Movie Reviews Dataset

1) Experiment with 20 topics, 500 iterations, alpha = 0.02, beta= 0.1

Note: for all the following experiments, if not specified, the same alpha and beta values were used. Only the parameters that are explicitly specified were changed, the rest are used in their default setting. Refer the documentation to know more about the default parameter settings.

Runtime of the experiment: 4 hours 2 mins. Approx. 29 secs/iteration (using numpy)

```
#---------------------------------------#
```

Topic *0* ~ Related to Austin Powers/Spy movies

```
[('spawn', 75), ('patch', 71), ('evil', 68), ('austin', 57), ('myers', 57), ('powers',
51), ('54', 50), ('carter', 46), ('dr', 44), ('washington', 39), ('prison', 37),
('melvin', 36), ('mike', 35), ('whale', 35), ('claire', 32), ('adams', 31), ('altman',
29), ('hurricane', 29), ('shane', 29), ('irene', 28), ('jakob', 28), ('jr', 26),
('hank', 26), ('charlie', 26), ('farrellys', 26)]

#---------------------------------------#
```

Topic *1* ~ these words somewhat feel like the ones used for a good movie review.. Not very informative to make a topic allocation

```
[('film', 5230), ('one', 2787), ('movie', 1537), ('like', 1519), ('story', 1393),
('character', 1220), ('characters', 1180), ('life', 1102), ('also', 1094), ('two',
1075), ('even', 1074), ('time', 1052), ('would', 1038), ('much', 1014), ('good', 969),
('way', 939), ('man', 868), ('best', 821), ('get', 795), ('first', 782), ('see', 771),
('make', 764), ('scene', 756), ('could', 752), ('many', 751)]

#---------------------------------------#
```

Topic *2* ~ Related to television series..

```
[('lynch', 68), ('nbsp', 58), ('carter', 53), ('seagal', 53), ('x-files', 52),
('series', 46), ('-4', 39), ('+4', 38), ('scale', 37), ('0', 36), ('mulder', 36),
('television', 34), ('existenz', 33), ('frank', 32), ('japanese', 28), ('duchovny',
28), ('frequency', 26), ('marty', 25), ('thirteenth', 25), ('betty', 24), ('bridge',
24), ('raimi', 23), ('amazing', 23), ('anderson', 23), ('greg', 23)]

#---------------------------------------#
```

Topic *3* ~ *probably related to Actors and characters.. Ben Affleck, Ben Stiller, Kevin Smith..*

```
[('jay', 84), ('bob', 65), ('smith', 60), ('toy', 59), ('kevin', 51), ('park', 50),
('silent', 47), ('woody', 42), ('vegas', 40), ('cauldron', 38), ('lucy', 36),
('chicken', 35), ('drug', 34), ('homer', 33), ('buzz', 32), ('rocky', 32), ('affleck',
30), ('ben', 29), ('garofalo', 28), ('toys', 28), ('phil', 28), ('las', 26),
('stiller', 26), ('taran', 26), ('hammer', 25)]

#---------------------------------------#
```

Topic *4* ~ not very sure but these words seem to be related to Disaster-related movies

```
[('girls', 100), ('spice', 65), ('harry', 64), ('impact', 58), ('deep', 56), ('helen',
46), ('armageddon', 45), ('freeman', 41), ('town', 37), ('comet', 37), ('tyler', 33),
('egoyan', 29), ('bus', 27), ('sweet', 27), ('paltrow', 27), ('monty', 27), ('watson',
27), ('martha', 26), ('emily', 25), ('leder', 25), ('hereafter', 25), ('songs', 24),
('peak', 24), ('morgan', 23), ('disaster', 22)]

#---------------------------------------#
```

Topic *5* ~ Teen movies related

```
[('school', 136), ('high', 90), ('sex', 85), ('american', 58), ('derek', 56),
('musical', 47), ('pie', 45), ('van', 43), ('hunting', 41), ('jim', 36), ('student',
35), ('damon', 35), ('hedwig', 34), ('college', 32), ('rock', 31), ('paul', 31),
('sweetback', 30), ('sean', 29), ('blonde', 29), ('affleck', 29), ('lambeau', 29),
('comedy', 28), ('matt', 28), ('parker', 28), ('eszterhas', 26)]

#--------------------------------------#
```

Topic *6* ~ related to Sci-fi thrillers

```
[('godzilla', 106), ('political', 87), ('bond', 71), ('bulworth', 60), ('lebowski',
54), ('dude', 52), ('president', 48), ('dog', 46), ('nuclear', 43), ('bobby', 39),
('broderick', 37), ('beatty', 37), ('american', 36), ('granger', 36), ('amistad', 34),
('spielberg', 31), ('york', 30), ('l', 29), ('campaign', 29), ('ship', 28), ('war',
28), ('ghost', 27), ('sphere', 27), ('goodman', 27), ('king', 26)]

#--------------------------------------#
```

Topic *7* ~ Related to vampires, witchcraft, horror movies

```
[('vampire', 96), ('blade', 72), ('vampires', 69), ('cage', 67), ('blair', 63),
('kevin', 56), ('devil', 56), ('cop', 52), ('witch', 49), ('baldwin', 43),
('carpenter', 42), ('8mm', 42), ('cruise', 42), ('woods', 42), ('arnold', 41),
('sixth', 40), ('snipes', 40), ('tom', 39), ('cole', 39), ('welles', 37), ('crow',
34), ('girl', 33), ('horror', 32), ('schumacher', 32), ('malkovich', 31)]

#--------------------------------------#
```

Topic *8* ~ Drama and romance related

```
[('love', 122), ('life', 101), ('shakespeare', 77), ('father', 69), ('family', 64),
('beautiful', 64), ('god', 60), ('angels', 53), ('son', 51), ('young', 45), ('boy',
45), ('fiennes', 44), ('elizabeth', 43), ('angel', 43), ('mother', 42), ('english',
42), ('finn', 41), ('romantic', 40), ('beauty', 35), ('anne', 35), ('francis', 34),
('fairy', 33), ('king', 32), ('husband', 32), ('anna', 32)]

#--------------------------------------#
```

Topic *9* ~ related to Star-wars

```
[('star', 217), ('wars', 138), ('trek', 118), ('lucas', 93), ('phantom', 80),
('menace', 75), ('jedi', 75), ('effects', 73), ('special', 71), ('jar', 48),
('neeson', 46), ('luke', 44), ('anakin', 38), ('liam', 38), ('trilogy', 37), ('darth',
36), ('obi-wan', 36), ('planet', 35), ('mcgregor', 34), ('insurrection', 34),
('george', 33), ('haunting', 33), ('skywalker', 32), ('house', 31), ('qui-gon', 31)]

#--------------------------------------#
```

Topic *10* ~ related to Romantic Movies and their characters

```
[('joe', 183), ('ship', 129), ('titanic', 96), ('harry', 77), ('simon', 75), ('shrek',
63), ('cameron', 55), ('rising', 51), ('allen', 50), ('jack', 49), ('rose', 47),
('anaconda', 40), ('snake', 38), ('kate', 34), ('dicaprio', 34), ('mike', 32),
('murphy', 32), ('poker', 32), ('football', 29), ('boat', 29), ('woody', 28), ('pitt',
28), ('ricky', 28), ('hopkins', 27), ('palmetto', 27)]

#--------------------------------------#
```

Topic *11* ~ Movie Review : such common words could be seen in a movie's review

```
[('film', 4209), ('movie', 4129), ('one', 2787), ('like', 2024), ('even', 1480),
('good', 1345), ('time', 1226), ('would', 1225), ('bad', 1182), ('get', 1128),
('much', 1008), ('really', 1000), ('first', 980), ('see', 957), ('plot', 947),
('well', 914), ('also', 868), ('could', 854), ('movies', 822), ('make', 821),
('films', 778), ('know', 777), ('character', 770), ('little', 752), ('two', 743)]
```

#--------------------------------------#

Topic *12* ~ Sci-fi fantasy Movies about extraterrestrial/Aliens: all the frequently occurring words do seem to be coherent with this topic.

```
[('alien', 340), ('aliens', 199), ('science', 158), ('planet', 155), ('space', 148),
('mars', 134), ('effects', 128), ('earth', 128), ('sci-fi', 99), ('special', 98),
('fiction', 97), ('computer', 92), ('ship', 88), ('humans', 81), ('crew', 78),
('troopers', 77), ('species', 74), ('human', 72), ('starship', 71), ('mission', 69),
('ripley', 65), ('matrix', 57), ('future', 55), ('verhoeven', 54), ('contact', 52)]
```

#--------------------------------------#

Topic *13 ~ Tarantino movies: all the topics seem to be coherent since Quentin Tarantino is famous for making unusual criminal movies. Some of his movie titles are even present in this topic. Eg Pulp Fiction, Boogie nights.*

```
[('tarantino', 113), ('jackie', 108), ('fiction', 74), ('pulp', 69), ('jack', 64),
('brown', 61), ('jackson', 53), ('ordell', 48), ('ray', 46), ('brooks', 45),
('robert', 43), ('boogie', 42), ('quentin', 40), ('clooney', 39), ('max', 38),
('sonny', 38), ('karen', 37), ('willis', 37), ('mr', 36), ('de', 36), ('crime', 34),
('nights', 34), ('niro', 33), ('scorsese', 32), ('pam', 29)]
```

#--------------------------------------#

Topic *14 ~* animated movies for children/family: all words seem to be thematically coherent

```
[('disney', 255), ('comedy', 142), ('animated', 132), ('funny', 131), ('family', 109),
('animation', 108), ('voice', 98), ('wedding', 97), ('mulan', 95), ('kids', 87),
('children', 86), ('king', 84), ('sandler', 77), ('julia', 69), ('crystal', 61),
('tarzan', 60), ('roberts', 53), ('eddie', 53), ('wrestling', 52), ('adults', 50),
('hanks', 50), ('humor', 49), ('little', 48), ('cartoon', 47), ('father', 46)]
```

#--------------------------------------#

Topic *15 ~* horror movies. Words in the topic are again thematically coherent since it talks about horror movies such as the Scream, Julie and plot/characters in them.

```
[('scream', 221), ('horror', 162), ('killer', 122), ('julie', 81), ('2', 70),
('slasher', 66), ('urban', 61), ('williamson', 60), ('teen', 59), ('tarzan', 57),
('last', 50), ('prinze', 49), ('summer', 46), ('legend', 46), ('sidney', 46),
('sequel', 43), ('college', 43), ('genre', 41), ('3', 41), ('craven', 40), ('freddie',
39), ('jr', 38), ('alex', 37), ('scary', 33), ('school', 33)]
```

#--------------------------------------#

Topic *16 ~ Filmmakers, their movies and actors: words seem to be somewhat coherent*

```
[('wild', 108), ('west', 76), ('annie', 56), ('kevin', 49), ('kelly', 46), ('grace',
45), ('sam', 41), ('hollow', 40), ('bacon', 39), ('zero', 38), ('smith', 36),
('argento', 36), ('campbell', 35), ('bats', 32), ('horse', 31), ('depp', 31), ('wood',
28), ('chris', 26), ('dillon', 26), ('burton', 26), ('ricci', 26), ('donnie', 26),
('eastwood', 26), ('richards', 25), ('suzie', 25)]
```

#--------------------------------------#

Topic *17 ~ Comedy : Krippendorf's Tribe, bowfinger.. Not all words are movie names, but they seem related to the plot/characters/actors of such movies*

```
[('flynt', 80), ('murphy', 69), ('apes', 66), ('babe', 48), ('eddie', 45), ('burton',
39), ('larry', 38), ('krippendorf', 38), ('jolie', 36), ('tribe', 35), ('general',
34), ('pig', 34), ('bowfinger', 33), ('martin', 31), ('daughter', 29), ('ape', 29),
('brenner', 29), ('leo', 28), ('wahlberg', 26), ('heston', 25), ('lawrence', 24),
('tim', 24), ('roth', 24), ('travolta', 24), ('casablanca', 22)]
```

```
#---------------------------------------#
```
Topic *18 ~ War movies and their characters: Private Ryan*
```
[('war', 193), ('battle', 95), ('men', 86), ('army', 80), ('ryan', 78), ('soldiers',
73), ('spielberg', 65), ('private', 64), ('jackal', 47), ('action', 46), ('jones',
45), ('hero', 44), ('joan', 42), ('carry', 42), ('military', 41), ('costner', 41),
('epic', 39), ('judd', 39), ('u', 38), ('fugitive', 37), ('law', 36), ('ashley', 36),
('gere', 35), ('williams', 32), ('bruce', 32)]
```

```
#---------------------------------------#
```
Topic *19 ~ Action Movies : all the words follow the theme*
```
[('jackie', 160), ('action', 158), ('truman', 145), ('chan', 144), ('batman', 125),
('carrey', 96), ('hong', 91), ('kong', 89), ('damme', 81), ('van', 72), ('martial',
72), ('mr', 69), ('chinese', 69), ('fight', 67), ('arts', 61), ('robin', 59),
('gibson', 53), ('master', 48), ('li', 45), ('fu', 42), ('lee', 40), ('kung', 40),
('jim', 38), ('cop', 36), ('partner', 36)]
```

2) Experiment with 1000 samples from the dataset 10 topics 250 iterations and 40 top words

Almost none of the topics follow any particular theme, unlike experiment 1.
Runtime of the experiment: 1 hour, 19.07s/iteration

```
#---------------------------------------#
Topic 0 - ...
[('joe', 110), ('godzilla', 101), ('harry', 92), ('york', 34), ('paul', 32),
('comedy', 31), ('matthew', 31), ('city', 30), ('noir', 29), ('nick', 28),
('broderick', 28), ('kate', 28), ('blonde', 25), ('young', 24), ('bob', 24),
('porter', 24), ('garofalo', 23), ('lucy', 23), ('women', 22), ('gorilla', 22),
('elizabeth', 22), ('jay', 21), ('daughter', 20), ('husband', 20), ('french', 20),
('palmetto', 20), ('greg', 20), ('friends', 19), ('lives', 19), ('heckerling', 19),
('perry', 19), ('mother', 18), ('max', 18), ('loser', 18), ('fish', 18), ('mona', 18),
('bye', 18), ('married', 17), ('pitt', 17), ('stiller', 17)]

#---------------------------------------#
Topic 1 - seems related to television series
[('murphy', 87), ('eddie', 83), ('nbsp', 58), ('comedy', 55), ('patch', 50),
('television', 46), ('sandler', 42), ('julia', 38), ('williams', 37), ('vegas', 36),
('drug', 36), ('flubber', 36), ('series', 35), ('wedding', 35), ('squad', 33),
('betty', 27), ('seagal', 27), ('douglas', 26), ('metro', 26), ('x-files', 26), ('g',
25), ('professor', 24), ('flying', 23), ('adam', 22), ('chris', 22), ('1900', 22),
('roberts', 22), ('beverly', 22), ('behind', 21), ('adams', 21), ('emily', 21),
('funny', 20), ('vacation', 20), ('scott', 19), ('steve', 19), ('attempts', 19),
('las', 19), ('crystal', 19), ('ricky', 19), ('duchovny', 19)]

#---------------------------------------#
Topic 2 - the words are not very coherent and it seems like they are from a mix of
horror and teenage movies genres
[('school', 108), ('horror', 99), ('scream', 93), ('high', 81), ('killer', 73),
('julie', 71), ('summer', 65), ('teen', 59), ('one', 56), ('know', 52), ('sequel',
52), ('slasher', 49), ('scary', 49), ('urban', 48), ('witch', 47), ('movie', 46),
('2', 44), ('sex', 44), ('prinze', 43), ('party', 39), ('blair', 39), ('kids', 37),
('3', 37), ('last', 35), ('girls', 34), ('first', 33), ('teenagers', 33), ('jennifer',
33), ('williamson', 33), ('kevin', 31), ('alicia', 30), ('charlie', 30),
```

('jawbreaker', 30), ('would', 29), ('original', 29), ('baseball', 29), ('genre', 28), ('house', 28), ('jr', 28), ('league', 28)]

#--------------------------------------#

Topic 3 - words that are usually used in a Movie review

[('film', 4250), ('movie', 3128), ('one', 2571), ('like', 1827), ('even', 1369), ('would', 1156), ('good', 1122), ('time', 1110), ('get', 1029), ('bad', 1005), ('much', 993), ('character', 910), ('could', 896), ('plot', 878), ('story', 878), ('characters', 864), ('two', 846), ('make', 790), ('really', 781), ('first', 771), ('also', 765), ('see', 759), ('way', 739), ('little', 720), ('well', 698), ('scene', 649), ('people', 648), ('never', 629), ('films', 626), ('scenes', 596), ('director', 592), ('know', 583), ('man', 570), ('another', 545), ('movies', 536), ('new', 535), ('big', 533), ('go', 531), ('made', 527), ('better', 527)]

#--------------------------------------#

Topic 4 - the words do not follow a particular theme.. They are from fantasy, crime, action, war.. Not coherent at all.

[('action', 105), ('jackie', 47), ('seagal', 44), ('john', 43), ('fight', 43), ('vampires', 42), ('vampire', 41), ('jolie', 40), ('west', 39), ('team', 38), ('general', 37), ('air', 36), ('chan', 36), ('james', 35), ('carpenter', 34), ('snipes', 34), ('woods', 34), ('crow', 33), ('son', 33), ('scenes', 31), ('kong', 31), ('brenner', 31), ('hong', 30), ('warrior', 30), ('lee', 29), ('daughter', 29), ('li', 29), ('fugitive', 28), ('crime', 28), ('travolta', 28), ('football', 27), ('ryan', 27), ('lawrence', 26), ('chinese', 26), ('jet', 25), ('war', 25), ('kudrow', 25), ('style', 24), ('libby', 23), ('editing', 21)]

#--------------------------------------#

Topic 5 - ...

[('girls', 53), ('tarzan', 50), ('willis', 48), ('spice', 47), ('lynch', 44), ('story', 35), ('richard', 34), ('jackal', 34), ('man', 32), ('krippendorf', 32), ('version', 28), ('simon', 28), ('tribe', 28), ('king', 27), ('jungle', 27), ('jakob', 27), ('message', 25), ('city', 25), ('novel', 24), ('lost', 23), ('jane', 23), ('robin', 23), ('hammer', 23), ('george', 21), ('francis', 21), ('mercury', 21), ('melvin', 20), ('williams', 19), ('subject', 19), ('u', 19), ('rising', 19), ('cisco', 18), ('claire', 17), ('grant', 17), ('government', 17), ('africa', 17), ('gere', 17), ('dreams', 16), ('political', 16), ('war', 16)]

#--------------------------------------#

Topic 6 - ...

[('van', 56), ('damme', 56), ('species', 47), ('fight', 41), ('stallone', 39), ('blade', 38), ('patrick', 36), ('charlie', 34), ('henstridge', 32), ('action', 28), ('ii', 28), ('knock', 26), ('gibson', 26), ('jerry', 25), ('trek', 25), ('sequel', 25), ('hollywood', 24), ('lambert', 23), ('todd', 23), ('data', 23), ('brother', 22), ('franklin', 22), ('soldier', 22), ('miller', 21), ('martha', 21), ('jean-claude', 20), ('carter', 20), ('bone', 20), ('sarah', 20), ('brothers', 19), ('winner', 19), ('connor', 19), ('schneider', 18), ('russell', 18), ('helen', 18), ('blues', 18), ('burn', 18), ('natasha', 17), ('sylvester', 17), ('mandingo', 17)]

#--------------------------------------#

Topic 7 - mix of extraterrestrial, disaster, sci-fi

[('alien', 120), ('space', 120), ('ship', 112), ('planet', 104), ('mars', 104), ('earth', 96), ('mission', 88), ('crew', 73), ('effects', 73), ('aliens', 64), ('deep', 64), ('virus', 54), ('team', 54), ('monster', 54), ('apes', 52), ('science', 50), ('humans', 49), ('snake', 48), ('human', 43), ('disaster', 42), ('cruise', 41), ('computer', 40), ('special', 39), ('harry', 38), ('giant', 38), ('sci-fi', 36),

```
('creature', 36), ('boat', 36), ('water', 35), ('anaconda', 35), ('rising', 33),
('park', 32), ('sinise', 31), ('impact', 31), ('dr', 30), ('fiction', 29), ('beast',
29), ('red', 28), ('affleck', 28), ('armageddon', 28)]

#--------------------------------------#
Topic 8 - ...
[('king', 51), ('wrestling', 48), ('54', 35), ('joan', 33), ('jimmy', 30),
('haunting', 28), ('wcw', 28), ('rock', 26), ('`the', 26), ('studio', 25), ('besson',
25), ('frank', 24), ('shane', 23), ('phillippe', 23), ('house', 22), ("o'donnell",
22), ('arquette', 22), ('bont', 22), ('horror', 21), ('jimmie', 21), ('bachelor', 20),
('de', 19), ('vince', 19), ('taylor', 19), ('murray', 19), ('myers', 19), ('neeson',
18), ('fans', 17), ('scary', 17), ('spirit', 17), ('seth', 17), ('married', 16),
('chucky', 16), ('skin', 15), ('animated', 15), ('club', 15), ('bride', 15), ('meyer',
15), ('liam', 15), ('rubell', 15)]

#--------------------------------------#
Topic 9 - ...
[('batman', 145), ('robin', 91), ('wild', 88), ('arnold', 80), ('mr', 75), ('smith',
70), ('spawn', 69), ('west', 65), ('disney', 55), ('schumacher', 54),
('schwarzenegger', 50), ('kevin', 44), ('carrey', 44), ('devil', 42), ('8mm', 39),
('joel', 37), ('verhoeven', 34), ('welles', 33), ('troopers', 33), ('jim', 31),
('freeze', 31), ('bats', 30), ('satan', 29), ('comic', 29), ('series', 28), ('gadget',
28), ('magoo', 28), ('cartoon', 27), ('evil', 27), ('starship', 27), ('alone', 27),
('kline', 27), ('webb', 27), ('gabriel', 26), ('scientist', 26), ('bugs', 25),
('cindy', 24), ('ace', 24), ('effects', 23), ('cage', 23)]
```

3) **For extra credit**, I also tried a different topic assignment initialization strategy by assigning all words in a particular document the same topic. In my opinion, the words related to the topics resulting from this initialization are not as thematically coherent as the ones which we get from random initialization. So, random initialization works better with the same 500 iterations.

In general, I would like to say that for topic modeling, one really needs to know the content of the dataset, otherwise uncovering the topics becomes a hassle. Detecting topics from the movie reviews was not an easy task at all (for me), maybe because I haven't watched so many movies yet.

```
alpha: 0.02
beta: 0.1
number of interations: 500
num topics to find: 20
num words to show per topic: 40
2021-01-28 14:55:00,608 - num_samples being used: 2000   num_unique_words: 46517
2021-01-28 14:55:00,761 - time taken to initialize counts: 0.01263570785522461 secs.
100%|████████████████████████████████████████████████| 500/500 [5:32:36<00:00, 39.91s/it]

#--------------------------------------#
Topic 0 - most words are war-action movies related. one needs a lot of
contextual knowledge to put all of these words in one particular domain.
[('impact', 39), ('comet', 37), ('event', 36), ('president', 35), ('war', 35),
('whale', 34), ('carlito', 33), ('chicken', 32), ('todd', 31), ('deep', 31),
('nuclear', 30), ('horizon', 30), ('warrior', 29), ('earth', 27), ('warriors', 27),
('russell', 26), ('country', 26), ('de', 26), ('america', 25), ('run', 25), ('pacino',
```

25), ('freeman', 25), ('banderas', 25), ('disaster', 25), ('maximus', 25), ('gladiator', 24), ('luis', 24), ('leader', 23), ('palma', 23), ('leder', 22), ('cuba', 22), ('casablanca', 21), ('tom', 20), ('cage', 19), ('franklin', 19), ('kombat', 19), ('13th', 18), ('rick', 18), ('gibson', 18), ('jolie', 18)]

#--------------------------------------#

Topic *1 - words look related to Tragedy/Drama Movie - Good Will Hunting, Matt Damon*

[('species', 54), ('nick', 45), ('hunting', 40), ('affleck', 40), ('sean', 34), ('damon', 32), ('rudy', 30), ('egoyan', 29), ('lambeau', 29), ('accident', 28), ('granger', 26), ('alicia', 23), ('sarah', 23), ('hereafter', 23), ('patrick', 23), ('bus', 22), ('ben', 22), ('marie', 22), ('jan', 22), ('children', 22), ('sweet', 21), ('justin', 20), ('jean', 20), ('lawyer', 20), ('town', 19), ('giles', 19), ('walsh', 19), ('ronin', 19), ('matt', 18), ('henstridge', 18), ('eve', 18), ('stretch', 18), ('exotica', 18), ('games', 17), ('musketeer', 17), ('madsen', 17), ('alien', 17), ('spencer', 16), ('tragedy', 16), ('college', 16)]

#--------------------------------------#

Topic *2 - Star Wars related*

[('wars', 163), ('star', 150), ('effects', 130), ('special', 124), ('tarzan', 113), ('earth', 82), ('phantom', 82), ('menace', 81), ('lucas', 80), ('planet', 80), ('jedi', 77), ('space', 66), ('young', 55), ('alien', 53), ('aliens', 52), ('sci-fi', 51), ('fiction', 47), ('jar', 47), ('science', 43), ('force', 42), ('george', 42), ('evil', 41), ('lost', 40), ('trilogy', 40), ('anakin', 38), ('darth', 37), ('zero', 36), ('obi-wan', 36), ('williams', 36), ('humans', 36), ('jane', 35), ('luke', 35), ('version', 34), ('carry', 34), ('ship', 34), ('queen', 34), ('original', 34), ('skywalker', 33), ('human', 33), ('jungle', 32)]

#--------------------------------------#

Topic *3 - the words do not seem to follow a specific topic since krripendorf tribe is a comedy movie while armageddon is sci-fi/disaster-related.. Then there are also words like cowboy, musical.*

[('joe', 169), ('american', 54), ('claire', 44), ('pitt', 40), ('krippendorf', 38), ('war', 38), ('death', 34), ('tribe', 33), ('armageddon', 32), ('willis', 31), ('kevin', 30), ('city', 30), ('bening', 28), ('political', 27), ('deuce', 26), ('spacey', 26), ('campaign', 25), ('tyler', 24), ('country', 24), ('cowboy', 23), ('eastwood', 23), ('president', 22), ('brooks', 22), ('york', 22), ('forlani', 22), ('advocate', 21), ('dream', 21), ('siege', 21), ('lumumba', 21), ('musical', 20), ('milton', 20), ('dreyfuss', 20), ('terrorist', 19), ('theron', 19), ('new', 18), ('mighty', 18), ('gorilla', 18), ('shandling', 18), ('washington', 18), ('men', 18)]

#--------------------------------------#

Topic *4 - again the topic is not easy to detect from the given words*

[('batman', 166), ('burton', 87), ('cage', 77), ('robin', 72), ('angels', 72), ('apes', 68), ('schumacher', 57), ('planet', 47), ('tim', 45), ('mr', 43), ('george', 43), ('8mm', 41), ('joel', 40), ('welles', 40), ('scorsese', 40), ('alex', 39), ('angel', 34), ('hollow', 34), ('destination', 32), ('freeze', 31), ('sleepy', 30), ('schwarzenegger', 29), ('series', 29), ('clooney', 27), ('ape', 27), ('danny', 27), ('snuff', 26), ('hammer', 26), ('sphere', 26), ('nicolas', 24), ('thurman', 24), ('carter', 23), ('plane', 23), ('leo', 23), ('francis', 23), ('boyle', 23), ('postman', 22), ('ivy', 22), ('seth', 21), ('1900', 21)]

#--------------------------------------#

Topic *5 - Action/thriller movies related*

[('jackie', 167), ('chan', 143), ('hong', 90), ('kong', 84), ('flynt', 80), ('fight', 80), ('martial', 71), ('action', 69), ('chinese', 68), ('arts', 61), ('austin', 57), ('evil', 53), ('myers', 53), ('van', 51), ('54', 50), ('mr', 49), ('powers', 48), ('american', 48), ('master', 46), ('li', 46), ('nights', 45), ('boogie', 44), ('larry', 43), ('fu', 42), ('lee', 40), ('kung', 39), ('damme', 38), ('carter', 36), ('jet', 36), ('nbsp', 35), ('mike', 35), ('hedwig', 34), ('drunken', 33), ('rush', 33), ('dr', 33), ('amistad', 31), ('spielberg', 30), ('club', 30), ('gattaca', 30), ('stunts', 30)]

#--#

Topic *6 - horror-fantasy related*

[('godzilla', 123), ('ship', 91), ('monster', 79), ('effects', 66), ('scary', 63), ('witch', 60), ('house', 60), ('blair', 59), ('babe', 55), ('deep', 54), ('crew', 48), ('special', 47), ('virus', 43), ('rising', 39), ('city', 39), ('horror', 39), ('creature', 38), ('x-files', 36), ('haunting', 36), ('anderson', 36), ('broderick', 36), ('pig', 34), ('york', 32), ('hill', 31), ('alien', 31), ('haunted', 30), ('mulder', 30), ('frankenstein', 29), ('duchovny', 29), ('gadget', 29), ('bont', 28), ('anaconda', 28), ('aliens', 25), ('edwards', 25), ('hank', 25), ('project', 23), ('scare', 22), ('geoffrey', 22), ('woods', 22), ('ocean', 22)]

#--#

Topic *7 - Action-thriler-comedy related*

[('murphy', 128), ('eddie', 87), ('wrestling', 50), ('gibson', 43), ('martin', 38), ('travolta', 36), ('ellie', 36), ('bowfinger', 34), ('mel', 33), ('contact', 32), ('mamet', 31), ('wcw', 31), ('jawbreaker', 31), ('brooks', 29), ('metro', 28), ('king', 28), ('ricky', 28), ('bobby', 27), ('foster', 26), ('liz', 26), ('g', 25), ('sean', 25), ('scott', 25), ('anderson', 25), ('league', 25), ('jimmy', 24), ('science', 24), ('battlefield', 24), ('mcgowan', 23), ('arquette', 22), ('clooney', 21), ('payback', 21), ('vince', 20), ('beverly', 20), ('caan', 20), ('bilko', 19), ('psychlos', 18), ('gordie', 18), ('steve', 18), ('ramsey', 18)]

#--#

Topic *8 - war-action/horror movies related*

[('scream', 206), ('horror', 139), ('ryan', 137), ('war', 105), ('killer', 90), ('troopers', 77), ('2', 75), ('starship', 67), ('williamson', 63), ('private', 61), ('hanks', 60), ('slasher', 57), ('julie', 56), ('summer', 53), ('verhoeven', 53), ('urban', 51), ('prinze', 50), ('soldiers', 49), ('battle', 47), ('sequel', 46), ('sidney', 44), ('legend', 43), ('robocop', 41), ('spielberg', 40), ('besson', 40), ('scary', 39), ('johnny', 39), ('freddie', 38), ('craven', 38), ('saving', 37), ('last', 36), ('kevin', 35), ('meg', 34), ('know', 33), ('bats', 33), ('college', 32), ('jr', 32), ('stab', 30), ('hewitt', 29), ('3', 29)]

#--#

Topic *9 real-life drama/documentary related*

[('simon', 85), ('life', 54), ('son', 48), ('beautiful', 40), ('political', 38), ('felix', 36), ('guido', 31), ('rocky', 31), ('sweetback', 30), ('brenner', 30), ('father', 29), ('stahl', 28), ('benigni', 26), ('family', 26), ('gavin', 25), ('history', 23), ('war', 23), ('man', 23), ('wallace', 23), ('rules', 22), ('historical', 22), ('stiller', 21), ('shark', 21), ('gibson', 21), ('braveheart', 21), ('sweet', 21), ('nazis', 21), ('jews', 20), ('pokemon', 20), ('violin', 20), ('holocaust', 20), ('french', 19), ('neil', 19), ('peebles', 19), ('spielberg', 19), ('subject', 19), ('camille', 18), ('west', 18), ('general', 18), ('darryl', 18)]

#--#

Topic *10 Shakespeare romance related movies and characters*

[('shakespeare', 83), ('spawn', 71), ('patch', 60), ('elizabeth', 58), ('paltrow', 45), ('tom', 45), ('love', 44), ('alice', 42), ('husband', 41), ('finn', 41), ('sixth', 41), ('cole', 40), ('romeo', 34), ('helen', 34), ('bill', 33), ('willis', 33), ('kate', 32), ('london', 31), ('ned', 31), ('friend', 30), ('oscar', 30), ('marriage', 29), ('gwyneth', 29), ('wife', 28), ('jesus', 27), ('anna', 27), ('monty', 26), ('juliet', 26), ('fiennes', 25), ('annie', 25), ('bob', 24), ('rush', 24), ('queen', 24), ('hamlet', 24), ('danes', 24), ('judith', 24), ('adams', 23), ('william', 23), ('kubrick', 23), ('crowe', 23)]

#--------------------------------------#

Topic *11 tarantino movies i.e. crime related*

[('tarantino', 111), ('jackie', 101), ('mike', 88), ('brown', 73), ('pulp', 69), ('fiction', 61), ('jackson', 52), ('ordell', 49), ('mary', 44), ('things', 42), ('sex', 41), ('brothers', 41), ('quentin', 41), ('wild', 39), ('dillon', 38), ('murray', 37), ('paulie', 37), ('bacon', 36), ('farrelly', 33), ('campbell', 31), ('vegas', 31), ('richards', 30), ('ted', 29), ('robert', 29), ('irene', 29), ('matt', 29), ('kelly', 29), ('grier', 28), ('charlie', 28), ('blue', 28), ('segment', 27), ('crystal', 27), ('henry', 27), ('farrellys', 26), ('pam', 26), ('crime', 25), ('eszterhas', 23), ('poker', 23), ('nomi', 23), ('sam', 23)]

#--------------------------------------#

Topic *12 truman Burbank show movie related*

[('truman', 155), ('carrey', 93), ('jim', 64), ('west', 62), ('wild', 62), ('smith', 49), ('jackal', 44), ('damme', 43), ('gere', 40), ('show', 28), ('webb', 28), ('frequency', 28), ('van', 27), ('garofalo', 26), ('ace', 26), ('ray', 25), ('knock', 24), ('quinn', 24), ('kline', 23), ('ford', 23), ('sethe', 23), ('willis', 23), ('gordon', 22), ('grant', 22), ('hudson', 22), ('donnie', 21), ('ventura', 21), ('beloved', 21), ('weir', 20), ('christmas', 20), ('burbank', 20), ('andrew', 18), ('gangster', 18), ('irish', 18), ('ricci', 17), ('christof', 17), ('tsui', 17), ('schneider', 17), ('martha', 17), ('heche', 15)]

#--------------------------------------#

Topic *13 - Adult movie related*

[('smith', 79), ('jay', 68), ('kevin', 64), ('bob', 58), ('derek', 58), ('gay', 53), ('sex', 51), ('pie', 49), ('american', 48), ('silent', 47), ('rock', 45), ('park', 43), ('sonny', 41), ('stuart', 41), ('flubber', 36), ('school', 33), ('god', 30), ('seth', 27), ('jason', 26), ('college', 25), ('lucy', 25), ('apostle', 25), ('amy', 25), ('professor', 24), ('dogma', 24), ('affleck', 24), ('rated', 24), ('duvall', 23), ('jim', 21), ('sara', 21), ('church', 20), ('r', 20), ('teenagers', 20), ('gloria', 19), ('norton', 19), ('ben', 19), ('brother', 19), ('alan', 19), ('x', 19), ('levy', 19)]

#--------------------------------------#

Topic *14 - animated movies for family/children related*

[('disney', 254), ('animated', 136), ('animation', 116), ('king', 106), ('mulan', 95), ('voice', 80), ('children', 65), ('shrek', 63), ('family', 54), ('kids', 54), ('princess', 50), ('songs', 50), ('prince', 49), ('bug', 43), ('cauldron', 41), ('feature', 41), ('fairy', 40), ('antz', 40), ('voiced', 38), ('army', 38), ('adults', 38), ('cartoon', 37), ('father', 35), ('snake', 34), ('giant', 33), ('british', 33), ('dragon', 33), ('japanese', 32), ('mrs', 32), ('mouse', 32), ('tale', 32), ('mermaid', 32), ('leila', 30), ('altman', 30), ('magoo', 29), ('marty', 28), ('jordan', 28), ('murphy', 28), ('fiona', 28), ('battle', 27)]

#--------------------------------------#

Topic *15 - extraterrestrial-vampire-witchcraft related*

[('mars', 120), ('vampire', 117), ('vampires', 77), ('blade', 73), ('mission', 72), ('carpenter', 65), ('bulworth', 60), ('arnold', 59), ('city', 52), ('crow', 49), ('beatty', 40), ('dark', 36), ('horror', 33), ('snipes', 32), ('ghosts', 31), ('aliens', 29), ('space', 29), ('margaret', 29), ('satan', 28), ('woods', 28), ('planet', 28), ('devil', 27), ('crystal', 26), ('redford', 26), ('gabriel', 26), ('reese', 25), ('horse', 24), ('byrne', 24), ('tim', 23), ('grace', 23), ('schwarzenegger', 23), ('red', 23), ('beau', 22), ('church', 22), ('wesley', 22), ('thomas', 21), ('robert', 21), ('quaid', 21), ('sutherland', 21), ('vacation', 21)]

#--------------------------------------#

Topic *16 - Romantic movies related*

[('wedding', 96), ('titanic', 90), ('sandler', 77), ('carter', 69), ('julia', 69), ('ship', 68), ('matrix', 59), ('cameron', 58), ('reeves', 58), ('dude', 55), ('lebowski', 55), ('big', 52), ('rose', 46), ('romantic', 44), ('singer', 44), ('keanu', 41), ('barrymore', 38), ('malkovich', 34), ('crawford', 32), ('kate', 32), ('roberts', 31), ('cusack', 30), ('coen', 29), ('floor', 28), ('washington', 28), ('hurricane', 27), ('goodman', 27), ('sam', 27), ('paxton', 26), ('momma', 24), ('jack', 24), ('robbie', 24), ('adam', 23), ('bowling', 23), ('brothers', 23), ('dicaprio', 23), ('buscemi', 23), ('friend', 23), ('tango', 23), ('winslet', 23)]

#--------------------------------------#

Topic *17 - not easy to detect a topic from these words, as they don't seem to be coherent, some words look related to the Palmetto movie and cast*

[('harry', 141), ('woody', 92), ('allen', 80), ('dog', 68), ('max', 67), ('toy', 62), ('seagal', 62), ('judd', 46), ('woo', 43), ('lee', 43), ('fugitive', 38), ('melvin', 36), ('hunt', 34), ('libby', 34), ('jones', 34), ('son', 33), ('ghost', 32), ('paul', 32), ('computer', 31), ('cop', 31), ('simon', 29), ('tommy', 28), ('buzz', 28), ('carol', 27), ('palmetto', 26), ('toys', 26), ('memphis', 25), ('nello', 24), ('chris', 23), ('jeopardy', 23), ('cruise', 23), ('annie', 23), ('ashley', 23), ('freeman', 22), ('harrelson', 21), ('kilmer', 21), ('october', 21), ('double', 20), ('shue', 20), ('homer', 20)]

#--------------------------------------#

Topic *18 - fiction drama related*

[('alien', 189), ('trek', 121), ('star', 112), ('series', 101), ('aliens', 89), ('bond', 72), ('spice', 67), ('ripley', 63), ('girls', 50), ('football', 43), ('jerry', 36), ('3', 36), ('argento', 35), ('insurrection', 34), ('school', 31), ('contact', 29), ('existenz', 28), ('james', 27), ('fincher', 27), ('jakob', 27), ('enterprise', 26), ('virtual', 25), ('data', 25), ('x-files', 25), ('fans', 25), ('weaver', 24), ("ba'ku", 23), ('nbsp', 23), ('planet', 23), ('picard', 22), ('science', 22), ('carver', 21), ('crew', 21), ('mercury', 20), ('government', 19), ('conspiracy', 18), ('hughes', 18), ('fiction', 17), ('coach', 17), ('lucas', 17)]

#--------------------------------------#

Topic *19 - words related to writing a Movie Review*

[('film', 9442), ('movie', 5671), ('one', 5580), ('like', 3535), ('even', 2551), ('good', 2313), ('time', 2275), ('would', 2259), ('story', 2140), ('much', 2017), ('character', 1994), ('also', 1957), ('get', 1921), ('characters', 1853), ('two', 1819), ('first', 1766), ('see', 1728), ('way', 1667), ('well', 1653), ('could', 1608), ('make', 1582), ('really', 1554), ('films', 1519), ('little', 1489), ('plot', 1458), ('people', 1441), ('life', 1424), ('bad', 1373), ('scene', 1369), ('never', 1359), ('man', 1315), ('best', 1284), ('many', 1268), ('new', 1259), ('scenes', 1243),

```
('movies', 1188), ('know', 1174), ('great', 1130), ('another', 1117), ('director',
1103)]
```

TREC-Incident Streams Dataset

In addition to the Movie Reviews Corpus, I carried out experiments with dataset(s) from TREC-Incident Streams Challenge. TREC-IS provides multiple Twitter datasets collected from a range of past wildfire, earthquake, flood, typhoon/hurricane, bombing and shooting events. For the purpose of topic modelling, I used the chileEarthquake2014 and australiaBushfire2013 datasets, preprocessed and combined them. The combined dataset (953 tweets) can be found under *data/pptweets_chileEarthquake2014_australiaBushfire2013.txt* and *preprocess.py* is the script I used for preprocessing the original datasets.

After running multiple experiments with several different parameters, the best results that I got was with **2000 iterations, 3 topics and top 20 words**. The words seem to be a bit more coherent than the two other experiments that follow. Note: I performed more than 3 experiments, x

1) num_samples being used: 953, num_unique_words: 2370
iterations : 2000, topics : 3, top words: 20
Runtime of the experiment: 17 mins, 1.92 secs/iteration

#-------------------------------------#
Topic 0 - ChileEarthquake
[('chile', 318.0), ('hquake', 257.0), ('tsunami', 84.0), ('magnitude', 64.0), ('hern', 56.0), ('coast', 54.0), ('quake', 47.0), ('strikes', 34.0), ('powerful', 31.0), ('dead', 27.0), ('hit', 24.0), ('people', 22.0), ('iquique', 21.0), ('warning', 21.0), ('news', 16.0), ('chileea', 15.0), ('via', 15.0), ('hits', 15.0), ('massive', 13.0), ('ale', 13.0)]

#-------------------------------------#
Topic 1 - Australia bushfires
[('fires', 190.0), ('australia', 149.0), ('fire', 105.0), ('nsw', 104.0), ('nswfires', 102.0), ('bush', 67.0), ('sydney', 58.0), ('amp', 32.0), ('south', 32.0), ('safe', 30.0), ('new', 29.0), ('thoughts', 26.0), ('wales', 25.0), ('stay', 24.0), ('bushfires', 24.0), ('climate', 23.0), ('state', 22.0), ('firefighters', 20.0), ('fighting', 19.0), ('affected', 18.0)]

#-------------------------------------#
Topic 2 - news about bushfires
[('nswfires', 199.0), ('fire', 108.0), ('nswrfs', 68.0), ('nsw', 56.0), ('sydney', 43.0), ('fires', 40.0), ('rfs', 26.0), ('blue', 25.0), ('emergency', 24.0), ('bushfire', 24.0), ('springwood', 23.0), ('amp',

22.0), ('bushfires', 22.0), ('live', 17.0), ('smoke', 17.0), ('update', 17.0), ('today', 17.0), ('warning', 16.0), ('mountains', 16.0), ('please', 14.0)]

2) iterations 1000, topics 3, top_words 20
num_samples being used: 953, num_unique_words: 2370
Runtime of the experiment: 8 mins 33 secs, 1.95 secs/iteration

#----------#
Topic 0 - chile earthquake
[('chile', 319.0), ('hquake', 257.0), ('tsunami', 84.0), ('magnitude', 64.0), ('hern', 55.0), ('coast', 54.0), ('quake', 47.0), ('strikes', 34.0), ('powerful', 31.0), ('dead', 27.0), ('hit', 25.0), ('people', 22.0), ('iquique', 21.0), ('warning', 21.0), ('news', 17.0), ('via', 15.0), ('chileea', 15.0), ('hits', 15.0), ('massive', 13.0), ('ale', 13.0)]

#----------#
Topic 1 - australia bushfires
[('fires', 184.0), ('australia', 150.0), ('fire', 103.0), ('nswfires', 103.0), ('nsw', 101.0), ('bush', 68.0), ('sydney', 54.0), ('safe', 33.0), ('amp', 32.0), ('south', 32.0), ('thoughts', 28.0), ('new', 27.0), ('wales', 25.0), ('climate', 23.0), ('stay', 23.0), ('bushfires', 23.0), ('photo', 20.0), ('affected', 20.0), ('fighting', 18.0), ('state', 18.0)]

#----------#
Topic 2 - news about bushfires
[('nswfires', 200.0), ('fire', 110.0), ('nswrfs', 68.0), ('nsw', 59.0), ('sydney', 47.0), ('fires', 46.0), ('emergency', 27.0), ('rfs', 26.0), ('blue', 25.0), ('bushfires', 23.0), ('bushfire', 22.0), ('amp', 22.0), ('springwood', 22.0), ('smoke', 19.0), ('today', 18.0), ('update', 18.0), ('mountains', 16.0), ('warning', 16.0), ('live', 16.0), ('burning', 15.0)]

3) niterations 1000, ntopics 4, top_words 20
num_samples being used: 953 num_unique_words: 2370
Runtime of the experiment: 8 mins 15 secs, 2 secs/iterations

#----------#
Topic 0 - chile earthquake
[('chile', 319.0), ('hquake', 257.0), ('tsunami', 84.0), ('magnitude', 64.0), ('hern', 55.0), ('coast', 54.0), ('quake', 47.0), ('strikes', 34.0), ('powerful', 31.0), ('dead', 27.0), ('hit', 24.0), ('warning', 22.0), ('people', 22.0), ('iquique', 21.0), ('news', 17.0), ('chileea', 15.0), ('hits', 14.0), ('via', 14.0), ('massive', 13.0), ('ale', 13.0)]

#----------#
Topic 1 - australia bush fires
[('fires', 133.0), ('australia', 100.0), ('nsw', 75.0), ('nswfires', 64.0), ('fire', 59.0), ('bush', 53.0), ('south', 34.0), ('safe', 32.0), ('sydney', 32.0), ('thoughts', 30.0), ('new', 27.0), ('wales', 25.0),

('stay', 24.0), ('state', 21.0), ('bushfires', 19.0), ('firefighters', 19.0), ('everyone', 18.0), ('amp', 18.0), ('emergency', 18.0), ('affected', 17.0)]

Topic 2 and topic 3 are both very similar, about the news of the fires.
#-------------------------------------#
Topic 2
[('nswfires', 160.0), ('fire', 97.0), ('nswrfs', 64.0), ('nsw', 51.0), ('fires', 41.0), ('amp', 26.0), ('blue', 23.0), ('springwood', 23.0), ('rfs', 23.0), ('emergency', 23.0), ('sydney', 21.0), ('today', 20.0), ('update', 17.0), ('mountains', 16.0), ('bushfires', 15.0), ('warning', 14.0), ('bushfire', 14.0), ('homes', 14.0), ('burning', 13.0), ('live', 13.0)]

#-------------------------------------#
Topic 3
[('nswfires', 79.0), ('fire', 57.0), ('fires', 56.0), ('sydney', 48.0), ('australia', 42.0), ('nsw', 34.0), ('climate', 20.0), ('smoke', 19.0), ('bush', 19.0), ('change', 17.0), ('abbott', 13.0), ('rain', 12.0), ('water', 12.0), ('bushfires', 12.0), ('amp', 11.0), ('sta', 11.0), ('auspol', 10.0), ('good', 10.0), ('day', 10.0), ('bushfire', 10.0)]

Some interesting observations/lessons learnt during the optimization process

1) using `defaultdict(list)` for maintaining topic assignments works faster than using a `variable 2D python list` or `defaultdict(int)`. The runtime reduced by approximately 4+ secs per iteration when the variable 2D vector data structure for topic assignments was replaced by defaultdict(...). There was also a reduction in the initialization runtime by the same factor.

The difference between defaultdict(int) and defaultdict(list) for topic assignment is not very significant and might also be system dependent (to be checked). But results from on my machine are as follows:

* For `defaultdict[tuple] = int` assignment, initialization took 3.42 secs and 33.64 secs/iteration.

* For `defaultdict[int] = [topic1, topic2,..]` assignment, initialization took 3.23 secs and 32.21 secs/ iteration.

2) using np.zeros(shape, dtype=int) is faster than np.zeros(shape, dtype=np.int32/np.int16) and np.zeros(shape) is the fastest. There was a significant reduction in initialization runtime from 8+ secs to 3+ secs and 10+secs/iteration when np.zeros(shape, dtype=np.int32/np.int16) was replaced with np.zeros(shape, dtype=int). With np.zeros(shape), the reduction was of around 1-2 secs/iteration, with no visible reduction in initialization runtime.

Both of these observations were made on the default settings with 2000 sentences of the movie review corpus.