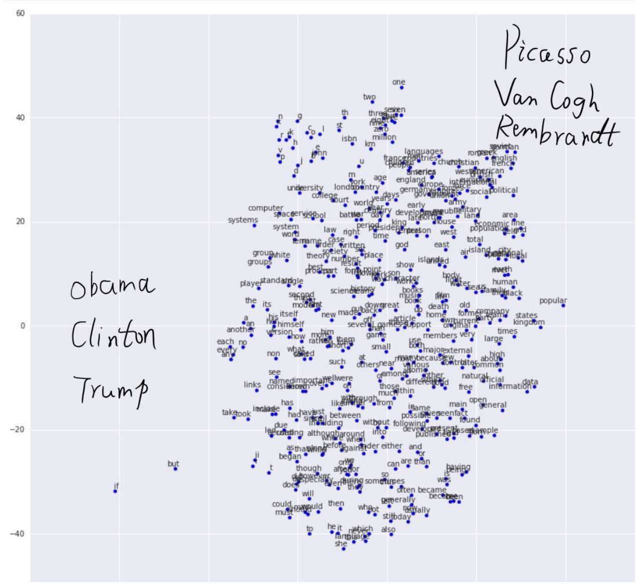
Meta for topic-modeling Example
https://github.com/meta-toolkit/metapy/blob/master/tutorials/5-topic-modeling.ipynb
Meta for Word2vec Example
https://meta-toolkit.org/word-embeddings.html
Tensor flow Word2vec Example:
https://github.com/tensorflow/tensorflow/blob/master/tensorflow/examples/udacity/5_word2vec.ipynb

I tried to work on both LDA and word embedding on Metapy, however, MeTA did not have the tutorial to implement the word embedding on python. Thus, I chose to use Tensor flow to train the model.

Expected result for topic modeling

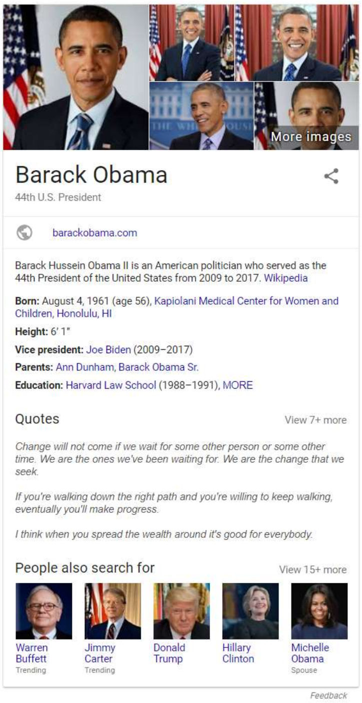
model.topic_distribution(“obama”) {Artist: 0.021341, Politic: 0.978659}
model.topic_distribution(“Picasso”) {Artist: 0.978659, Politic: 0.021341}

Expected result for word2vec



5. How to represent the result?

The google graph show the result as:



Barack Obama

44th U.S. President

barackobama.com

Barack Hussein Obama II is an American politician who served as the 44th President of the United States from 2009 to 2017. Wikipedia

Born: August 4, 1961 (age 56), Kapiolani Medical Center for Women and Children, Honolulu, HI

Height: 6′ 1″

Vice president: Joe Biden (2009–2017)

Parents: Ann Dunham, Barack Obama Sr.

Education: Harvard Law School (1988–1991), MORE

Quotes View 7+ more

Change will not come if we wait for some other person or some other time. We are the ones we've been waiting for. We are the change that we seek.

If you're walking down the right path and you're willing to keep walking, eventually you'll make progress.

I think when you spread the wealth around it's good for everybody.

People also search for View 15+ more

Warren Buffett
Trending

Jimmy Carter
Trending

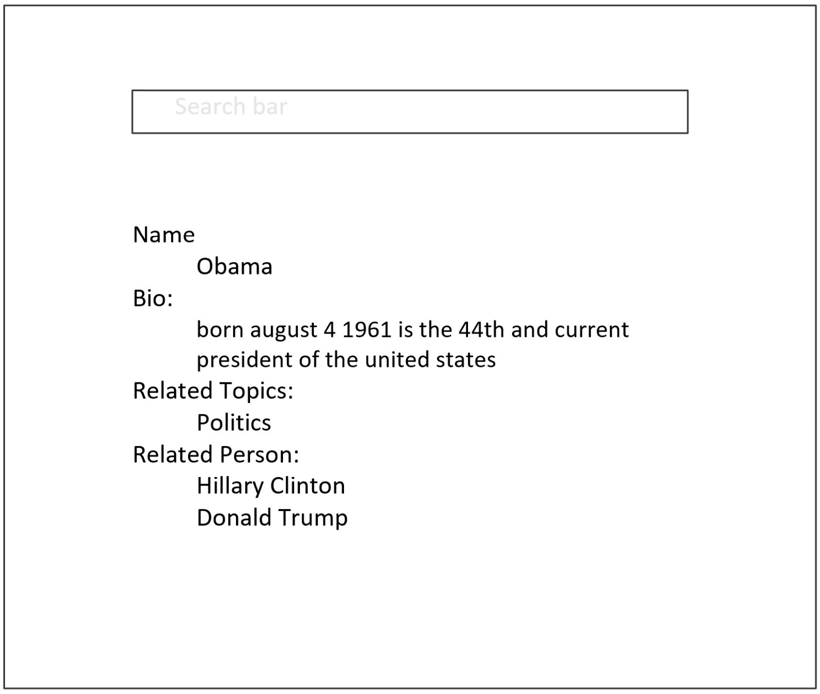
Donald Trump

Hillary Clinton

Michelle Obama
Spouse

Feedback

I would use Bootstrap to show the result as:



Search bar

Name
    Obama
Bio:
    born august 4 1961 is the 44th and current president of the united states
Related Topics:
    Politics
Related Person:
    Hillary Clinton
    Donald Trump