Q4. (40 pts) Please finish the following text mining mini-project. Please submit your source code together with your PDF file. All text data are provided in the folder “Data”.

In [34]:

```
import os

folder_path = '/Users/varun/Downloads/Data'

files = []

for filename in os.listdir(folder_path):
    if filename.endswith('.txt'):
        file_path = os.path.join(folder_path, filename)
        with open(file_path, 'r', encoding="ISO-8859-1") as file:
            contents = file.read()
            #print(f'---{filename}---')
            #print(contents)
            files.append(contents)
```

In [35]:

```
files[0]
```

```
Out[35]: '                  THE HOLE THAT WAS TOO NARROW\n\n   Once upon a time. . . a
         stoat was so greedy that he would eat anything that\ncame his way. But he w
         as punished for his greed. He found some old stale eggs\nin a barn and, as
         usual, gobbled the lot. However, he soon started to feel\nagonizing pains i
         n his tummy, his eyes grew dim and he broke out in a cold \nsweat. For days
         , he lay between life and death, then the fever dropped. The\nfirst time he
         dared climb a tree to rob a nest, thin and weak with his \ntrousers danglin
         g over an empty stomach, he became dizzy and fell. That is how\nhe twisted
         his ankle. Sick with hunger, he limped about in search of food, but\nthat m
         ade him feel even hungrier than before. Then good luck came his way.\nAltho
         ugh wary of venturing too close to human habitations, he was so hungry he\n
         went up to a tavern on the outskirts of the village. The air was full of \n
         lovely smells and the poor stoat felt his mouth watering as he pictured all
         \nthe nice things inslde. An inviting smell coming from a crack in the wall
         \nseemed to be stronger than the others. Thrusting his nose into the crack,
         he \nwas greeted by a waft of delicious scents. The stoat frantically clawe
         d at the\ncrack with his paws and teeth, trying to widen it. Slowly the pla
         ster between \nthe blocks of rubble began to crumble, till all he had to do
         was move a stone.\nShoving with all his might, the stoat made a hole. And t
         hen a really wonderful\nsight met his gaze. He was inside the pantry, where
         hams, salamis, cheeses, \nhoney, jam and nuts were stored. Overwhelmed by i
         t all, the stoat could not \nmake up his mind what to taste first. He jumpe
         d from one thing to another, \nmunching all the time, till his tummy was fu
         ll. Satisfied at last, he fell \nasleep. Then he woke again, had another fe
         ast and went back to sleep. With all\nthis food, his strength returned, and
         next day, the stoat was strong enough to\nclimb up to the topmost shelves a
         nd select the tastiest delicacies. By this \ntime, he was just having a nib
         ble here and a nibble there. But he never \nstopped eating: he went on and
         on and on. By now, he was very full indeed, as \nhe chattered to himself: "
         Salami for starters . . . no, the ham\'s better! Some\nsoft cheese and a sp
         ot of mature cheese as well . . . I think I\'ll have a \npickled sausage to
         o . . ."\n   In only a few days, the stoat had become very fat and his trou
         ser button \nhad popped off over a bulging tummy. But of course, the stoat\
         's fantastic luck\ncould not last for ever.  \n   One afternoon, the stoat
         froze in mid-munch at the creak of a door. Heavy \nfootsteps thumped down t
         he stairs, and the stoat looked helplessly round. Fear\nof discovery sent h
         im hunting for a way to escape. He ran towards the hole in \nthe wall throu
         gh which he had come. But though his head and shoulders entered \nthe hole,
         his tummy, which had grown much larger since the day he had come in,\nslmpl
         y would not pass. The stoat was in a dangerous position: he was stuck! \nTw
         o thick hands grabbed him by the tail.\n   "You horrid little robber! So yo
         u thought you\'d get away, did you? I\'ll \nsoon deal with you!" \n   Stran
         ge though it may sound, the only thought in the greedy stoat\'s head \nwas
         a longing to be starving of hunger again . . .\n'
```

```
In [ ]:
```

a. (10 pts) Preprocess all documents by considering the following typical steps:

[1]. Convert to lowercase

In [36]:

```
for i in range(len(files)):
    files[i] = files[i].lower()

print('Have converted all the words from uppercase to lowercase. Printing
```

```
Have converted all the words from uppercase to lowercase. Printing only fir
st document                    the hole that was too narrow

   once upon a time. . . a stoat was so greedy that he would eat anything t
hat
came his way. but he was punished for his greed. he found some old stale eg
gs
in a barn and, as usual, gobbled the lot. however, he soon started to feel
agonizing pains in his tummy, his eyes grew dim and he broke out in a cold
sweat. for days, he lay between life and death, then the fever dropped. the
first time he dared climb a tree to rob a nest, thin and weak with his
trousers dangling over an empty stomach, he became dizzy and fell. that is
how
he twisted his ankle. sick with hunger, he limped about in search of food,
but
that made him feel even hungrier than before. then good luck came his way.
although wary of venturing too close to human habitations, he was so hungry
he
went up to a tavern on the outskirts of the village. the air was full of
lovely smells and the poor stoat felt his mouth watering as he pictured all
the nice things inslde. an inviting smell coming from a crack in the wall
seemed to be stronger than the others. thrusting his nose into the crack, h
e
was greeted by a waft of delicious scents. the stoat frantically clawed at
the
crack with his paws and teeth, trying to widen it. slowly the plaster betwe
en
the blocks of rubble began to crumble, till all he had to do was move a sto
ne.
shoving with all his might, the stoat made a hole. and then a really wonder
ful
sight met his gaze. he was inside the pantry, where hams, salamis, cheeses,
honey, jam and nuts were stored. overwhelmed by it all, the stoat could not
make up his mind what to taste first. he jumped from one thing to another,
munching all the time, till his tummy was full. satisfied at last, he fell
asleep. then he woke again, had another feast and went back to sleep. with
all
this food, his strength returned, and next day, the stoat was strong enough
to
climb up to the topmost shelves and select the tastiest delicacies. by this
time, he was just having a nibble here and a nibble there. but he never
stopped eating: he went on and on and on. by now, he was very full indeed,
as
he chattered to himself: "salami for starters . . . no, the ham's better! s
ome
soft cheese and a spot of mature cheese as well . . . i think i'll have a
pickled sausage too . . ."
   in only a few days, the stoat had become very fat and his trouser button
had popped off over a bulging tummy. but of course, the stoat's fantastic l
uck
```

```
could not last for ever.
   one afternoon, the stoat froze in mid-munch at the creak of a door. heav
y
footsteps thumped down the stairs, and the stoat looked helplessly round. f
ear
of discovery sent him hunting for a way to escape. he ran towards the hole
in
the wall through which he had come. but though his head and shoulders enter
ed
the hole, his tummy, which had grown much larger since the day he had come
in,
slmply would not pass. the stoat was in a dangerous position: he was stuck!
two thick hands grabbed him by the tail.
   "you horrid little robber! so you thought you'd get away, did you? i'll
soon deal with you!"
   strange though it may sound, the only thought in the greedy stoat's head
was a longing to be starving of hunger again . . .
```

In [ ]:

```

```

[2]. Remove stop words

In [37]:

```python
import nltk
from nltk.corpus import stopwords

# Download the stop words list (only need to do this once)
#nltk.download('stopwords')
```

In [38]:

```python
# Load the stop words list
stop_words = set(stopwords.words('english'))

print("Total number of words of first document before removing stopwords",l

for i in range(1):

    words = files[i].split()

    filtered_words = [word for word in words if word.lower() not in stop_wo

    files[i]= ' '.join(filtered_words)

print("Total number of words of first document after removing stopwords",le
print('\n')
print(files[0])
```

```
Total number of words of first document before removing stopwords 614
Total number of words of first document after removing stopwords 324


hole narrow upon time. . . stoat greedy would eat anything came way. punish
ed greed. found old stale eggs barn and, usual, gobbled lot. however, soon
started feel agonizing pains tummy, eyes grew dim broke cold sweat. days, l
ay life death, fever dropped. first time dared climb tree rob nest, thin we
ak trousers dangling empty stomach, became dizzy fell. twisted ankle. sick
hunger, limped search food, made feel even hungrier before. good luck came
way. although wary venturing close human habitations, hungry went tavern ou
tskirts village. air full lovely smells poor stoat felt mouth watering pict
ured nice things inslde. inviting smell coming crack wall seemed stronger o
thers. thrusting nose crack, greeted waft delicious scents. stoat frantical
ly clawed crack paws teeth, trying widen it. slowly plaster blocks rubble b
egan crumble, till move stone. shoving might, stoat made hole. really wonde
rful sight met gaze. inside pantry, hams, salamis, cheeses, honey, jam nuts
stored. overwhelmed all, stoat could make mind taste first. jumped one thin
g another, munching time, till tummy full. satisfied last, fell asleep. wok
e again, another feast went back sleep. food, strength returned, next day,
stoat strong enough climb topmost shelves select tastiest delicacies. time,
nibble nibble there. never stopped eating: went on. now, full indeed, chatt
ered himself: "salami starters . . . no, ham's better! soft cheese spot mat
ure cheese well . . . think i'll pickled sausage . . ." days, stoat become
fat trouser button popped bulging tummy. course, stoat's fantastic luck cou
ld last ever. one afternoon, stoat froze mid-munch creak door. heavy footst
eps thumped stairs, stoat looked helplessly round. fear discovery sent hunt
ing way escape. ran towards hole wall come. though head shoulders entered h
ole, tummy, grown much larger since day come in, slmply would pass. stoat d
angerous position: stuck! two thick hands grabbed tail. "you horrid little
robber! thought get away, you? i'll soon deal you!" strange though may soun
d, thought greedy stoat's head longing starving hunger . . .
```

In [ ]:

```

```

[3]. Remove Punctuation

In [39]:

```python
import string

for i in range(len(files)):


# Define a translation table to remove punctuation
    translator = str.maketrans('', '', string.punctuation)

# Use the translate method to remove punctuation
    files[i] = files[i].translate(translator)

# Print the result

print("Removed all the punctuations:\n",files[0])
```

```
Removed all the punctuations:
 hole narrow upon time   stoat greedy would eat anything came way punished
greed found old stale eggs barn and usual gobbled lot however soon started
feel agonizing pains tummy eyes grew dim broke cold sweat days lay life dea
th fever dropped first time dared climb tree rob nest thin weak trousers da
ngling empty stomach became dizzy fell twisted ankle sick hunger limped sea
rch food made feel even hungrier before good luck came way although wary ve
nturing close human habitations hungry went tavern outskirts village air fu
ll lovely smells poor stoat felt mouth watering pictured nice things inslde
inviting smell coming crack wall seemed stronger others thrusting nose crac
k greeted waft delicious scents stoat frantically clawed crack paws teeth t
rying widen it slowly plaster blocks rubble began crumble till move stone s
hoving might stoat made hole really wonderful sight met gaze inside pantry
hams salamis cheeses honey jam nuts stored overwhelmed all stoat could make
mind taste first jumped one thing another munching time till tummy full sat
isfied last fell asleep woke again another feast went back sleep food stren
gth returned next day stoat strong enough climb topmost shelves select tast
iest delicacies time nibble nibble there never stopped eating went on now f
ull indeed chattered himself salami starters    no hams better soft cheese
spot mature cheese well    think ill pickled sausage    days stoat become f
at trouser button popped bulging tummy course stoats fantastic luck could l
ast ever one afternoon stoat froze midmunch creak door heavy footsteps thum
ped stairs stoat looked helplessly round fear discovery sent hunting way es
cape ran towards hole wall come though head shoulders entered hole tummy gr
own much larger since day come in slmply would pass stoat dangerous positio
n stuck two thick hands grabbed tail you horrid little robber thought get a
way you ill soon deal you strange though may sound thought greedy stoats he
ad longing starving hunger
```

In [ ]:

```python

```

[4]. Single Characters

In [40]:

```python
import re

for i in range(len(files)):

    regex = r'\b\w\b'

    files[i] = re.sub(regex, "", files[i])

# Print the processed document
print("Removed all the single characters:\n",files[0])
```

```
Removed all the single characters:
 hole narrow upon time   stoat greedy would eat anything came way punished
greed found old stale eggs barn and usual gobbled lot however soon started
feel agonizing pains tummy eyes grew dim broke cold sweat days lay life dea
th fever dropped first time dared climb tree rob nest thin weak trousers da
ngling empty stomach became dizzy fell twisted ankle sick hunger limped sea
rch food made feel even hungrier before good luck came way although wary ve
nturing close human habitations hungry went tavern outskirts village air fu
ll lovely smells poor stoat felt mouth watering pictured nice things inslde
inviting smell coming crack wall seemed stronger others thrusting nose crac
k greeted waft delicious scents stoat frantically clawed crack paws teeth t
rying widen it slowly plaster blocks rubble began crumble till move stone s
hoving might stoat made hole really wonderful sight met gaze inside pantry
hams salamis cheeses honey jam nuts stored overwhelmed all stoat could make
mind taste first jumped one thing another munching time till tummy full sat
isfied last fell asleep woke again another feast went back sleep food stren
gth returned next day stoat strong enough climb topmost shelves select tast
iest delicacies time nibble nibble there never stopped eating went on now f
ull indeed chattered himself salami starters    no hams better soft cheese
spot mature cheese well    think ill pickled sausage    days stoat become f
at trouser button popped bulging tummy course stoats fantastic luck could l
ast ever one afternoon stoat froze midmunch creak door heavy footsteps thum
ped stairs stoat looked helplessly round fear discovery sent hunting way es
cape ran towards hole wall come though head shoulders entered hole tummy gr
own much larger since day come in slmply would pass stoat dangerous positio
n stuck two thick hands grabbed tail you horrid little robber thought get a
way you ill soon deal you strange though may sound thought greedy stoats he
ad longing starving hunger
```

In [ ]:

[5]. Stemming and Lemmatization, e.g., change “playing” and “played” to play

In [41]:

```python
import nltk
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk.tokenize import word_tokenize
#nltk.download('punkt')
#nltk.download('wordnet')
```

In [42]:

```
# Initialize a Porter stemmer and lemmatizer
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

for i in range(len(files)):

# Tokenize the sentence into individual words
    words = word_tokenize(files[i])

    stemmed_words = [stemmer.stem(word) for word in words]

    lemmatized_words = [lemmatizer.lemmatize(word) for word in words]

#It is advisable to use lemmatization instead of stemming as the words are
    files[i] = " ".join(lemmatized_words)

print("Stemmed sentence:\n\n", " ".join(stemmed_words))
print("\nLemmatized sentence:\n\n", " ".join(lemmatized_words))
```

```
Stemmed sentence:

 the weep princess onc upon time greedi emperor forc hi subject to pay heav
i tax not onli the poor were squeez but the nobl in thi immens empir were h
ighli tax too at last tire of be crush by tax the nobl held protest meet wh
en the emperor heard about thi he took fright for he fear rebellion so he s
ent out thi proclam to put an end to their complaint the nobleman that can
make my daughter sarah smile again for she mourn the loss of her fianc will
never pay tax again thi caus an uproar at the protest meet most of the prin
c decid there wa no need now to complain for each wa quit sure he would suc
ceed where other might fail so off they went to get readi to tri and make s
arah smile but some of the nobl warn their fellow that with hi word the emp
eror wa not realli abolish ani tax at all from that day on long process of
nobl knight troop from all over the empir to the palac to tri and consol th
e weep princess the crowd cheer them as they pass but when they return with
bow head the same crowd boo and whistl at their failur the day went by and
the list of defeat knight grew longer indian circassian arab and turk from
all over the provinc came bold young men bounc with confid and hope but the
minut the princess set eye on them she just wept and wept the emperor wa de
light for each failur meant anoth taxpay even the common folk seem content
to see that the rich too did not alway get what they want the onli unhappi
person among them wa sarah who went on weep one day mongol princ seem to be
on the point of win smile he thrum hi balalaika for hour play first sad tun
e then more cheer one till he finish by play merri jig the princess sat for
age stare at him eye and the onlook thought she wa about to smile instead s
he burst into flood of tear to everyon disappoint kurdish chief fame for hi
humour who had alreadi kept the court in fit of laughter tri to steal smile
from sarah with hi witti remark but the princesss dark eye fill with tear n
oblemen came from as far away as persia but in vain the onli person who had
not yet appear wa omar the chief of the tiniest farthest away provinc brigh
t intellig young man he had cleverli got the better of certain greedi ambit
i rel that tri to take away hi power when he succeed hi uncl as chief the e
mperor messeng had taken long time to reach thi remot realm and though omar
set out at onc on hear the news he rode for mani day on hi fine black hors
then one even he reach the palac when the tire and dusti travel explain to
the stabl boy whi he had come they laugh in scorn but they had order to obe
```

```
y so they told him to enter it late they said and you wont see the princess
till tomorrow the emperor other daughter howev were soon told of the new ar
riv he the most handsom of them all exclaim one of the servant so marika th
e emperor youngest and prettiest daughter with her sister peek through wind
ow at the sleep omar next morn the emperor order the newcom to be led befor
sarah the court crowd round to watch unlik all the other suitor omar did no
th at all to amus the princess he stare at sarah without say word and she s
tare back with an empti look on her face the two young peopl stare silent a
t each other then omar went back to the emperor and said sire give me your 
sceptr and will solv the problem of sarah surpris at such an odd request th
e emperor follow omar into sarah room the other princess cluster round smil
e and admir the handsom young man with deep bow to sarah omar straighten up
and dealt her blow on the head with the sceptr scream fill the air the empe
ror threw up hl arm in rage and hi daughter fled in all direct the guard dr
ew their sword then the whole room stop in amaz for out of sarah head which
had been chop off by the blow roll broken spring and piec of metal the prin
cess that never smile wa doll perfect dolll and nobodi had ever been awar o
f it except omar the onli princess that couldnt stop laugh wa marika the em
peror glare at her be quiet he order but he too saw the funni side of it fo
r the crafti emperor had been make use of sarah the doll as way of guarante
himself steadi flow of tax from all hi subject and now man more cun than hi
mself had expos hi trick the emperor had sudden thought he would rid himsel
f of the cheeki marika and gain an astut soninlaw abl to help him hold onto
hi kingdom you should be put to death for thi insol he said but im go to sp
are your life if you marri my youngest daughter of cours you wont need to p
ay tax smile at happi marika omar nod silent down in the depth of hi mind h
e wa think one day dear fatherinlaw ill be sit on your imperi throne and he
wa few year later

Lemmatized sentence:

 the weeping princess once upon time greedy emperor forced his subject to p
ay heavy tax not only the poor were squeezed but the noble in this immense 
empire were highly taxed too at last tired of being crushed by tax the nobl
e held protest meeting when the emperor heard about this he took fright for
he feared rebellion so he sent out this proclamation to put an end to their
complaint the nobleman that can make my daughter sarah smile again for shes
mourning the loss of her fiance will never pay tax again this caused an upr
oar at the protest meeting most of the prince decided there wa no need now 
to complain for each wa quite sure he would succeed where others might fail
so off they went to get ready to try and make sarah smile but some of the n
oble warned their fellow that with his word the emperor wa not really aboli
shing any tax at all from that day on long procession of noble knight troop
ed from all over the empire to the palace to try and console the weeping pr
incess the crowd cheered them a they passed but when they returned with bow
ed head the same crowd booed and whistled at their failure the day went by 
and the list of defeated knight grew longer indian circassian arab and turk
from all over the province came bold young men bouncing with confidence and
hope but the minute the princess set eye on them she just wept and wept the
emperor wa delighted for each failure meant another taxpayer even the commo
n folk seemed contented to see that the rich too did not always get what th
ey wanted the only unhappy person among them wa sarah who went on weeping o
ne day mongol prince seemed to be on the point of winning smile he thrummed
his balalaika for hour playing first sad tune then more cheerful one till h
e finished by playing merry jig the princess sat for age staring at him eye
d and the onlooker thought she wa about to smile instead she burst into flo
od of tear to everyones disappointment kurdish chief famed for his humour w
```

```
ho had already kept the court in fit of laughter tried to steal smile from
sarah with his witty remark but the princess dark eye filled with tear nobl
eman came from a far away a persia but in vain the only person who had not
yet appeared wa omar the chief of the tiniest farthest away province bright
intelligent young man he had cleverly got the better of certain greedy ambi
tious relative that tried to take away his power when he succeeded his uncl
e a chief the emperor messenger had taken long time to reach this remote re
alm and though omar set out at once on hearing the news he rode for many da
y on his fine black horse then one evening he reached the palace when the t
ired and dusty traveller explained to the stable boy why he had come they l
aughed in scorn but they had order to obey so they told him to enter it lat
e they said and you wont see the princess till tomorrow the emperor other d
aughter however were soon told of the new arrival he the most handsome of t
hem all exclaimed one of the servant so marika the emperor youngest and pre
ttiest daughter with her sister peeked through window at the sleeping omar
next morning the emperor ordered the newcomer to be led before sarah the co
urt crowded round to watch unlike all the other suitor omar did nothing at
all to amuse the princess he stared at sarah without saying word and she st
ared back with an empty look on her face the two young people stared silent
ly at each other then omar went back to the emperor and said sire give me y
our sceptre and will solve the problem of sarah surprised at such an odd re
quest the emperor followed omar into sarah room the other princess clustere
d round smiling and admiring the handsome young man with deep bow to sarah
omar straightened up and dealt her blow on the head with the sceptre scream
filled the air the emperor threw up hl arm in rage and his daughter fled in
all direction the guard drew their sword then the whole room stopped in ama
zement for out of sarah head which had been chopped off by the blow rolled
broken spring and piece of metal the princess that never smiled wa doll per
fect dolll and nobody had ever been aware of it except omar the only prince
ss that couldnt stop laughing wa marika the emperor glared at her be quiet
he ordered but he too saw the funny side of it for the crafty emperor had b
een making use of sarah the doll a way of guaranteeing himself steady flow
of tax from all his subject and now man more cunning than himself had expos
ed his trick the emperor had sudden thought he would rid himself of the che
eky marika and gain an astute soninlaw able to help him hold onto his kingd
om you should be put to death for this insolence he said but im going to sp
are your life if you marry my youngest daughter of course you wont need to
pay tax smiling at happy marika omar nodded silently down in the depth of h
is mind he wa thinking one day dear fatherinlaw ill be sitting on your impe
rial throne and he wa few year later
```

In [ ]:

```

```

[6]. Converting Numbers, e.g., “1000” to one thousand

In [46]:

```
import re
from num2words import num2words

for i in range(len(files)):

    words = files[i].split()
    for i in range(len(words)):
        if words[i].isnumeric() and len(words[i]) <= 4:
            words[i] = num2words(int(words[i]))
    files[j]=' '.join(words)
```

In [54]:

```
print(files[6])
```

```
manly wager by lucillus dedicated to testosterone there wa pair of warrior
who thought they were so cool bbbain and magnus were their name the king an
d prince of fool now a to which is greater come listen to my tale and will
tell you of the time these mighty warrior failed late one night bet they ma
de very manly boast so many maiden each could bed but who could get the mos
t and so they set out for to prove who wa the biggest prick and just how st
upid they could act and get away with it bain and magnus wanted to show who
wa the best and each man wa determined to win this manly test by fair mean
and by foul many maiden they would lay then prove it all by boasting in ver
y manly way bain went into town now and found likely inn he wa sure the mai
d would swoon a soon a they saw him he preened and pranced and pampered to
show his better side and practiced his sincerity to hide the fact he lied a
nd sure enough the spell he wove had all the lady there dying for the chanc
e to run their finger through his hair please lady take number pretty bain
he then did say for will serve you all upstairs until the break of day so b
ain think he stud now and many maid agree he care not for discretion in fac
t he charged fee all the lady they were waiting to take their turn in bed t
hen boast to one another im his only love they said but magnus think he cle
ver of that he is so sure he followed bain to see how he would all the lady
lure he saw lady that he knew whose jealous husband cruel would kill to kee
p his lovely wife a miser keep jewel so straight away he went to tell this
jealous hulking man of just what bain wa going to do and his wife part in t
he plan and so he thought he could be sure to win their manly bet this sure
ly wa a clever a any man could get now magnus he is lazy a if you didnt kno
w he thought he had it made now and wished to see the show so he went into
the tavern and waited for to see very jealous husband and his victim soonto
be upstairs bain wa grooming he made the lady wait while magnus tried his b
est to hide and leave bain to his fate soon the jealous husband had gathere
d to his side a many friend a he could find to help him take his bride but
magnus wa impatient and quite horny now a well so he slipped out through th
e back to stable by their smell he wa sure that he could ream some very lon
ely horse then be back inside in time to see bain thrashed by manly force b
ut even for old magnus thing sometimes work out well for bain chose for his
first lay the faithless wife from hell she could not wait for foreplay but
jumped upon his steed and started quick to ride him to service her deep nee
d know now what youre thinking how typical it seems for bain to end up with
maid while magnus horse ream but justice it soon entered into this merry tu
ne the husband and his many friend had come and none too soon up the stair
they charged a one and burst into the room then looked bain and saw right t
here his own impending doom and naked a jaybird he took his only chance wen
t leaping out the window without even his pant now bain had not yet finishe
```

```
d his very manly chore his manhood still wa rigid and hard now to ignore bu
t his luck did not desert him for below him now he saw thatched roof coming
quickly made of soft and yielding straw and magnus in the stable had found
horse to pork wa pounding deep into her tail and leaned into his work when
crashing through the rooftop came bain with his stiff spear and found poor
magnus most exposed and fell into his rear mighty squeal of pain and glee w
a heard for mile around and far away some pig got hard just thinking of tha
t sound and so we have sandwich of two men and horse it hard now to imagine
how thing could turn out worse and bain who wa stuck deep now wa trying to
escape but magnus had recovered and started screaming rape out came the jea
lous husband who could not believe his eye followed closely by the lady who
laughed until they cried now if man were able to die of shame alone then su
rely now our hero would be deader than stone but ala in their position an e
mbarrassment from hell they could not defend themselves and their fate now
will tell they were taken to the wood and then tied upside down their cloth
es were burned before their eye and all went back to town you think the sto
ry over but there one more thing to see who had won the wager and the great
est prick would be for a they were ahanging an argument ensued bain said im
the winner and still the coolest dude but magnus he retorted at least finis
hed mine so shut up bain you loser hate it when you whine shindar drinking
song last night went out drinking and met lively crew so we wandered off to
gether to hoist an ale or two we came upon little place where the lady were
most fair and so we all decided that we should tarry there oh were bold and
handsome bastard til the morning anyway when the rising sun will smite our
eye and make u curse the day the lass they all love u for a long a we can p
ay but when our coin run out they will send u on our way so well hoist anot
her ale and well sing another song and well crawl from bar to bar until the
coming of the dawn behind the bar with golden hair there stood beautious la
ss with ruby lip and eye of blue and this exquisite nose well maddog wa foo
lish lad whose mood did not soon pas he lept across that little bar and tri
ed to pinch her nose chorus we grabbed young maddog by the cloak and everyt
hing wa fine until the lass lad showed up with friend or nine will not bore
you with the tale suffice to say we had them beat and then the city watch s
howed up to make the night complete chorus now fezzik he wa not to bright h
ed smashed guardsman head so the magistrate said take him out and hang him
til he dead the lass that maddog tried to pinch appealed to the bench and t
his morning ive been told he got married to the wench chorus his honor look
ed me in the eye and said what should do so looked right back at him said l
et hoist an ale or two so off we wandered into town with guard or two hed b
rung along the way there wa brawl and come the morning he wa hung final cho
rus jester of the shindar one thousand, nine hundred and ninety-two
```

b. Extract TF-IDF features for all documents. You may consider each unique word as a token and compute the frequency of each word. This will give you a TF-IDF vector for each document.

In [53]:

```
#As the documents have already been pre-processed we will be utilizing it

from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer()

# Compute TF-IDF vectors for the query sentences
query_matrix = vectorizer.fit_transform(files)
print(query_matrix)
```

```
  (0, 33756)    0.045205679582165655
  (0, 21081)    0.04820335940497572
  (0, 33087)    0.020599996570631954
  (0, 22058)    0.02166859170714967
  (0, 34102)    0.023226322927822625
  (0, 9205)     0.02735671472337309
  (0, 3301)     0.015371042255792717
  (0, 15017)    0.015371042255792717
  (0, 35852)    0.03207102181015885
  (0, 30074)    0.04612021880206683
  (0, 20912)    0.01563221112891138
  (0, 17361)    0.0507529174506282
  (0, 40463)    0.037284342086654064
  (0, 35072)    0.031537748211055795
  (0, 15501)    0.030521466193142607
  (0, 16261)    0.017570970663093346
  (0, 35752)    0.028935170855799774
  (0, 37148)    0.015765068913288324
  (0, 34259)    0.02958824235629397
  (0, 27203)    0.027923391553498395
  (0, 9025)     0.02981450006103463
  (0, 25862)    0.025970089174123766
  (0, 32551)    0.06659221874891681
  (0, 18126)    0.011609929240013174
  (0, 32192)    0.020397767561849703
  :	:
  (248, 36077)  0.03187618866204422
  (248, 19120)  0.019339098897722434
  (248, 25177)  0.012382952109133623
  (248, 31209)  0.023871022595308544
  (248, 27122)  0.012584899297691046
  (248, 1806)   0.011403351007445826
  (248, 39385)  0.03415283034264758
  (248, 3934)   0.008502347199707527
  (248, 12334)  0.009062968132857677
  (248, 11768)  0.013386307495688107
  (248, 13504)  0.00890851913736004
  (248, 9221)   0.012936581884507775
  (248, 20694)  0.009678456343376695
  (248, 9147)   0.0399380349050666
  (248, 15714)  0.014215594902265326
  (248, 12779)  0.01828327533210323
  (248, 33026)  0.010757299069351588
  (248, 17469)  0.011403351007445826
  (248, 2215)   0.14533021758911963
  (248, 39170)  0.008720640089723852
  (248, 5745)   0.017441280179447703
  (248, 40122)  0.016107709516126542
  (248, 15673)  0.047999253549353235
  (248, 36089)  0.013636482508853246
  (248, 38114)  0.00926151711360725
```

In [ ]:

c. (10 pts) Compute the pairwise cosine similarities of all documents, save in a matrix (.txt) and submit together with your PDF solution.

In [61]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

vectorizer = TfidfVectorizer()

# create a TfidfVectorizer object and then doing fit and transform
tfidf_features = vectorizer.fit_transform(files)

# compute the pairwise cosine similarities
cos_sim_matrix = cosine_similarity(tfidf_features)

#cosine similarity matrix to a text file
np.savetxt('/Users/varun/Desktop/cosine_similarity_matrix.txt', cos_sim_mat
```

In [ ]:

d. (10 pts) Among these documents, find the top-10 retrieval results (using file names) for the following two sentences. The retrieval can be done by computing the cosine similarity between the “query” and “reference” documents in the data.

[1]. Once upon a time . . . there were three little pigs, who left their mummy and daddy to see the world.

[2]. There once lived a poor tailor, who had a son called Aladdin, a careless, idle boy who would do nothing but play all day long in the streets with little idle boys like himself.

In [62]:

```
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer


translator = str.maketrans('', '', string.punctuation)
```

```
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()



files1 = ['Once upon a time . . . there were three little pigs, who left tl
    'There once lived a poor tailor, who had a son called Aladdin, a carele
]

# Preprocess the query sentences
query_docs = []

for i in range(len(files1)):

    files1[i] = files1[i].lower()
    #print(files1[i])

    words = files1[i].split()

    filtered_words = [word for word in words if word.lower() not in stop_wo

    files1[i]= ' '.join(filtered_words)


    files1[i] = files1[i].translate(translator)


    words1 = word_tokenize(files1[i])

    lemmatized_words = [lemmatizer.lemmatize(word) for word in words1]

    files1[i] = " ".join(lemmatized_words)

    #print(files1[i])


    words = files1[i].split()
    for j in range(len(words)):
        if words[i].isnumeric() and len(words[j]) <= 4:
            words[j] = num2words(int(words[j]))
    files[j]=' '.join(words)
    query_docs.append(files[j])


#print(query_docs)

# Compute TF-IDF vectors for the query sentences
query_matrix = vectorizer.transform(query_docs)

# Compute cosine similarities between query sentences and reference documen
cos_sim_query_ref = cosine_similarity(query_matrix, tfidf_features)
cos_sim_query_ref
```

Out[62]:

```
array([[1.26677149e-02, 3.19602241e-02, 4.57985593e-03, 1.02859282e-02,
        2.12059853e-02, 3.11345465e-02, 2.09728042e-02, 9.81456350e-03,
        2.53749508e-02, 1.00000000e+00, 1.36992474e-02, 9.89078235e-03,
        9.68158126e-03, 7.38458016e-03, 2.28880421e-02, 3.51488864e-02,
        2.21686902e-02, 8.03429943e-03, 1.42888360e-02, 1.71328926e-02,
        1.98687507e-02, 1.98657732e-02, 6.53157521e-03, 6.90207574e-02,
        1.14257465e-02, 1.19057578e-02, 1.89212388e-02, 1.00959342e-02,
        1.05827870e-02, 2.12232669e-02, 8.81376477e-03, 6.73455754e-03,
        2.46522097e-02, 1.58601954e-02, 7.24313862e-03, 9.82308775e-03,
        4.68350836e-03, 1.36972458e-02, 2.11715485e-02, 8.88118446e-03,
        7.32389438e-03, 3.91253712e-03, 1.16419560e-02, 1.33077825e-02,
        1.43560286e-02, 1.86218632e-02, 1.75765444e-02, 1.06675681e-02,
        9.61404718e-03, 1.29180607e-02, 1.83514693e-02, 3.28371346e-02,
        1.41578410e-02, 7.88061648e-03, 2.61310345e-02, 6.58773266e-03,
        1.78277454e-02, 1.04340240e-02, 7.98311438e-03, 1.07304004e-02,
        1.01198611e-02, 2.00016595e-02, 8.44635502e-03, 8.50946565e-03,
        1.16365814e-02, 1.77116504e-02, 6.01180849e-03, 1.05086173e-02,
        2.24018648e-02, 1.43482232e-02, 1.14865484e-02, 9.75285461e-03,
        9.02929736e-03, 1.25868513e-02, 6.33663612e-03, 1.90960763e-02,
        1.08975361e-02, 6.40835497e-03, 5.14736115e-03, 8.06609882e-03,
        2.69690996e-02, 1.30465040e-02, 6.26302025e-03, 5.67414723e-03,
        2.99723800e-02, 1.74505900e-02, 1.05936682e-02, 2.17010501e-02,
        4.46839695e-03, 0.00000000e+00, 2.04106911e-02, 1.80405107e-02,
        5.72238983e-03, 1.06978614e-02, 1.41726752e-02, 2.02337640e-02,
        4.11368631e-03, 1.27980862e-02, 1.21703206e-02, 8.94801993e-03,
        7.54817666e-03, 2.91339331e-02, 2.24590096e-02, 2.20369313e-02,
        2.02348506e-02, 5.34062008e-03, 1.37622742e-02, 1.97967839e-02,
        4.26653916e-03, 1.66179571e-02, 1.67826583e-02, 5.29317042e-03,
        1.05290639e-02, 2.50043494e-03, 1.15872147e-02, 1.42881173e-02,
        2.81727327e-02, 5.05996026e-03, 6.82814529e-03, 1.50221777e-02,
        1.53412146e-02, 1.57622008e-02, 3.46162959e-03, 1.68577778e-02,
        1.62566327e-02, 1.58111228e-02, 1.45616392e-02, 1.98796961e-02,
        3.53781746e-02, 6.26118965e-03, 6.60035695e-03, 2.11258976e-02,
        1.26084787e-02, 4.68718314e-02, 6.36201252e-03, 1.48601043e-02,
        4.68666904e-03, 1.73543839e-02, 3.81136456e-02, 1.65083960e-02,
        1.04914101e-02, 3.43568205e-02, 3.24450776e-01, 1.24701140e-02,
        2.43412127e-02, 1.44676631e-02, 1.35890304e-02, 3.46722682e-02,
        5.62854849e-03, 1.10849253e-02, 2.03256553e-02, 0.00000000e+00,
        7.71533624e-03, 1.02976182e-02, 0.00000000e+00, 1.01884717e-02,
        2.50776688e-02, 1.41728254e-02, 1.09245723e-02, 2.63721044e-02,
        1.39812076e-02, 1.59690707e-02, 2.44159995e-02, 2.00068242e-02,
        1.86258452e-02, 0.00000000e+00, 4.03580388e-03, 1.15778996e-02,
        6.04972724e-03, 8.50315990e-03, 1.57599389e-02, 1.06204731e-02,
        7.49028539e-03, 2.47327691e-03, 9.53168919e-03, 1.26775379e-02,
        2.72318707e-02, 1.66902212e-02, 9.39398683e-03, 2.32856805e-02,
        2.04612315e-02, 2.05731938e-02, 1.30797766e-02, 9.94227885e-03,
        2.87010998e-02, 1.52874459e-02, 2.64437277e-02, 1.97695202e-02,
        1.85331747e-02, 1.86774429e-02, 6.98625828e-03, 2.43650061e-02,
        1.22999636e-02, 6.96990228e-03, 1.00609295e-02, 3.32335125e-03,
        4.11644934e-03, 7.67477129e-03, 1.97172586e-02, 2.39179738e-02,
        1.57008549e-02, 1.54460413e-02, 2.62332772e-02, 2.88863942e-02,
        1.87371199e-02, 9.88358592e-04, 2.74786531e-02, 1.23584247e-02,
        2.73078183e-02, 4.80844860e-03, 7.20230336e-03, 1.01371102e-02,
        2.38100425e-02, 4.60154047e-03, 1.26004367e-02, 1.60013034e-02,
        1.59328969e-02, 2.45665848e-02, 1.60037953e-02, 1.32979248e-02,
        7.21005878e-03, 1.32026834e-02, 2.85318562e-02, 1.53293739e-02,
        3.76672090e-02, 3.62774668e-02, 5.94851117e-03, 1.91328029e-02,
        2.46222913e-02, 1.19589187e-02, 5.31853734e-03, 1.20454631e-02,
        1.64504717e-02, 1.68487206e-02, 1.16272088e-02, 2.40587848e-02,
        2.53817989e-02, 5.89690672e-03, 2.42159654e-02, 1.82968948e-02,
        1.65395334e-02, 1.67200623e-02, 5.27356144e-02, 4.11060519e-02,
        2.94517517e-02, 4.40765438e-03, 8.81181687e-03, 1.32462500e-02,
        6.26302025e-03],
       [1.48430925e-02, 2.43433672e-02, 3.60352507e-03, 1.04090787e-02,
        1.78178849e-02, 2.56628306e-02, 8.58092599e-03, 6.51177911e-03,
        1.82267840e-02, 1.71328926e-02, 2.08041008e-02, 2.50910991e-02,
        1.00596278e-02, 1.13559666e-02, 2.14362857e-02, 2.01937253e-02,
        4.51304865e-03, 2.85374714e-02, 1.36936981e-02, 1.00000000e+00,
        1.27093794e-02, 1.39033106e-02, 4.36296401e-03, 2.05096572e-02,
        6.99951147e-03, 2.49409586e-02, 1.40385629e-02, 3.53227237e-03,
        9.25743895e-03, 1.81570336e-02, 4.65476592e-03, 1.75840663e-02,
        1.73205213e-02, 1.68000585e-02, 1.88016447e-02, 9.48822113e-03,
        4.23655053e-03, 1.13730814e-02, 2.23529809e-02, 1.10282595e-02,
        2.58158056e-02, 1.12445209e-02, 1.00309925e-02, 7.58185192e-03,
        1.92597279e-02, 1.17482160e-02, 2.02463656e-02, 9.12760600e-03,
        1.85436238e-02, 1.54831567e-02, 3.29103723e-02, 1.40513805e-02,
        1.04043866e-02, 1.07427781e-02, 1.62030429e-02, 9.13026967e-03,
        1.44100907e-02, 2.88878427e-03, 1.47146940e-02, 1.67300795e-02,
        1.29344012e-02, 2.11176686e-02, 9.00746583e-03, 1.01189848e-02,
        8.37679348e-03, 2.17082124e-02, 3.32512744e-03, 1.14199244e-02,
        2.09155146e-02, 6.07953729e-03, 1.10612022e-02, 8.53286243e-03,
        7.82472460e-03, 3.73190042e-02, 1.21783952e-02, 9.91795945e-03,
        2.64369359e-02, 5.70416290e-03, 3.06596263e-03, 1.06031500e-02,
        1.70402752e-02, 9.36101717e-03, 1.73658043e-02, 8.99517722e-03,
        2.35237254e-02, 2.51918269e-02, 2.66919312e-02, 9.69774504e-03,
        9.55777231e-03, 4.93198482e-03, 1.63497000e-02, 2.12213018e-02,
        2.01928668e-02, 1.87234403e-02, 6.99621900e-03, 2.93413355e-03,
        7.90705749e-03, 1.49103026e-02, 8.38937832e-03, 1.25844619e-03,
        1.62528163e-02, 1.69663974e-02, 1.56372548e-02, 1.45027388e-02,
        1.71335525e-02, 5.28735031e-03, 7.78423857e-03, 4.69837238e-03,
        0.00000000e+00, 1.25384186e-02, 9.08583519e-03, 4.58398705e-03,
        2.09736069e-02, 8.99295832e-03, 1.14090126e-02, 1.94007887e-02,
        5.28350379e-03, 7.67998418e-03, 3.99163528e-03, 1.64277929e-02,
        7.28808223e-03, 2.03651135e-02, 1.37321113e-02, 1.18722652e-02,
        1.61897367e-02, 9.51115689e-03, 9.72528566e-03, 1.13164545e-02,
        6.79721923e-02, 2.69901386e-01, 5.93261781e-03, 1.62830442e-02,
        1.24470995e-02, 2.41848825e-02, 1.16994906e-02, 1.31507393e-02,
        7.03399904e-03, 1.66484771e-02, 3.73192335e-02, 2.65516065e-02,
        1.60286955e-02, 1.79241933e-02, 3.36876794e-02, 1.27175314e-02,
        2.76417032e-02, 1.54406510e-02, 2.53840243e-01, 1.73477051e-02,
        1.40416217e-02, 1.33149477e-02, 1.15733315e-02, 1.86329984e-02,
        6.87293458e-03, 1.48849060e-02, 3.22780244e-02, 2.23961391e-03,
        1.23875470e-02, 2.88966112e-02, 0.00000000e+00, 2.09828037e-02,
        1.13062001e-02, 2.03210544e-02, 3.05133461e-02, 6.73985032e-03,
        1.29449798e-02, 0.00000000e+00, 2.25370432e-03, 1.00798742e-02,
        1.89926447e-02, 1.20029237e-02, 6.75867735e-03, 1.17998784e-02,
        2.77715046e-03, 0.00000000e+00, 1.37728346e-02, 5.80463260e-03,
        2.57266416e-02, 1.04605624e-02, 1.43109362e-02, 1.63838137e-02,
        1.81833456e-02, 1.28467445e-02, 1.74658599e-02, 4.40324984e-03,
        2.91970262e-02, 1.22835564e-02, 1.47209090e-02, 4.06242111e-03,
        1.57575077e-02, 2.38177553e-02, 1.67792509e-02, 1.58726604e-02,
        1.32296961e-02, 1.58012742e-02, 3.53997100e-02, 9.69784440e-03,
        3.65710784e-03, 1.18384072e-02, 2.23756174e-02, 1.14117332e-02,
        9.81927360e-03, 1.35116611e-02, 8.58362849e-03, 1.81757986e-02,
        1.75036635e-02, 3.62003046e-03, 8.67357135e-03, 3.43757228e-03,
        1.96514222e-02, 8.49187897e-03, 1.48991364e-02, 9.38463658e-03,
        1.65990854e-02, 5.08384277e-03, 3.56903772e-02, 2.29720453e-02,
        2.33509379e-02, 1.70627387e-02, 1.71912805e-02, 1.05898023e-02,
        1.43778981e-02, 1.32394418e-02, 1.64494223e-02, 1.73226417e-02,
        3.96520472e-02, 1.42274574e-02, 4.38812385e-03, 7.68333197e-03,
        1.25495888e-02, 2.88747260e-02, 1.64562055e-02, 8.18735209e-03,
        1.93854228e-02, 1.93665457e-02, 8.38458591e-03, 1.34416352e-02,
        1.28932892e-02, 9.17420034e-03, 2.11635198e-02, 1.93635240e-02,
        1.34045433e-02, 1.69676131e-02, 1.31519579e-02, 2.12926409e-02,
        2.37892531e-02, 5.63551927e-03, 3.30264221e-02, 1.42089934e-02,
        1.73658043e-02]])
```

In [ ]: