In [9]:
```
# Download data if you haven't already
import graphlab as gl
import os

if os.path.exists('LDA_3'):
    docs_3 = gl.SFrame('LDA_3')
    docs_3.dropna(how='all')
else:
    docs_3 = gl.SFrame.read_csv('C:/Users/Niranjan/Documents/Spring2016/
BigData/LDA/data_folder/File_No_3.csv', header=False)
    docs_3.save('LDA_3')
    docs_3.dropna(how='all')

# Remove stopwords and convert to bag of words
docs_3 = gl.text_analytics.count_words(docs_3['X1'])
docs_3 = docs_3.dict_trim_by_keys(gl.text_analytics.stopwords(), exclude
=True)

# Learn topic model
model_3= gl.topic_model.create(docs_3,
                                    num_topics=20,
                                    num_iterations=50)

#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=10)['word
s'][1]

#saving the model
model_3.save('my_model_3')
```

Unable to parse line "Finally, a championship parade on a weekend Will be screaming at the and then screaming at Katy Perry   Robyn concer t.,"

Unable to parse line "News Venezuela, Chavez Brace for Uncertain Futur e,"

Unable to parse line "RT Glad to see another day, live it up ,"

Unable to parse line "we are the future ,,"

Unable to parse line "Cloud 9 ,,"

Unable to parse line "Kross Wordz, Eurydice, Church, G, Paul D, Keith R odgers, Renaissance, Gate City Youth Slam Team, & URL,"

Unable to parse line "Low on cash? no worries, MoSoul in the Park this Saturday at Festival Park at 3 7 p.m. is,"

Unable to parse line "RT Bout draw Conn, Richmond, Mass, Gate City in that order Leggo ,"

Unable to parse line "The Festival is next week   Just reminder that the dopest poets are coming to Greensboro, N.C.   ,"

Unable to parse line "For The First Time, Developing Countries Spending The Most On Renewables,"

Unable to parse line "Keep the team in your prayers   They are doing a great job of representing Greensboro, N.C.,"

Unable to parse line "News At Camp No Worries, cancer patients are jus t kids again | Philadelphia ...,"

164975 lines failed to parse correctly

Finished parsing file C:\Users\Niranjan\Documents\Spring2016\BigData\LD A\data_folder\File_No_3.csv

Parsing completed. Parsed 100 lines in 11.82 secs.

Unable to parse line "News Venezuela, Chavez Brace for Uncertain Futur e,"

Unable to parse line "Finally, a championship parade on a weekend Will be screaming at the and then screaming at Katy Perry   Robyn concer t.,"

Unable to parse line "News At Camp No Worries, cancer patients are jus t kids again | Philadelphia ...,"

Unable to parse line "RT Glad to see another day, live it up ,"

Unable to parse line "For The First Time, Developing Countries Spending The Most On Renewables,"

Unable to parse line "News A s, Yankees teaming up to fight pediatric cancer | New York Daily ...,"

Unable to parse line "Put in your Oscar votes already, Streep will swee p again The Iron Lady Teaser Trailer,"

Unable to parse line "we are the future ,,"

Unable to parse line "News A s, Yankees teaming up to fight pediatric cancer,"

Unable to parse line "Cloud 9 ,,"

Unable to parse line "RT Dear Obama and Democrats great job betraying and alienating and the people who supported you, chasing after the peop le who ne ...,"

Unable to parse line "News Irish softball coach, daughter makes emotio nal donation to Memorial Hospital,"

Read 563410 lines. Lines per second: 38100.8

Read 1134196 lines. Lines per second: 37185.4

Read 1704342 lines. Lines per second: 34286.1

Read 2279862 lines. Lines per second: 31000.2

Read 3428220 lines. Lines per second: 39288

Read 4011122 lines. Lines per second: 37302.1

Read 4593137 lines. Lines per second: 40231.5

Read 5157615 lines. Lines per second: 42234.3

Read 5727159 lines. Lines per second: 40006.9

Read 6295150 lines. Lines per second: 42351.2

Read 6873175 lines. Lines per second: 42686.8

Read 7450722 lines. Lines per second: 44348.7

Read 8023722 lines. Lines per second: 42865.6

Read 8619684 lines. Lines per second: 41577

Read 9207169 lines. Lines per second: 40075.8

Read 9790074 lines. Lines per second: 40738.8

```
2799543 lines failed to parse correctly

Finished parsing file C:\Users\Niranjan\Documents\Spring2016\BigData\LD
A\data_folder\File_No_3.csv

Parsing completed. Parsed 10007330 lines in 240.6 secs.

Learning a topic model

       Number of documents  10007330

           Vocabulary size   1429632

   Running collapsed Gibbs sampling

+-----------+---------------+----------------+-----------------+

| Iteration | Elapsed Time  | Tokens/Second  | Est. Perplexity |

+-----------+---------------+----------------+-----------------+

| 10        | 5m 13s        | 3.17922e+006   | 0               |

| 20        | 9m 50s        | 3.16148e+006   | 0               |

| 30        | 13m 43s       | 3.301e+006     | 0               |

| 40        | 17m 31s       | 3.23236e+006   | 0               |

| 50        | 21m 39s       | 3.0705e+006    | 0               |

+-----------+---------------+----------------+-----------------+

------------------------------------------------------
Inferred types from first line of file as
column_type_hints=[str,str]
If parsing fails due to incorrect types, you can correct
the inferred type list above and pass it to read_csv in
the column_type_hints argument
------------------------------------------------------
['win', 'game', 'gift', 'tonight', 'chance', 'tickets', 'enter', 'read
y', 'save', 'card']
```

In [10]:
```
print model_3.get_topics()
```

```
+-------+---------+------------------+
| topic |   word  |      score       |
+-------+---------+------------------+
|   0   |   day   | 0.0423762136743  |
|   0   |   free  | 0.0342499623065  |
|   0   |  happy  |  0.033401095456  |
|   0   |   don   | 0.0125981097679  |
|   0   |  today  | 0.0122341800383  |
|   1   |   win   | 0.0230610551047  |
|   1   |   game  |  0.013851781633  |
|   1   |   gift  | 0.00983608910781 |
|   1   | tonight | 0.00976821486424 |
|   1   |  chance | 0.00915764700069 |
+-------+---------+------------------+
[100 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 an
d columns.
```

In [11]:
```
print model_3.get_topics(output_type='topic_words')
```

```
+--------------------------------+
|             words              |
+--------------------------------+
| [day, free, happy, don, today] |
| [win, game, gift, tonight,...  |
| [things, people, learn, li...  |
| [news, state, &, house, love]  |
| [rt, tonight, day, don, feel]  |
|    [&, 2, follow, ll, week]    |
|   [rt, don, ..., job, life]    |
| [facebook, twitter, posted...  |
| [lol, today, top, real, daily] |
|      [video, 2, 1, 3, rt]      |
+--------------------------------+
[20 rows x 1 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 an
d columns.
```

In [21]:
```
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][1]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][2]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][3]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][4]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][5]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][6]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][7]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][8]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][9]
#printing the topics
print model_3.get_topics(output_type='topic_words',num_words=100)['word
s'][10]
```

```
['win', 'game', 'gift', 'tonight', 'chance', 'tickets', 'enter', 'rea
dy', 'save', 'card', '2', 'giving', 'contest', 'sale', 'water', 'pla
y', 'hot', 'watching', '1', 'worth', 'drink', 'giveaway', 'won', 'cod
e', 'season', 'winner', 'hour', 'coupon', '$10', 'credit', 'movie',
'll', 'days', 'answer', 'star', 'trip', 'free', 'lady', '7', 'tix',
'winning', '$50', 'hours', '$100', 'pack', 'super', 'amazon', 'luck
y', 'copy', 'winners', 'cool', 'room', 'cards', '$20', 'discount', 'g
onna', 'yet?', 'ticket', '$25', 'price', 'receive', 'watch', 'heat',
'kid', 'pair', 'box', 'ball', 'baseball', 'ray', 'offer', 'prize', 'd
vd', 'bought', 'shop', 'tomorrow', 'party', 'signed', 'announced', 't
ill', 'mini', 'entered', 'purchase', 'month', 'coupons', 'store', 're
staurant', 'grab', 'spot', 'bag', 'hair', 'stop', 'miles', '20%', 'ga
ga', 'deals', 'entry', 'horse', 'order', 'trivia', 'forex']
['things', 'people', 'learn', 'life', 'great', 'health', 'home', 'sta
rt', '...', 'pay', 'channel', 'working', 'american', 'don', 'succes
s', 'dog', 'turn', 'child', 'training', 'book', 'secret', 'mom', 'rea
d', 'children', 'subscribed', 'deal', 'building', 'money', 'youtub
e.', '\xe2\x80\x93', 'fun', 'dream', 'share', 'school', 'pet', 'stude
nts', 'extra', 'professional', 'simple', 'community', 'design', 'insu
rance', 'air', 'successful', 'advice', 'mind', 'development', 'onlin
e', 'teeth', 'cash', 'states', 'industry', 'united', 'planning', 'stu
dy', 'awesome', 'attention', 'issues', 'research', 'control', 'person
al', 'age', 'thoughts', 'information', 'key', 'system', 'helps', 'poi
nt', 'selling', 'family', 'survey', 'college', 'favorite', 'shows',
'weekend', 'act', 'thought', 'yellow', 'growing', 'dogs', 'wait', 'ha
rd', 'force', 'sell', 'perfect', 'trick', 'computer', 'summer', 'teac
h', 'program', 'check', 'discovered', 'friends', 'experience', 'smar
t', 'wedding', 'body', 'books', 'baby', 'give']
['news', 'state', '&', 'house', 'love', 'million', 'u.s.', 'white',
'fire', '|', '2011', 'press', 'bill', '...', 'release', '~', 'year',
'public', 'county', 'market', 'budget', 'law', 'made', 'case', 'educa
tion', 'breaking', 'national', 'debt', 'photo', 'vote', 'gold', 'livi
ng', 'review', 'senate', 'rate', 'federal', 'conference', 'jim', 'sou
th', 'mortgage', 'major', 'ap', 'ny', 'world', 'california', 'privat
e', 'dollar', 'gop', 'times', 'prices', 'left', 'red', 'named', 'cong
rats', 'housing', 'war', 'yahoo', 'celebrity', 'report', 'union', 'ec
onomy', 'rates', 'loan', 'council', 'pre', 'reports', 'official', 'pr
oject', 'property', 'funding', 'fund', 'calls', 'ohio', 'make', 'cit
y', 'agency', 'chief', 'sold', 'university', 'crisis', 'french', 'jo
b', 'missed', 'the...', 'governor', 'department', 'reuters', 'passe
d', 'recovery', 'earrings', 'rick', 'fair', 'glad', 'water', 'foreclo
sure', 'congress', 'quarter', 'reform', 'coach', 'hospital']
['rt', 'tonight', 'day', 'don', 'feel', '.', 'hope', '>', 've', 'gonn
a', '?', 'heart', 'wow', '<', 'happy', 'day.', 'playing', 'enjoy', 'n
ow.', '...', 'person', 'lot', 'doesn', 'friends', 'eat', 'give', 'chu
rch', 'thinking', 'catch', 'week', 'kids', 'heard', 'you?', 'summer',
'kind', 'thing', 'father', 'rest', 'omg', 'rsvp', 'world', 'fall', 'm
e.', 'list', 'forward', 'won', 'place', 'debt', 'hold', 'retweet', 'c
oncert', 'days', 'care', 'nice', '7', 'story', 'support', 'series',
'young', 'john', 'meet', '2', 'presents', 'true', 'guest', 'haven',
'listening', 'enjoying', 'son', 'play', 'yeah', 'missing', 'vote', 't
hat.', 'truth', 'today', 'tv', 'busy', 'dad', 'text', 'perfect', 'mak
es', 'tonight.', 'living', 'fan', 'national', 'up.', 'ceiling', 'love
```

```
d', 'found', 'lost', 'loves', 'coming', 'brother', 'episode', '<<',
'lol.', 'awesome.', 'slow', 'on.']
['&', '2', 'follow', 'll', 'week', 'find', 'call', 'friday', 'love',
'visit', 'make', 'beach', 'art', 'share', 'details', 'page', 'lates
t', '2011', '.', 'long', 'dj', 'listen', 'short', 'plan', 'golf', 'ch
ristmas', '10', 'hit', 'tomorrow', 'text', 'deals', 'july', 'link',
'>', 'friends', 'meet', '1st', 'weather', 'back', 'hop', 'pics', 'jun
e', 'area', '18', 'add', 'free', 'hot', 'website', '9', 'hip', 'set',
'sale', 'family', 'artists', 'santa', 'travel', 'ladies', 'mix', 'hos
ted', 'april', 'atlanta', 'design', 'list', 'newsletter', 'city', '1
1', 'start', 'forward', 'closed', 'holiday', 'blog', 'paid', 'miami',
'museum', 'lots', 'mark', 'button', 'local', 'here.', 'beats', 'remi
x', 'products', 'arts', 'wednesday', 'site', 'rts', 'seattle', 'bac
k.', 'dinner', 'step', 'public', 'doesn', 'island', 'project', 'col
d', 'calendar', 'warning', 'ft.', 'bay', 'hotel']
['rt', 'don', '...', 'job', 'life', 'summer', 'girl', 'big', 'peopl
e', '.', 'it.', 'you.', 'time', 'coffee', 'miss', 'pretty', 've', 'ob
ama', 'bad', 'week', 'vegas', 'give', 'reading', 'needed', 'upcomin
g', 'coming', 'amazing', '..', 'tea', 'court', 'las', 'show', 'interv
iew', 'means', 'human', 'true', 'lol', 'fat', 'isn', 'read', 'preside
nt', 'nice', 'morning.', 'questions', 'dm', 'town', 'stop', 'sweet',
'now.', 'remember', 'end', 'blogs', 'skin', 'body', 'favorite', 'coo
l', 'trust', 'working', 'thought', 'spring', 'moving', 'success', 'ha
ha', 'late', 'jesus', 'realize', 'moment', 'left', 'break', 'scott',
'virtual', 'song', 'hear', 'yeah', 'spot', 'blood', 'enjoy', 'team',
'god', 'walk', 'thanksgiving', 'matter', 'dance', 'freedom', 'head',
'past', 'rock', 'heart', 'touch', 'faith', 'story', 'do.', 'loving',
'fight', 'assistant', 'shows', 'deep', 'lord', 'ca', 'answers']
['facebook', 'twitter', 'posted', 'photo', 'photos', 'album', '...',
'make', 'rt', 'money', 'center', 'part', 'end', 'film', 'ready', 'gre
at', 'news', 'added', 'years', '\xe2\x80\x93', 'fans', 'find', 'excit
ed', 've', 'week', 'st.', 'friend', 'things', 'super', 'lot', 'accoun
t', 'give', '2010', 'unlocked', 'international', 'users', 'badge',
'0', '|', '10', 'fan', 'checked', 'harry', 'community', 'group', 'fes
tival', 'security', 'wait', 'finally', 'vs.', 'set', 'airport', 'coll
ege', '17', '50', 'times', '24', 'ebay', 'auto', 'potter', 'it?', 'ch
icago', 'software', 'francisco', 'student', 'update', 'place', 'featu
re', '100', 'run', 'city', 'big', 'bid', 'jackson', 'louis', 'tweet',
'cool', 'record', 'taking', 'dc', 'medical', 'directory', 'society',
'fantastic', 'half', 'hard', 'star', 'tv', 'tour', '28', 'spam', 'com
puter', 'washington', 'walk', 'ads', 'deck', 'thomas', 'young', 'cont
act', 'visitors']
['lol', 'today', 'top', 'real', 'daily', 'back', 'stories', 'ya', '\x
e2\x96\xb8', 'll', 'damn', '...', 'da', 'news', 'big', 'san', 'haha',
'.', 'gotta', 'estate', 'bad', 'lil', '\xe2\x80\x9c', 'put', 'guess',
'police', 'kno', '5', 'yea', ';', 'boy', 'city', 'diego', 'york', 'ma
d', 'told', 'shut', 'homie', 'pic', 'ill', 'idk', 'missed', 'though
t', 'girls', 'up.', 'stop', 'crazy', 'love', 'it.', 'imma', 'man', 'b
day', 'head', 'ima', '2', 'dont', 'shout', 'called', 'ma', 'local',
'reason', 'nation', 'listening', 'cute', '....', 'ain', 'sex', 'forgo
t', 'job', 'sum', 'hahaha', 'dm', 'nigga', '\xe2\x99\xab', 'shot',
'<', 'reasons', 'def', 'hold', 'eyes', 'yall', '10', 'doin', 'home',
'tryna', 'yeah', 'review', 'sit', 'late', 'bruh', 'music', 'face', 'c
```

```
omin', 'smoke', 'dat', 'luv', 'tho', 'wassup', 'nap', 'team']
['video', '2', '1', '3', 'rt', 'live', 'youtube', 'love', '&', 'tim
e', '4', '...', 'black', 'world', 'women', 'll', 'show', 'wait', 'upl
oaded', 'men', 'watch', 'energy', 'dress', 'red', 'buy', 'west', 'siz
e', 'find', 'cup', 'hair', ';', 'ladies', 'don', 'week', 'tomorrow',
'chat', 'line', 'side', 'minutes', 'send', '2nd', 'bad', 'tonight',
'man', 'green', 'classic', 'favorited', 'rock', 'months', 'now.', 'ni
ce', 'place', '9', 'east', 'double', 'wins', 'tune', 'videos', 'perfe
ct', 'class', 'cut', 'colors', 'sizes', 'south', 'stop', 'wear', 'coa
st', 'park', 'round', '6', 'bag', 'pink', '30%', 'tweet', 'no.', 'pic
k', 'room', 'style', 'save', '5', 'fall', 'series', 'end', 'due', 'em
ail', '[pic]', 'north', 'soccer', 'taylor', 'studio', 'sexy', 'nigh
t', 'sarah', 'past', 'short', 'big', 'flip', 'print', 'review', 'min
e']
['rt', 'good', '...', 'great', 'morning', 'time', 'hope', 'lol', 'he
y', 'god', 'sounds', 'guys', 'nice', 'work', 'followers', 'night', 'w
eekend', 'thx', 'http', 'ready', 'fun', 'luck', 'friends', 'love', 'u
r', 'early', 'cont', 'watching', 'yeah', 'congrats', 'you.', 'miss',
'enjoy', 'man', 'ppl', 'shout', 'back', 'makes', 'idea', 'bro', 'pu
t', 'feeling', 'fam', 'thing', 'feels', '<', 'funny', 'tweets', 'lon
g', 'tweeps', 'post', 'amy', 'heard', 'favorite', 'road', 'doesn', 'f
olks', 'haven', 'night.', 'found', 'ty', 'sharing', 'spent', 'movie',
'breakfast', 'blog', 'gm', 'busy', 'tonight.', 'blessed', 'meeting',
'pic', 'extended', 'words', 'world', 'htt', 'winehouse', 'too.', 'par
k', 'cheers', 'whats', 'wrong', 'smart', 'good.', 'didnt', 'bed', 'we
ather', 'club', 'peeps', 'hate', 'ideas', 'dead', 'one.', 'question',
'today?', 'out.', 'met', 'sooo', 'days', 'woke']
```

In [17]:
```
print model_3['topics']
```

```
+-------------------------------+------------+
|      topic_probabilities      | vocabulary |
+-------------------------------+------------+
| [1.24032576159e-06, 3.0032... |  dolphins  |
| [3.02518478437e-08, 0.0002... |    swim    |
| [5.77810293814e-06, 3.0032... |    ets     |
| [3.02518478437e-08, 3.0032... |    ...     |
| [3.02518478437e-08, 0.0001... |   waves    |
| [8.80328772251e-06, 3.0032... |    news    |
| [3.02518478437e-08, 9.3101... |    west    |
| [3.02518478437e-08, 3.0032... |  roxbury   |
| [3.02518478437e-08, 6.9375... |  swimmers  |
| [0.00707109716683, 0.00048... |    make    |
+-------------------------------+------------+
[1429632 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 an
d columns.
```

In [19]:
```
print model_3.get_topics()
```

```
+-------+---------+------------------+
| topic |   word  |      score       |
+-------+---------+------------------+
|   0   |   day   | 0.0423762136743  |
|   0   |   free  | 0.0342499623065  |
|   0   |  happy  |  0.033401095456  |
|   0   |   don   | 0.0125981097679  |
|   0   |  today  | 0.0122341800383  |
|   1   |   win   | 0.0230610551047  |
|   1   |   game  |  0.013851781633  |
|   1   |   gift  | 0.00983608910781 |
|   1   | tonight | 0.00976821486424 |
|   1   |  chance | 0.00915764700069 |
+-------+---------+------------------+
[100 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 an
d columns.
```