Document retrieval from Wikipedia data

In [3]:
```
import turicreate
```

Load some text data from Wikipedia

In [4]:
```
people = turicreate.SFrame('people_wiki.sframe')
```

In [5]:
```
people
```

Out[5]:

URI	name	text
<http://dbpedia.org/resource/Digby_Morrell> ...	Digby Morrell	digby morrell born 10 october 1979 is a former ...
<http://dbpedia.org/resource/Alfred_J._Lewy> ...	Alfred J. Lewy	alfred j lewy aka sandy lewy graduated from ...
<http://dbpedia.org/resource/Harpdog_Brown> ...	Harpdog Brown	harpdog brown is a singer and harmonica player who ...
<http://dbpedia.org/resource/Franz_Rottensteiner> ...	Franz Rottensteiner	franz rottensteiner born in waidmannsfeld lower ...
<http://dbpedia.org/resource/G-Enka> ...	G-Enka	henry krvits born 30 december 1974 in tallinn ...
<http://dbpedia.org/resource/Sam_Henderson> ...	Sam Henderson	sam henderson born october 18 1969 is an ...
<http://dbpedia.org/resource/Aaron_LaCrate> ...	Aaron LaCrate	aaron lacrate is an american music producer ...
<http://dbpedia.org/resource/Trevor_Ferguson> ...	Trevor Ferguson	trevor ferguson aka john farrow born 11 november ...
<http://dbpedia.org/resource/Grant_Nelson> ...	Grant Nelson	grant nelson born 27 april 1971 in london ...
<http://dbpedia.org/resource/Cathy_Caruth> ...	Cathy Caruth	cathy caruth born 1955 is frank h t rhodes ...

[59071 rows x 3 columns]

Note: Only the head of the SFrame is printed.

You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

Explore data

Taking a look at the entry for President Obama

In [6]:
```
obama = people[people['name'] == 'Barack Obama']
```

In [7]:
```
obama
```

Out[7]:

URI	name	text
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...

[? rows x 3 columns]

Note: Only the head of the SFrame is printed. This SFrame is lazily evaluated.

You can use sf.materialize() to force materialization.

In [8]:
```
obama['text']
```

Out[8]:
```
dtype: str
Rows: ?
['barack hussein obama ii brk husen bm born august 4 1961 is the 44th and cur
rent president of the united states and the first african american to hold th
e office born in honolulu hawaii obama is a graduate of columbia university a
nd harvard law school where 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 th
e 13th district in the illinois senate from 1997 to 2004 running unsuccessful
ly for the united states house of representatives in 2000in 2004 obama receiv
ed national attention during his campaign to represent illinois in the united
states senate with his victory in the march democratic party primary his keyn
ote 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 sufficie
nt delegates in the democratic party primaries to receive the presidential no
mination he then defeated republican nominee john mccain in the general elect
ion and was inaugurated as president on january 20 2009 nine months after his
election obama was named the 2009 nobel peace prize laureateduring his first
two years in office obama signed into law economic stimulus legislation in re
sponse to the great recession in the form of the american recovery and reinve
stment act of 2009 and the tax relief unemployment insurance reauthorization
and job creation act of 2010 other major domestic initiatives in his first te
rm included the patient protection and affordable care act often referred to
as obamacare the doddfrank wall street reform and consumer protection act and
the dont ask dont tell repeal act of 2010 in foreign policy obama ended us mi
litary involvement in the iraq war increased us troop levels in afghanistan s
igned the new start arms control treaty with russia ordered us military invol
vement 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 ho
use of representatives as the democratic party lost a total of 63 seats and a
fter a lengthy debate over federal spending and whether or not to raise the n
ations debt limit obama signed the budget control act of 2011 and the america
n taxpayer relief act of 2012obama was reelected president in november 2012 d
efeating republican nominee mitt romney and was sworn in for a second term on
january 20 2013 during his second term obama has promoted domestic policies r
elated to gun control in response to the sandy hook elementary school shootin
g and has called for full equality for lgbt americans while his administratio
n has filed briefs which urged the supreme court to strike down the defense o
f marriage act of 1996 and californias proposition 8 as unconstitutional in f
oreign policy obama ordered us military involvement in iraq in response to ga
ins made by the islamic state in iraq after the 2011 withdrawal from iraq con
tinued the process of ending us combat operations in afghanistan and has soug
ht to normalize us relations with cuba', ... ]
```

Explore the entry for actor George Clooney

In [9]:
```
clooney = people[people['name'] == 'George Clooney']
clooney['text']
```

Out[9]:
```
dtype: str
Rows: ?
['george timothy clooney born may 6 1961 is an american actor writer producer
director and activist he has received three golden globe awards for his work
as an actor and two academy awards one for acting and the other for producing
clooney made his acting debut on television in 1978 and later gained wide rec
ognition in his role as dr doug ross on the longrunning medical drama er from
1994 to 1999 for which he received two emmy award nominations while working o
n er he began attracting a variety of leading roles in films including the su
perhero film batman robin 1997 and the crime comedy out of sight 1998 in whic
h he first worked with a director who would become a longtime collaborator st
even soderbergh in 1999 clooney took the lead role in three kings a wellrecei
ved war satire set during the gulf warin 2001 clooneys fame widened with the
release of his biggest commercial success the heist comedy oceans eleven the
first of the film trilogy a remake of the 1960 film with frank sinatra as dan
ny ocean he made his directorial debut a year later with the biographical thr
iller confessions of a dangerous mind and has since directed the drama good n
ight and good luck 2005 the sports comedy leatherheads 2008 the political dra
ma the ides of march 2011 and the comedydrama war film the monuments men 2014
he won an academy award for best supporting actor for the middle east thrille
r syriana 2005 and subsequently earned best actor nominations for the legal t
hriller michael clayton 2007 the comedydrama up in the air 2009 and the drama
the descendants 2011 in 2013 he received the academy award for best picture f
or producing the political thriller argo alongside ben affleck and grant hesl
ov he is the only person ever to be nominated for academy awards in six categ
oriesclooney is sometimes described as one of the most handsome men in the wo
rld in 2005 tv guide ranked clooney no 1 on its 50 sexiest stars of all time
list in 2009 he was included in times annual time 100 as one of the most infl
uential people in the world clooney is also noted for his political activism
and has served as one of the united nations messengers of peace since january
31 2008 his humanitarian work includes his advocacy of finding a resolution f
or the darfur conflict raising funds for the 2010 haiti earthquake 2004 tsuna
mi and 911 victims and creating documentaries such as sand and sorrow to rais
e awareness about international crises he is also a member of the council on
foreign relations', ... ]
```

Word counts for Obama acticle

In [10]:
```
obama['word_count'] = turicreate.text_analytics.count_words(obama['text'])
```

In [11]:
```
obama
```

Out[11]:

URI	name	text	word_count
<http://dbpedia.org/resou rce/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	{'normalize': 1.0, 'sought': 1.0, 'combat': ...

[1 rows x 4 columns]

In [12]:
```
print (obama['word_count'])
```

[{'normalize': 1.0, 'sought': 1.0, 'combat': 1.0, 'continued': 1.0, 'unconsti
tutional': 1.0, '8': 1.0, 'californias': 1.0, '1996': 1.0, 'marriage': 1.0, '
defense': 1.0, 'down': 1.0, 'proposition': 1.0, 'court': 1.0, 'supreme': 1.0,
'urged': 1.0, 'which': 1.0, 'briefs': 1.0, 'administration': 1.0, 'while': 1.
0, 'americans': 1.0, 'called': 1.0, 'cuba': 1.0, 'gun': 1.0, 'related': 1.0,
'policies': 1.0, 'promoted': 1.0, '2013': 1.0, 'second': 2.0, 'romney': 1.0,
'filed': 1.0, '2012': 1.0, 'reelected': 1.0, 'taxpayer': 1.0, 'budget': 1.0,
'nations': 1.0, 'raise': 1.0, 'spending': 1.0, 'over': 1.0, 'lengthy': 1.0, '
gains': 1.0, 'seats': 1.0, '63': 1.0, 'total': 1.0, 'lost': 1.0, 'regained':
1.0, 'whether': 1.0, 'close': 1.0, 'patient': 1.0, 'by': 1.0, 'sandy': 1.0, '
after': 4.0, 'presidential': 2.0, 'november': 2.0, 'obama': 9.0, 'election':
3.0, 'august': 1.0, 'mccain': 1.0, 'primary': 2.0, 'he': 7.0, 'united': 3.0,
'with': 3.0, 'current': 1.0, 'campaign': 3.0, 'degree': 1.0, 'obamacare': 1.
0, 'convention': 1.0, 'house': 2.0, 'during': 2.0, 'victory': 1.0, 'worked':
1.0, 'troop': 1.0, 'senate': 3.0, 'islamic': 1.0, 'law': 6.0, 'district': 1.
0, '13th': 1.0, 'national': 2.0, 'sworn': 1.0, 'representing': 1.0, 'democrat
ic': 4.0, '20': 2.0, 'that': 1.0, 'process': 1.0, 'the': 40.0, '1961': 1.0, '
2004': 3.0, 'debate': 1.0, 'from': 3.0, 'three': 1.0, 'into': 1.0, 'at': 2.0,
'is': 2.0, 'rights': 1.0, 'withdrawal': 1.0, 'often': 1.0, 'attorney': 1.0, '
civil': 1.0, 'taught': 1.0, 'represent': 1.0, 'january': 3.0, 'laureateduring
': 1.0, 'running': 1.0, '1992': 1.0, 'organizer': 1.0, 'debt': 1.0, 'american
': 3.0, 'unsuccessfully': 1.0, 'president': 4.0, 'july': 1.0, 'and': 21.0, 'a
rms': 1.0, 'hussein': 1.0, '2010': 2.0, 'african': 1.0, 'operations': 1.0, 't
wo': 1.0, 'won': 1.0, 'has': 4.0, 'representatives': 2.0, 'where': 1.0, 'cons
titutional': 1.0, '44th': 1.0, 'his': 11.0, 'first': 3.0, 'death': 1.0, 'rece
ived': 1.0, '1997': 1.0, 'us': 6.0, '2012obama': 1.0, 'limit': 1.0, 'barack':
1.0, 'lgbt': 1.0, 'against': 1.0, 'job': 1.0, '2007': 1.0, 'receive': 1.0, 'e
lementary': 1.0, 'brk': 1.0, 'earning': 1.0, 'initiatives': 1.0, 'born': 2.0,
'shooting': 1.0, 'mitt': 1.0, 'consumer': 1.0, 'was': 5.0, 'named': 1.0, 'pri
ze': 1.0, 'office': 2.0, 'school': 3.0, 'bm': 1.0, 'honolulu': 1.0, 'hawaii':
1.0, 'as': 6.0, 'hold': 1.0, 'nominee': 2.0, 'a': 7.0, 'party': 3.0, 'reform
': 1.0, 'columbia': 1.0, 'years': 1.0, 'for': 4.0, 'john': 1.0, 'ending': 1.
0, 'served': 2.0, 'began': 1.0, 'husen': 1.0, '2011': 3.0, 'in': 30.0, 'illin
ois': 2.0, 'hook': 1.0, 'harvard': 2.0, 'months': 1.0, 'community': 1.0, 'ter
m': 3.0, 'nobel': 1.0, 'defeating': 1.0, '4': 1.0, 'chicago': 2.0, 'before':
1.0, 'foreign': 2.0, 'hillary': 1.0, 'unemployment': 1.0, 'to': 14.0, 'rodham
': 1.0, 'clinton': 1.0, 'libya': 1.0, '2008': 1.0, 'relations': 1.0, 'suffici
ent': 1.0, 'delegates': 1.0, 'primaries': 1.0, 'repeal': 1.0, 'new': 1.0, 'no
mination': 1.0, '2009': 3.0, 'form': 1.0, 'made': 1.0, 'peace': 1.0, 'defeate
d': 1.0, 'military': 4.0, 'republican': 2.0, 'doddfrank': 1.0, 'address': 1.
0, 'general': 1.0, 'inaugurated': 1.0, 'on': 2.0, 'nine': 1.0, 'signed': 3.0,
'ii': 1.0, 'economic': 1.0, 'included': 1.0, 'levels': 1.0, 'review': 1.0, 'l
egislation': 1.0, 'ask': 1.0, 'resulted': 1.0, 'graduate': 1.0, 'response':
3.0, 'great': 1.0, 'full': 1.0, 'recession': 1.0, '2000in': 1.0, 'recovery':
1.0, 'reinvestment': 1.0, 'act': 8.0, 'tax': 1.0, 'relief': 2.0, 'major': 1.
0, 'affordable': 1.0, 'reauthorization': 1.0, 'insurance': 1.0, 'creation':
1.0, 'state': 1.0, 'attention': 1.0, 'keynote': 1.0, 'other': 1.0, 'domestic
': 2.0, 'equality': 1.0, 'of': 18.0, 'protection': 2.0, 'march': 1.0, 'start
': 1.0, 'care': 1.0, 'afghanistan': 2.0, 'university': 2.0, 'laden': 1.0, 'ru
ssia': 1.0, 'wall': 1.0, 'referred': 1.0, 'street': 1.0, 'then': 1.0, 'dont':
2.0, 'tell': 1.0, 'policy': 2.0, 'ended': 1.0, 'involvement': 3.0, 'federal':
1.0, 'iraq': 4.0, 'terms': 1.0, 'war': 1.0, 'or': 1.0, 'treaty': 1.0, 'strike
': 1.0, 'not': 1.0, 'increased': 1.0, 'control': 4.0, 'ordered': 3.0, 'states
': 3.0, 'operation': 1.0, 'osama': 1.0, 'stimulus': 1.0, 'bin': 1.0, 'republi
cans': 1.0}]

Find most common words in Obama article

In [13]:
```
obama.stack('word_count',new_column_name=['word','count'])
```

Out[13]:

URI	name	text	word	count
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	normalize	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	sought	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	combat	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	continued	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	unconstitutional	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	8	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	californias	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	1996	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	marriage	1.0
<http://dbpedia.org/resource/Barack_Obama> ...	Barack Obama	barack hussein obama ii brk husen bm born august ...	defense	1.0

[273 rows x 5 columns]

Note: Only the head of the SFrame is printed.

You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

In [14]:
```
obama_word_count_table = obama[['word_count']].stack('word_count', new_column
```

In [15]:
```
obama_word_count_table
```

Out[15]:

word	count
normalize	1.0
sought	1.0
combat	1.0
continued	1.0
unconstitutional	1.0

8	1.0
californias	1.0
1996	1.0
marriage	1.0
defense	1.0

[273 rows x 2 columns]

Note: Only the head of the SFrame is printed.

You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

In [16]:
```
obama_word_count_table.sort('count',ascending=False)
```

Out[16]:

word	count
the	40.0
in	30.0
and	21.0
of	18.0
to	14.0
his	11.0
obama	9.0
act	8.0
a	7.0
he	7.0

[273 rows x 2 columns]

Note: Only the head of the SFrame is printed.

You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

Compute TF-IDF for the entire corpus of articles

In [17]:
```
people['word_count'] = turicreate.text_analytics.count_words(people['text'])
```

In [18]:
```
people
```

Out[18]:

URI	name	text	word_count
<http://dbpedia.org/resource/Digby_Morrell> ...	Digby Morrell	digby morrell born 10 october 1979 is a former ...	{'melbourne': 1.0, 'parade': 1.0, ...
<http://dbpedia.org/resource/Alfred_J._Lewy> ...	Alfred J. Lewy	alfred j lewy aka sandy lewy graduated from ...	{'time': 1.0, 'each': 1.0, 'hour': 1.0, ...

<http://dbpedia.org/resource/Harpdog_Brown> ...	Harpdog Brown	harpdog brown is a singer and harmonica player who ...	{'society': 1.0, 'hamilton': 1.0, 'to': ...
<http://dbpedia.org/resource/Franz_Rottensteiner> ...	Franz Rottensteiner	franz rottensteiner born in waidmannsfeld lower ...	{'kurdlawitzpreis': 1.0, 'awarded': 1.0, '2004': ...
<http://dbpedia.org/resource/G-Enka> ...	G-Enka	henry krvits born 30 december 1974 in tallinn ...	{'curtis': 1.0, '2007': 1.0, 'cent': 1.0, ...
<http://dbpedia.org/resource/Sam_Henderson> ...	Sam Henderson	sam henderson born october 18 1969 is an ...	{'asses': 1.0, 'sic': 1.0, 'toilets': 1.0, ...
<http://dbpedia.org/resource/Aaron_LaCrate> ...	Aaron LaCrate	aaron lacrate is an american music producer ...	{'streamz': 1.0, 'including': 1.0, ...
<http://dbpedia.org/resource/Trevor_Ferguson> ...	Trevor Ferguson	trevor ferguson aka john farrow born 11 november ...	{'concordia': 1.0, 'creative': 1.0, ...
<http://dbpedia.org/resource/Grant_Nelson> ...	Grant Nelson	grant nelson born 27 april 1971 in london ...	{'heavies': 1.0, 'new': 1.0, 'brand': 1.0, ...
<http://dbpedia.org/resource/Cathy_Caruth> ...	Cathy Caruth	cathy caruth born 1955 is frank h t rhodes ...	{'2002': 1.0, 'harvard': 1.0, 'twentieth': 1.0, ...

[59071 rows x 4 columns]

Note: Only the head of the SFrame is printed.

In [19]:
```
people['tfidf'] = turicreate.text_analytics.tf_idf(people['text'])
```

In [20]:
```
people
```

Out[20]:

URI	name	text	word_count
<http://dbpedia.org/resource/Digby_Morrell> ...	Digby Morrell	digby morrell born 10 october 1979 is a former ...	{'melbourne': 1.0, 'parade': 1.0, ...
<http://dbpedia.org/resource/Alfred_J._Lewy> ...	Alfred J. Lewy	alfred j lewy aka sandy lewy graduated from ...	{'time': 1.0, 'each': 1.0, 'hour': 1.0, ...
<http://dbpedia.org/resource/Harpdog_Brown> ...	Harpdog Brown	harpdog brown is a singer and harmonica player who ...	{'society': 1.0, 'hamilton': 1.0, 'to': ...
<http://dbpedia.org/resource/Franz_Rottensteiner> ...	Franz Rottensteiner	franz rottensteiner born in waidmannsfeld lower ...	{'kurdlawitzpreis': 1.0, 'awarded': 1.0, '2004': ...
<http://dbpedia.org/resource/G-Enka> ...	G-Enka	henry krvits born 30 december 1974 in tallinn ...	{'curtis': 1.0, '2007': 1.0, 'cent': 1.0, ...
<http://dbpedia.org/resource/Sam_Henderson> ...	Sam Henderson	sam henderson born october 18 1969 is an ...	{'asses': 1.0, 'sic': 1.0, 'toilets': 1.0, ...
<http://dbpedia.org/resource/Aaron_LaCrate> ...	Aaron LaCrate	aaron lacrate is an american music producer ...	{'streamz': 1.0, 'including': 1.0, ...

<http://dbpedia.org/resource/Trevor_Ferguson> ...	Trevor Ferguson	trevor ferguson aka john farrow born 11 november ...	{'concordia': 1.0, 'creative': 1.0, ...
<http://dbpedia.org/resource/Grant_Nelson> ...	Grant Nelson	grant nelson born 27 april 1971 in london ...	{'heavies': 1.0, 'new': 1.0, 'brand': 1.0, ...
<http://dbpedia.org/resource/Cathy_Caruth> ...	Cathy Caruth	cathy caruth born 1955 is frank h t rhodes ...	{'2002': 1.0, 'harvard': 1.0, 'twentieth': 1.0, ...

tfidf
{'melbourne': 3.8914310119380633, ...
{'time': 1.3253342074200498, ...
{'society': 2.4448047262085693, ...
{'kurdlawitzpreis': 10.986495389225194, ...
{'curtis': 5.299520032885375, ...
{'asses': 9.600201028105303, 's ...
{'streamz': 10.986495389225194, ...
{'concordia': 6.250296940830698, ...
{'heavies': 8.907053847545358, 'n ...
{'2002': 1.8753125887822302, ...

Examine the TF-IDF for the Obama article

In [21]:
```
obama = people[people['name'] == 'Barack Obama']
obama[['tfidf']].stack('tfidf',new_column_name=['word','tfidf']).sort('tfidf'
```

Out[21]:

word	tfidf
obama	43.2956530720749
act	27.67822262297991
iraq	17.747378587965535
control	14.887060845181308
law	14.722935761763422
ordered	14.533373950913514
military	13.115932778499415
involvement	12.784385241175055

response	12.784385241175055
democratic	12.410688697332166

[273 rows x 2 columns]

Note: Only the head of the SFrame is printed.

You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

Examine the TF-IDF for Clooney

In [22]:
```
clooney = people[people['name'] == 'George Clooney']
clooney[['tfidf']].stack('tfidf',new_column_name=['word','tfidf']).sort('tfid
```

Out[22]:

word	tfidf
clooney	30.47679823695488
thriller	19.64459743254604
drama	13.544372218899177
comedydrama	12.973371437789858
er	12.782751078181208
actor	11.832160900443771
categoriesclooney	10.986495389225194
heslov	10.986495389225194
producingclooney	10.986495389225194
comedy	10.481205264908446

[239 rows x 2 columns]

Note: Only the head of the SFrame is printed.

You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

Manually evaluate the distance between certain people's articles

In [23]:
```
clinton = people[people['name'] == 'Bill Clinton']
beckham = people[people['name'] == 'David Beckham']
```

Is Obama closer to Clinton or to Beckham?

In [24]:
```
turicreate.distances.cosine(obama['tfidf'][0],clinton['tfidf'][0])
```

Out[24]: 0.8339854936884277

In [25]:
```
turicreate.distances.cosine(obama['tfidf'][0],beckham['tfidf'][0])
```

Out[25]: 0.9791305844747478

Apply nearest neighbors for retrieval of Wikipedia articles

Build the NN model

In [26]:
```
knn_model = turicreate.nearest_neighbors.create(people,features=['tfidf'],lab
```

```
Starting brute force nearest neighbors model training.
Validating distance components.
Initializing model data.
Initializing distances.
Done.
```

Use model for retrieval... for example, who is closest to Obama?

In [27]:
```
knn_model.query(obama)
```

```
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 6.517ms      |
| Done         |         | 100         | 342.916ms    |
+--------------+---------+-------------+--------------+
```

Out[27]:

query_label	reference_label	distance	rank
0	Barack Obama	0.0	1
0	Joe Biden	0.7941176470588236	2
0	Joe Lieberman	0.7946859903381642	3
0	Kelly Ayotte	0.8119891008174387	4
0	Bill Clinton	0.8138528138528138	5

[5 rows x 4 columns]

Other examples of retrieval

In [28]:
```
swift = people[people['name'] == 'Taylor Swift']
```

In [29]:
```
knn_model.query(swift)
```

```
Starting pairwise querying.
```

```
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 5.504ms      |
| Done         |         | 100         | 363.37ms     |
+--------------+---------+-------------+--------------+
```

Out[29]:

	query_label	reference_label	distance	rank
0	0	Taylor Swift	0.0	1
1	0	Carrie Underwood	0.7623188405797101	2
2	0	Alicia Keys	0.7647058823529411	3
3	0	Jordin Sparks	0.7696335078534031	4
4	0	Leona Lewis	0.7761194029850746	5

[5 rows x 4 columns]

In [30]:
```
jolie = people[people['name'] == 'Angelina Jolie']
```

In [31]:
```
knn_model.query(jolie)
```

```
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 10.299ms     |
| Done         |         | 100         | 366.896ms    |
+--------------+---------+-------------+--------------+
```

Out[31]:

	query_label	reference_label	distance	rank
0	0	Angelina Jolie	0.0	1
1	0	Brad Pitt	0.7840236686390533	2
2	0	Julianne Moore	0.7958579881656804	3
3	0	Billy Bob Thornton	0.80306905370844	4
4	0	George Clooney	0.8046875	5

[5 rows x 4 columns]

In [32]:
```
arnold = people[people['name'] == 'Arnold Schwarzenegger']
```

In [33]:
```
knn_model.query(arnold)
```

```
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
```

```
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 11.642ms     |
| Done         |         | 100         | 346.312ms    |
+--------------+---------+-------------+--------------+
```

Out[33]:

query_label	reference_label	distance	rank
0	Arnold Schwarzenegger	0.0	1
0	Jesse Ventura	0.8189189189189189	2
0	John Kitzhaber	0.8246153846153846	3
0	Lincoln Chafee	0.8338762214983714	4
0	Anthony Foxx	0.8339100346020761	5

[5 rows x 4 columns]

In [34]:
```
elton_john = people[people['name'] == 'Elton John']
```

In [35]:
```
elton_john
```

Out[35]:

URI	name	text	word_count	tfidf
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 1.0, 'social': 1.0, ...	{'movements': 5.030658019760364, ...

[? rows x 5 columns]

Note: Only the head of the SFrame is printed. This SFrame is lazily evaluated.

You can use sf.materialize() to force materialization.

In [36]:
```
elton_john['text']
```

Out[36]:
```
dtype: str
Rows: ?
['sir elton hercules john cbe born reginald kenneth dwight 25 march 1947 is a
n english singer songwriter composer pianist record producer and occasional a
ctor he has worked with lyricist bernie taupin as his songwriter partner sinc
e 1967 they have collaborated on more than 30 albums to datein his fivedecade
career elton john has sold more than 300 million records making him one of th
e bestselling music artists in the world he has more than fifty top 40 hits i
ncluding seven consecutive no 1 us albums 58 billboard top 40 singles 27 top
10 four no 2 and nine no 1 for 31 consecutive years 19702000 he had at least
one song in the billboard hot 100 his single something about the way you look
tonightcandle in the wind 1997 sold over 33 million copies worldwide and is t
he bestselling single of all time he has received six grammy awards five brit
awards winning two awards for outstanding contribution to music and the first
brits icon in 2013 for his lasting impact on british culture an academy award
a golden globe award a tony award a disney legend award and the kennedy cente
r honors in 2004 in 2004 rolling stone ranked him number 49 on its list of 10
0 influential musicians of the rock and roll era in 2008 billboard ranked him
the most successful male solo artist on the billboard hot 100 top alltime art
ists third overallelton john was inducted into the rock and roll hall of fame
in 1994 is an inductee into the songwriters hall of fame and is a fellow of t
```

```
he british academy of songwriters composers and authors having been named a c
ommander of the order of the british empire in 1996 john received a knighthoo
d from elizabeth ii for services to music and charitable services in 1998 joh
n has performed at a number of royal events such as the funeral of princess d
iana at westminster abbey in 1997 the party at the palace in 2002 and the que
ens diamond jubilee concert outside buckingham palace in 2012he has been heav
ily involved in the fight against aids since the late 1980s in 1992 he establ
ished the elton john aids foundation and a year later began hosting the annua
l academy award party which has since become one of the highestprofile oscar
parties in the hollywood film industry since its inception the foundation has
raised over 200 million john who announced he was bisexual in 1976 and has be
en openly gay since 1988 entered into a civil partnership with david furnish
on 21 december 2005 and after gay marriage became legal in england wed furnis
h on 21 december 2014 he continues to be a champion for lgbt social movements
```

In [37]:
```
elton_john['word_count'] = turicreate.text_analytics.count_words(elton_john['
```

In [38]:
```
elton_john['word_count']
```

Out[38]:
```
dtype: dict
Rows: 1
[{'movements': 1.0, 'social': 1.0, 'champion': 1.0, 'wed': 1.0, 'legal': 1.0,
'became': 1.0, 'after': 1.0, '2005': 1.0, 'december': 2.0, 'furnish': 2.0, '1
988': 1.0, 'gay': 2.0, 'openly': 1.0, '1976': 1.0, 'bisexual': 1.0, '200': 1.
0, 'raised': 1.0, 'industry': 1.0, 'film': 1.0, 'hollywood': 1.0, 'parties':
1.0, 'become': 1.0, 'which': 1.0, 'hosting': 1.0, 'year': 1.0, 'established':
1.0, '1992': 1.0, '1980s': 1.0, 'lgbt': 1.0, 'against': 1.0, 'involved': 1.0,
'2014': 1.0, 'entered': 1.0, '2012he': 1.0, 'you': 1.0, 'nine': 1.0, 'academy
': 3.0, 'something': 1.0, 'artists': 2.0, 'samesex': 1.0, 'single': 2.0, 'sol
o': 1.0, 'overallelton': 1.0, 'no': 3.0, 'era': 1.0, 'performed': 1.0, 'for':
5.0, 'foundation': 2.0, 'john': 7.0, 'such': 1.0, '27': 1.0, 'and': 15.0, 'wo
rld': 1.0, 'billboard': 4.0, '1': 2.0, 'highestprofile': 1.0, 'actor': 1.0, '
aids': 2.0, 'since': 5.0, 'consecutive': 2.0, 'songwriter': 2.0, '300': 1.0,
'including': 1.0, '40': 2.0, 'palace': 2.0, 'tonightcandle': 1.0, 'in': 18.0,
'datein': 1.0, 'singer': 1.0, 'four': 1.0, '1994': 1.0, 'his': 4.0, 'of': 13.
0, 'the': 27.0, 'one': 3.0, 'continues': 1.0, '21': 2.0, 'rolling': 1.0, 'sev
en': 1.0, 'diana': 1.0, 'us': 1.0, '1997': 2.0, '10': 1.0, 'hits': 1.0, 'reco
rds': 1.0, 'million': 3.0, 'order': 1.0, 'david': 1.0, 'civil': 1.0, 'pianist
': 1.0, '100': 3.0, 'him': 3.0, 'he': 7.0, 'march': 1.0, 'look': 1.0, 'at':
4.0, 'received': 2.0, 'reginald': 1.0, 'song': 1.0, 'british': 3.0, 'golden':
1.0, 'an': 3.0, 'bernie': 1.0, 'way': 1.0, 'all': 1.0, 'ii': 1.0, 'hercules':
1.0, '1967': 1.0, 'collaborated': 1.0, '33': 1.0, 'partnership': 1.0, 'awards
': 3.0, 'fifty': 1.0, '1947': 1.0, 'cbe': 1.0, 'singles': 1.0, 'has': 9.0, 'h
ad': 1.0, 'producer': 1.0, 'music': 3.0, 'career': 1.0, 'marriage': 2.0, 'ken
neth': 1.0, 'with': 2.0, 'fellow': 1.0, 'elton': 3.0, 'fight': 1.0, 'worked':
1.0, 'record': 1.0, 'having': 1.0, 'be': 1.0, 'sold': 2.0, 'making': 1.0, 'mo
st': 1.0, 'buckingham': 1.0, 'sir': 1.0, 'least': 1.0, 'to': 4.0, 'as': 2.0,
'have': 1.0, 'hot': 2.0, 'inducted': 1.0, 'taupin': 1.0, 'they': 1.0, 'bestse
lling': 2.0, 'partner': 1.0, 'on': 6.0, 'more': 3.0, '2002': 1.0, 'brit': 1.
0, '30': 1.0, 'lasting': 1.0, 'composers': 1.0, 'award': 5.0, 'royal': 1.0, '
fivedecade': 1.0, 'artist': 1.0, 'westminster': 1.0, 'occasional': 1.0, 'copi
es': 1.0, 'worldwide': 2.0, 'charitable': 1.0, 'england': 1.0, 'began': 1.0,
'top': 4.0, 'time': 1.0, '58': 1.0, 'wind': 1.0, 'over': 2.0, 'six': 1.0, 'is
': 4.0, 'culture': 1.0, 'later': 1.0, 'its': 2.0, 'grammy': 1.0, 'five': 1.0,
'winning': 1.0, 'two': 1.0, 'english': 1.0, 'outstanding': 1.0, '2008': 1.0,
'influential': 1.0, 'contribution': 1.0, 'disney': 1.0, 'first': 1.0, 'funera
l': 1.0, 'brits': 1.0, 'albums': 2.0, 'icon': 1.0, 'than': 3.0, '2013': 1.0,
'inception': 1.0, 'a': 10.0, '1998': 1.0, 'globe': 1.0, '1996': 1.0, 'tony':
1.0, 'legend': 1.0, '2': 1.0, 'center': 1.0, 'who': 1.0, 'from': 1.0, 'annual
```

```
': 1.0, 'honors': 1.0, '25': 1.0, 'fame': 2.0, 'events': 1.0, '2004': 2.0, 's
tone': 1.0, 'ranked': 2.0, 'number': 2.0, 'years': 1.0, 'queens': 1.0, 'born
': 1.0, '49': 1.0, 'musicians': 1.0, 'kennedy': 1.0, 'rock': 2.0, 'been': 3.
0, 'princess': 1.0, 'roll': 2.0, 'successful': 1.0, 'male': 1.0, 'about': 1.
0, 'alltime': 1.0, 'list': 1.0, 'third': 1.0, 'was': 2.0, 'into': 3.0, 'impac
t': 1.0, 'hall': 2.0, 'inductee': 1.0, 'announced': 1.0, '19702000': 1.0, '31
': 1.0, 'named': 1.0, 'dwight': 1.0, 'songwriters': 2.0, 'heavily': 1.0, 'aut
hors': 1.0, 'commander': 1.0, 'oscar': 1.0, 'empire': 1.0, 'lyricist': 1.0, '
knighthood': 1.0, 'elizabeth': 1.0, 'services': 2.0, 'abbey': 1.0, 'late': 1.
0, 'party': 2.0, 'composer': 1.0, 'diamond': 1.0, 'jubilee': 1.0, 'concert':
1.0, 'outside': 1.0}]
```

In [39]:
```
elton_john_table = elton_john.stack('word_count',new_column_name=['word','cou
```

Top word count words for Elton John

In [40]:
```
elton_john_table.sort('count',ascending=False)
```

Out[40]:

URI	name	text	tfidf	word	count
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	the	27.0
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	in	18.0
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	and	15.0
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	of	13.0
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	a	10.0
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	has	9.0
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	he	7.0
<http://dbpedia.org/resource/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	john	7.0

<http://dbpedia.org /resou rce/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	on	6.0
<http://dbpedia.org /resou rce/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 5.030658019760364, ...	award	5.0

Top TF-IDF words for Elton John

In [41]:
```
elton_john = people[people['name'] == 'Elton John']
```

In [42]:
```
elton_john
```

Out[42]:

URI	name	text	word_count	tfidf
<http://dbpedia.org /resou rce/Elton_John> ...	Elton John	sir elton hercules john cbe born reginald ken ...	{'movements': 1.0, 'social': 1.0, ...	{'movements': 5.030658019760364, ...

[? rows x 5 columns]
Note: Only the head of the SFrame is printed. This SFrame is lazily evaluated.
You can use sf.materialize() to force materialization.

In [43]:
```
elton_john[['tfidf']].stack('tfidf',new_column_name=['word','tfidf']).sort('t
```

Out[43]:

word	tfidf
furnish	18.38947183999428
elton	17.482320270031995
billboard	17.30368095754203
john	13.93931279239831
songwriters	11.250406447031539
overallelton	10.986495389225194
tonightcandle	10.986495389225194
fivedecade	10.293348208665249
19702000	10.293348208665249
aids	10.262846934045534

[255 rows x 2 columns]
Note: Only the head of the SFrame is printed.
You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.

The cosine distance between 'Elton John's and 'Victoria

Beckham's articles (represented with TF-IDF) falls within which range?

In [48]:
```
victoria_beckham = people[people['name'] == 'Victoria Beckham']
elton_john = people[people['name'] == 'Elton John']
```

In [49]:
```
turicreate.distances.cosine(elton_john['tfidf'][0],victoria_beckham['tfidf'][
```

Out[49]:
```
0.9567006376655429
```

In [50]:
```
paul_mccartney = people[people['name'] == 'Paul McCartney']
```

In [51]:
```
turicreate.distances.cosine(elton_john['tfidf'][0],paul_mccartney['tfidf'][0]
```

Out[51]:
```
0.8250310029221779
```

Nearest neighbours model

In [54]:
```
knn_model_2 = turicreate.nearest_neighbors.create(people,features=['tfidf'],l
```

```
Starting brute force nearest neighbors model training.
Validating distance components.
Initializing model data.
Initializing distances.
Done.
```

In [55]:
```
knn_model_2.query(elton_john)
```

```
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 8.166ms      |
| Done         |         | 100         | 344.165ms    |
+--------------+---------+-------------+--------------+
```

Out[55]:

query_label	reference_label	distance	rank
0	Elton John	-2.220446049250313e-16	1
0	Rod Stewart	0.7172196678927374	2
0	George Michael	0.7476009989692848	3
0	Sting (musician)	0.7476719544306141	4
0	Phil Collins	0.7511932487904706	5

[5 rows x 4 columns]

In [57]:
```
knn_model_3 = turicreate.nearest_neighbors.create(people,features=['word_coun
```

```
Starting brute force nearest neighbors model training.
Validating distance components.
Initializing model data.
Initializing distances.
Done.
```

In [58]:
```
knn_model_3.query(elton_john)
```

```
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 4.636ms      |
| Done         |         | 100         | 328.247ms    |
+--------------+---------+-------------+--------------+
```

Out[58]:

query_label	reference_label	distance	rank
0	Elton John	2.220446049250313e-16	1
0	Cliff Richard	0.16142415258967036	2
0	Sandro Petrone	0.16822542751041114	3
0	Rod Stewart	0.16832716558706107	4
0	Malachi O'Doherty	0.177315545978884	5

[5 rows x 4 columns]

Who is the nearest cosine-distance neighbor to 'Victoria Beckham' using raw word counts?

In [59]:
```
knn_model_3.query(victoria_beckham)
```

```
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 10.682ms     |
| Done         |         | 100         | 302.302ms    |
+--------------+---------+-------------+--------------+
```

Out[59]:

query_label	reference_label	distance	rank
0	Victoria Beckham	-2.220446049250313e-16	1
0	Mary Fitzgerald (artist)	0.20730703611504997	2
0	Adrienne Corri	0.21450978278754795	3

0	Beverly Jane Fry	0.21746646874079278	4
0	Raman Mundair	0.21769547499150488	5

[5 rows x 4 columns]

Who is the nearest cosine-distance neighbor to 'Victoria Beckham' using TF-IDF?

In [60]:
```
knn_model_2.query(victoria_beckham)
```

```
Starting pairwise querying.
+--------------+---------+-------------+--------------+
| Query points | # Pairs | % Complete. | Elapsed Time |
+--------------+---------+-------------+--------------+
| 0            | 1       | 0.00169288  | 11.631ms     |
| Done         |         | 100         | 377.216ms    |
+--------------+---------+-------------+--------------+
```

Out[60]:

query_label	reference_label	distance	rank
0	Victoria Beckham	1.1102230246251565e-16	1
0	David Beckham	0.5481696102632145	2
0	Stephen Dow Beckham	0.7849867068283364	3
0	Mel B	0.8095855234085036	4
0	Caroline Rush	0.81982642291868	5

[5 rows x 4 columns]

In [ ]:
```
```