Report on
Comparing Corpora with Corpus Statistics

Task 1- Choosing the data

Documents which I have selected for this assignment are from Gutenberg project namely:

- **Speeches & Letters of Abraham Lincoln,1832-1865**
- **Mark Twain's Letters 1901-1906(Volume V)**
- **Mark Twain's Speeches**

I merged Mark Twain's Letters (5th volume) and Speeches into a single file and then I did a comparative study with Speeches & Letters of Abraham Lincoln,1832-1865.

Reason for selecting above documents was to compare writing styles of two different personalities-

- One being 16th president of USA (also a lawyer and politician) and other being a well-known philosopher and writer.
- Speeches & Letters of Abraham Lincoln were limited to political, administrative and legal matters. Moreover, Abraham Lincoln speeches were prepared with care whereas addressing of Mark Twain ranged in variety. He spoke what was in his mind untrammeled by literary conventions.
- Audience of Abraham Lincoln addressing was not just limited to Congress or office holder's but in some case, it was entire nation whereas Mark Twain's addressing subjected to individual's or social gathering like dinner parties.
- Also, these speeches and letter belong to different times-Abraham Lincoln writings belong to mid-19th century from 1832-1865 whereas large collection of Mark Twain writings belong to early 20th century.

Above reasoning is based on reading preface/introduction of these documents. I would try to examine whether this holds true based on frequency distribution analysis or not?

Note-Huge collection of Mark Twain letters archived online are divided into six parts. Due to limited time frame, I restrained my scope of study to 5th volume of Mark Twain's Letters.

Description of Task 1:

First, I obtained URL's to all three ASCII text files from Project Gutenberg.
http://www.gutenberg.org/cache/epub/14721/pg14721.txt
http://www.gutenberg.org/files/3188/3188-0.txt
http://www.gutenberg.org/files/3197/3197-0.txt

Then I downloaded all three files from Project Gutenberg by using following code in Jupyter notebook:

```
import nltk
from urllib import request
url1 ="http://www.gutenberg.org/cache/epub/14721/pg14721.txt"
response1 = request.urlopen(url1)
file1 = response1.read().decode('utf8')
print(file1[:129])
```

Similarly, I downloaded other two documents and store them in variable file1 and file2.

As part of pre-processing I did following steps:

1. **Getting actual content-**
   Actual content of Project Guternberg books are delimited by the phrases "START OF THIS PROJECT GUTENBERGBOOK" and "END OF THIS PROJECT GUTENBERGBOOK" respectively. Content in between these delimeters is the actual content of the ebook. In order to process and compare the actual content of above documents more efficiently, we will remove extra data.

   ```
   import re
   start = re.search(r"START OF THE PROJECT GUTENBERG EBOOK", file1).start()
   stop = re.search(r"END OF THE PROJECT GUTENBERG EBOOK", file1).end()
   AbrahamLincoln= file1[start:stop]
   AbrahamLincoln
   ```

   Note-Depending on url used, delimeter phrase can vary- "START OF THE PROJECT GUTENBERGBOOK" and "END OF THE PROJECT GUTENBERGBOOK"

2. **Merging file 2 and 3 using string concatenation function and storing the output in variable MarkTwain**
   ```
   MarkTwain=text2 + text3
   MarkTwain
   ```

3. **Tokenization and Changing to lower case(Normalization)**
   I tried three different types of tokenizers-
   - **Countvectorizer**
     ```
     from sklearn.feature_extraction.text import CountVectorizer
     altokens=CountVectorizer().build_tokenizer()(AbrahamLincoln)
     alwords = [w.lower( ) for w in altokens]
     print(len(altokens))
     alwords
     ```

     This divided sentences into words similar to wordpunct tokenizer and removed all single character words like 'can't' resulted into 'can' and 'U.S.A.' was completed removed.

   - **word_punct tokenizer**
     ```
     punctaltokens = nltk.wordpunct_tokenize(AbrahamLincoln)
     punctalwords = [w.lower( ) for w in punctaltokens]
     print(len(punctalwords))
     ```

     This divided word like 'U.S.A.' into six words- 'u','.','s','.','a','.' and 'can't' into three words- 'can',''','t'

   - **word tokenizer**
     ```
     waltokens = nltk.word_tokenize(AbrahamLincoln)
     walwords = [w.lower( ) for w in waltokens]
     print(len(walwords))
     ```
     This approach produced some unnecessary words but unique words like U.S.A were not removed. In order to remove unnecessary words like non-alphanumeric words and stoplist, I used following code:

```
def alpha_filter(w):
  # pattern to match word of non-alphabetical characters
  pattern = re.compile('^[^a-z]+$')
  if (pattern.match(w)):
    return True
  else:
    return False

stopwords = nltk.corpus.stopwords.words('english')

walwords = [w for w in walwords if not alpha_filter(w)]
print(len(walwords))

stoppedwalwords = [w for w in walwords if not w in stopwords]
print(len(stoppedwalwords))
stoppedwalwords
```

Note-I normalized the text to lowercase so that the distinction between The and the is ignored. I converted all tokens into lower case as many functions of nltk are case-sensitive.

4. **Stemming and lemmatization**
   I tried three stemmers-Lancaster, Porter and Snowball stemmer. I also examined document using WordNet lemmatizer.

```
porter = nltk.PorterStemmer()
lancaster = nltk.LancasterStemmer()
wnl=nltk.WordNetLemmatizer()
from nltk.stem import SnowballStemmer
snowball_stemmer = SnowballStemmer('english')

crimePstem = [porter.stem(t) for t in stoppedwalwords]
print('Porter\n', crimePstem[:200])

crimeLstem = [lancaster.stem(t) for t in stoppedwalwords]
print('Lancaster\n', crimeLstem[:200])

crimeLemma = [wnl.lemmatize(t) for t in stoppedwalwords]
print('WordNet Lemmatizer\n', crimeLemma[:200])

crimeSnstem = [snowball_stemmer.stem(t) for t in stoppedwalwords]
print('Snowball Stemmer\n', crimePstem[:200])
```

I found that Lancaster stemmer was severe on some words like event and ever resulted into ev whereas Snowball stemmer hardly changed any word compared to other two stemmers.
The WordNet lemmatizer only removes affixes if the resulting word is in its dictionary like lying remains same instead of changing to lie. So, I decided to use combination of WordNet lemmatization and Porter stemming.

Note -Based on results obtained for task 2, I updated stop words list on order to perform some meaningful analysis in task 3.

Task 2 Frequency Distribution -
List the top 50 words by frequency (normalized by the length of the document):
Note- I divided the word count by total length of document into normalize word frequency distribution.

For Abraham Lincoln document-

one	0.000629176765760394
us	0.00044332762879732376
men	0.00042396834369700395
right	0.00038912163051642827
man	0.00038524977349636434
may	0.00036976234541610846
constitution	0.00034265934627566075
union	0.00032136413266530895
state	0.000315556347135213
government	0.00030587670458505307
great	0.0002632862773643495
free	0.00025747849183425356
made	0.0002536066348141896
time	0.00024586292077406166
question	0.00023811920673393373
let	0.00023811920673393373
said	0.00022650363567374183
think	0.00022456770716370987
ever	0.00021875992163361392
way	0.00021682399312358193
slave	0.00021682399312358193
new	0.00021488806461354996
never	0.00021101620759348598
much	0.00021101620759348598
wrong	0.00021101620759348598
without	0.00020520842206339003
make	0.00020520842206339003
know	0.00019552877951323013
law	0.00019552877951323013
lincoln	0.0001781054229229423
even	0.0001781054229229423
good	0.0001781054229229423
well	0.00017616949441291032
take	0.00016261799484268644
first	0.00015874613782262248
power	0.00015874613782262248
yet	0.00015874613782262248
whether	0.00015487428080255853
war	0.00015100242378249457
speech	0.0001471305667624306
come	0.00014519463825239862
congress	0.00014325870974236663

nation	0.00014132278123233464
case	0.00014132278123233464
thing	0.00014132278123233464
better	0.00013745092421227069
get	0.0001355149957022387
among	0.00013357906719220673
subject	0.00013164313868217474
public	0.00013164313868217474

For Mark Twain documents-

year	0.0005597285176054307
know	0.00047112827486889275
time	0.00046269015651303195
man	0.0004176868586151079
go	0.0004134677994371775
thing	0.00038815344436959523
day	0.00037549626683580406
like	0.00035018191176822177
good	0.0003262739097599496
new	0.0003220548505820192
never	0.00031783579140408884
great	0.00031502308528546857
clemen	0.0003093976730482281
well	0.00028548967103995590
come	0.0002770515526840951
old	0.0002742388465654749
think	0.0002714261404468546
way	0.00026861343432823436
mark	0.0002545499037351331
alway	0.0002503308445572027
life	0.00023204825478617105
all	0.00023064190172686092
two	0.00022220378337100015
first	0.0002151720180744495
word	0.00019970213442203812
peopl	0.00019970213442203812
may	0.00019970213442203812
look	0.00019126401606617735
littl	0.00018563860382893684
came	0.00018282589771031658
york	0.00018001319159169634
long	0.00018001319159169634
much	0.00017720048547307607
want	0.00017720048547307607
let	0.00017579413241376594
ever	0.00017579413241376594
mani	0.00017579413241376594
ask	0.00017579413241376594
world	0.0001729814262951457
tell	0.00017157507323583556

ago	0.0001687623671172153
friend	0.0001659496609985950 6
read	0.0001617306018206646 6
work	0.0001617306018206646 6
put	0.0001603242487613545 5
use	0.0001589178957020444 2
anyth	0.0001575115426427342 9
place	0.0001575115426427342 9
give	0.0001561051895834241 5
everi	0.0001546988365241140 5

List the top 50 bigrams by frequencies:

For Abraham Lincoln document-

(('let', 'us'), 0.002553191489361702)
(('dred', 'scott'), 0.0023028785982478098)
(('men', 'equal'), 0.0012015018773466834)
(('nebraska', 'bill'), 0.001051314142678348)
(('slave', 'state'), 0.000951188986232791)
(('abraham', 'lincoln'), 0.0009011264080100125)
(('right', 'wrong'), 0.000851063829787234)
(('among', 'us'), 0.0008010012515644555)
(('one', 'man'), 0.0008010012515644555)
(('public', 'mind'), 0.0008010012515644555)
(('white', 'men'), 0.0006508135168961202)
(('mr.', 'lincoln'), 0.0006007509386733417)
(('question', 'whether'), 0.0006007509386733417)
(('government', 'live'), 0.0005506883604505632)
(('new', 'york'), 0.0005506883604505632)
(('state', 'constitution'), 0.0005506883604505632)
(('white', 'man'), 0.0005506883604505632)
(('a.', 'lincoln'), 0.0005006257822277847)
(('constitution', 'right'), 0.0005006257822277847)
(('lecompton', 'constitution'), 0.0005006257822277847)
(('one', 'thing'), 0.0005006257822277847)
(('save', 'union'), 0.0005006257822277847)
(('support', 'constitution'), 0.0005006257822277847)
(('dear', 'sir'), 0.0004505632040050062 6)
(('free', 'state'), 0.0004505632040050062 6)
(('go', 'back'), 0.0004505632040050062 6)
(('new', 'jersey'), 0.0004505632040050062 6)
(('public', 'sentiment'), 0.0004505632040050062 6)
(('springfield', 'ill.'), 0.0004505632040050062 6)
(('_from', 'address'), 0.0004005006257822277 7)
(('_from', 'letter'), 0.0004005006257822277 7)
(('amongst', 'us'), 0.0004005006257822277 7)
(('care', 'whether'), 0.0004005006257822277 7)
(('constitution', 'union'), 0.0004005006257822277 7)
(('first', 'place'), 0.0004005006257822277 7)
(('good', 'men'), 0.0004005006257822277 7)
(('man', 'may'), 0.0004005006257822277 7)

(('mind', 'rest'), 0.00040050062578222777)
(('north', 'south'), 0.00040050062578222777)
(('one', 'thousand'), 0.00040050062578222777)
(('question', 'court'), 0.00040050062578222777)
(('right', 'constitution'), 0.00040050062578222777)
(('state', 'union'), 0.00040050062578222777)
(('civil', 'war'), 0.00035043804755944933)
(('come', 'end'), 0.00035043804755944933)
(('great', 'britain'), 0.00035043804755944933)
(('half', 'free'), 0.00035043804755944933)
(('half', 'slave'), 0.00035043804755944933)
(('men', 'free'), 0.00035043804755944933)
(('much', 'better'), 0.00035043804755944933)

For Mark Twain documents-
(('mr', 'clemen'), 0.0027753976828223375)
(('new', 'york'), 0.0022724732704988266)
(('year', 'ago'), 0.0014715195767984203)
(('year', 'old'), 0.0007078195432701263)
(('new', 'england'), 0.0005588049025816787)
(('fourth', 'juli'), 0.0005215512424095668)
(('let', 'u'), 0.0005029244123235108)
(('five', 'year'), 0.000428417091979287)
(('dear', 'mr'), 0.00040979026189323 1)
(('first', 'time'), 0.00040979026189323 1)
(('mr', 'roger'), 0.00040979026189323 1)
(('seven', 'year'), 0.00040979026189323 1)
(('two', 'year'), 0.00039116343180717507)
(('mr', 'carnegi'), 0.0003725366017211191)
(('three', 'year'), 0.00035390977163506315)
(('twichel', 'hartford'), 0.00033528294154900717)
(('two', 'three'), 0.00033528294154900717)
(('unit', 'state'), 0.00033528294154900717)
(('joan', 'arc'), 0.00031665611146295124)
(('twenti', 'five'), 0.00031665611146295124)
(('dear', 'joe'), 0.00029802928137689526)
(('introduc', 'mr'), 0.00029802928137689526)
(('long', 'ago'), 0.00029802928137689526)
(('mani', 'year'), 0.00029802928137689526)
(('mr', 'choat'), 0.00029802928137689526)
(('dinner', 'given'), 0.00027940245129083933)
(('good', 'deal'), 0.00027940245129083933)
(('ladi', 'gentleman'), 0.00027940245129083933)
(('twenti', 'year'), 0.00027940245129083933)
(('know', 'anyth'), 0.00026077562120478334)
(('robert', 'fulton'), 0.00026077562120478334)
(('fifti', 'year'), 0.00024214879111872734)
(('four', 'hundr'), 0.00024214879111872734)
(('friend', 'mine'), 0.00024214879111872734)
(('human', 'race'), 0.00024214879111872734)

(('long', 'time'), 0.0002421487911187274)
(('mani', 'time'), 0.0002421487911187274)
(('old', 'friend'), 0.0002421487911187274)
(('rev', 'twichel'), 0.0002421487911187274)
(('san', 'francisco'), 0.0002421487911187274)
(('sincer', 'clemen'), 0.0002421487911187274)
(('tell', 'truth'), 0.0002421487911187274)
(('thirti', 'year'), 0.0002421487911187274)
(('address', 'dinner'), 0.00022352196103267146)
(('citi', 'new'), 0.00022352196103267146)
(('forti', 'year'), 0.00022352196103267146)
(('good', 'mani'), 0.00022352196103267146)
(('per', 'cent'), 0.00022352196103267146)
(('two', 'hundr'), 0.00022352196103267146)
(('anybodi', 'els'), 0.0002048951309466155)

List the top 50 bigrams by their Mutual Information scores (using min frequency 5):
For Abraham Lincoln document-
(('h.', 'herndon'), 10.679919878518422)
(('william', 'h.'), 10.457527457181975)
(('ill.', 'oct.'), 9.679919878518419)
(('john', 'brown'), 9.387138129290573)
(('dred', 'scott'), 8.741320423182563)
(('wo', "n't"), 8.71029352756194)
(('sworn', 'support'), 8.691992710818994)
(('herndon', 'washington'), 8.50999487707611)
(('springfield', 'ill.'), 8.24251456621112)
(('life', 'pursuit'), 8.235135035845527)
(('nebraska', 'bill'), 8.1434722060659)
(('rest', 'belief'), 8.106453016635095)
(('last', 'night'), 8.065210034403213)
(('mind', 'rest'), 7.388365432674576)
(('new', 'jersey'), 7.31846341944442)
(('new', 'york'), 7.229458413385675)
(('repeal', 'missouri'), 7.207793187032758)
(('great', 'britain'), 7.177419537989239)
(('civil', 'war'), 7.086395364293841)
(('due', 'regard'), 6.946565537904592)
(('washington', 'jefferson'), 6.83192297196347)
(('member', 'congress'), 6.676917390356897)
(('lecompton', 'constitution'), 6.6597733054066435)
(('without', 'consent'), 6.643877128592891)
(('public', 'mind'), 6.6228306863116)
(("n't", 'care'), 6.56345213923267)
(('good', 'faith'), 6.255893596012321)
(('public', 'sentiment'), 6.177419537989236)
(('north', 'south'), 6.1557046758843015)
(('care', 'whether'), 5.98875797396534)
(('amongst', 'us'), 5.966246972505335)
(('spread', 'place'), 5.955633418422959)

(('two', 'ago'), 5.898560164993761)
(('go', 'back'), 5.845156234167103)
(('washington', 'march'), 5.744460130713129)
(('government', 'live'), 5.613178327642167)
(('idea', 'wrong'), 5.570698635906426)
(('speech', 'springfield'), 5.538907568991348)
(('come', 'end'), 5.521490515913939)
(('day', 'year'), 5.431992365074835)
(('black', 'white'), 5.361603037183436)
(('sacred', 'right'), 5.350796282226854)
(('half', 'slave'), 5.310686068852704)
(('slave', 'half'), 5.310686068852704)
(('_from', 'speech'), 5.251420119433012)
(('brave', 'men'), 5.227060913804609)
(('men', 'equal'), 5.217076825231988)
(('let', 'us'), 5.155589427774887)
(('half', 'free'), 5.062758555409117)
(('support', 'constitution'), 5.031742082793601)

For Mark Twain documents-
(('mortim', 'durand'), 13.381840327682625)
(('oliv', 'wendel'), 13.381840327682625)
(('waldorf', 'astoria'), 13.381840327682625)
(('gutenberg', 'ebook'), 12.89641350051238)
(('madam', 'bernhardt'), 12.89641350051238)
(('warren', 'hast'), 12.89641350051238)
(('project', 'gutenberg'), 12.674021079175933)
(('quarri', 'farm'), 12.533843421127674)
(('anti', 'doughnut'), 12.53384342112767)
(('helen', 'keller'), 12.118805921848827)
(('plymouth', 'rock'), 12.003328704428894)
(('san', 'francisco'), 12.003328704428892)
(('eng', 'nye'), 11.981302398098894)
(('fifth', 'ave'), 11.855771516015036)
(('di', 'quarto'), 11.796877826961463)
(('joan', 'arc'), 11.533843421127674)
(('riverdal', 'hudson'), 11.410986673342139)
(('wendel', 'holm'), 11.311450999791226)
(('fifth', 'avenu'), 11.244336803932686)
(('tom', 'sawyer'), 11.18020646651297)
(('villa', 'di'), 11.173947476041292)
(('fred', 'grant'), 11.118805921848828)
(('tom', 'reed'), 11.02820337306792)
(('seventieth', 'birthday'), 10.932939376537496)
(('robert', 'fulton'), 10.900098547320235)
(('adam', 'diari'), 10.865825180678959)
(('eve', 'diari'), 10.865825180678959)
(('declar', 'independ'), 10.833403702986578)
(('st', 'loui'), 10.566264898820052)
(('hair', 'pin'), 10.533843421127674)

(('havana', 'cigar'), 10.486537706349317)
(('innoc', 'abroad'), 10.22572112576534)
(('plug', 'hat'), 10.195973782371288)
(('georg', 'harvey'), 10.128859586512752)
(('colonel', 'harvey'), 10.106833280182752)
(('quarto', 'florenc'), 10.05991223279526)
(('carnegi', 'hall'), 10.031343080598488)
(('accid', 'insur'), 9.966802828403779)
(('john', 'hay'), 9.961076217376819)
(('fourth', 'juli'), 9.79601933539142)
(('simplifi', 'spell'), 9.753809105069578)
(('human', 'be'), 9.726488499070069)
(('admir', 'harrington'), 9.703768422569983)
(('appear', 'committe'), 9.685846514572724)
(('green', 'watermelon'), 9.640758625044183)
(('per', 'cent'), 9.39694722007283)
(('unit', 'state'), 9.393913159983198)
(('bench', 'show'), 9.391893992346542)
(('rev', 'twichel'), 9.27363757790565)
(('disappear', 'literatur'), 9.244336803932686)

Description of Task 2:

a) Briefly state why you chose the processing options that you did.
Already described in Task 1
Word-tokenizer-> Normalized text(lower case)->alpha filter->stopwords->Porter stemmer->WordNet Lemmatizer-> generated frequency distribution-> updated stopwords->generated frequency distribution

Note-For Mark Twain documents I used countvectorizer instead of word tokenizer as bigram list and PMI contained lot of single character words.

b) Are there any problems with the word or bigram lists that you found? Could you get a better list of bigrams? How are the top 50 bigrams by frequency different from the top 50 bigrams scored by Mutual Information?

Some of the bigrams (raw frequency) are generally (did not render any useful information) associ ated to my document topic like (('_from', 'address'), 0.00040050062578222777), (('_from', 'letter') , 0.00040050062578222777).
Also, I found some redundancy in this bigram list like ('a.', 'lincoln'), ('mr.', 'lincoln'), ('abraham', ' lincoln'). PMI version did not generate any such redundancy.
I tried to generate more efficient list by adding above words in stopword list but it resulted in generation of more ordinary list like (one,one),(one,two). This led me to the conclusion that filters like minimum frequency 5 improved list significantly.

Top 50 bigrams by frequency differ from the top 50 bigrams scored by Mutual Information in following way:

- PMI list scores vary in range whereas bigram (raw frequency) scores were same for many pairs.

- Raw frequency pairs are highly collocated but the expressions are also very infrequent like north south, let us, among us, right wrong, year ago, year old, long ago etc. This is not the case with PMI as we have applied filters like ignoring all bigrams which occur less than 5 times in the corpus

c) If you modify the stop word list, or expand the methods of filtering, describe that here.
Word frequency distribution for Abraham Lincoln document originally resulted into following:

state	0.0008208336882535601
one	0.0006524079078807777
slaveri	0.0006040096951299783
upon	0.0005885222670497224
would	0.0005672270534393706
say	0.0005323803402587949
peopl	0.0004955976985681873
right	0.00047430248495783546
constitut	0.00045300727134748366
u	0.00044332762879732376
men	0.00042396834369700395
shall	0.00042009648667693997
judg	0.00040848091561674813
slave	0.00039105755902646026
man	0.00038524977349636434
may	0.00036976234541610846
govern	0.0003465312032957247
dougla	0.00034072341776562876
union	0.00032136413266530895
law	0.000315556347135213
make	0.00031362041862518 1
's	0.00030781263309508506
nation	0.00030200484756498914
time	0.00030200484756498914
question	0.0002768377693457336
great	0.00027490184842454137
think	0.0002632862773643495
free	0.0002632862773643495
declar	0.0002613503488543175
made	0.0002536066348141896
could	0.0002516707063041576
let	0.00023811920673393373
must	0.00023424734971386977
speech	0.0002303754926938058 1
thing	0.0002303754926938058 1
way	0.0002303754926938058 1
said	0.00022650363567374183
know	0.00022456770716370987
wrong	0.00022456770716370987
decis	0.00022069585014364588
ever	0.00021875992163361392
territori	0.00021682399312358193

new	0.00021488806461354996
year	0.00021488806461354996
never	0.00021101620759348598
much	0.00021101620759348598
place	0.00020908027908345402
without	0.00020520842206339003
unit	0.00019940063653329409
countri	0.00019746470802326212

So, I updated stop word list to -
morestopword=['upon','would','say','u','shall','\\s','could','must']
Similarly, I added more words to stop list based on word frequency obtained in Mark Twain document.

d) You may choose to also run top trigram lists, and include them in the analysis in part 3.
Since some of the bigrams don't give clear indication/purpose of their usage so I generated trigram so that I can match common pairs in bigram and trigram list and get a clear meaning of author's thought/style of writing.
For Abraham Lincoln documents (raw frequency)-
(('dred', 'scott', 'decis'), 0.0008201973599897476)
(('constitut', 'unit', 'state'), 0.0005126233499935922)
(('men', 'creat', 'equal'), 0.0005126233499935922)
(('cours', 'ultim', 'extinct'), 0.0003332051774958349)
(('frame', 'govern', 'live'), 0.0002563116749967961)
(('end', 'slaveri', 'agit'), 0.0002306805074971165)
(('father', 'frame', 'govern'), 0.0002306805074971165)
(('care', 'whether', 'slaveri'), 0.0002050493399974369)
(('public', 'mind', 'rest'), 0.0002050493399974369)
(('whether', 'slaveri', 'vote'), 0.0002050493399974369)
(('belief', 'cours', 'ultim'), 0.00017941817249775727)
(('half', 'slave', 'half'), 0.00017941817249775727)
(('mind', 'rest', 'belief'), 0.00017941817249775727)
(('presid', 'unit', 'state'), 0.00017941817249775727)
(('repeal', 'missouri', 'compromis'), 0.00017941817249775727)
(('rest', 'belief', 'cours'), 0.00017941817249775727)
(('slave', 'half', 'free'), 0.00017941817249775727)
(('slaveri', 'vote', 'vote'), 0.00017941817249775727)
(('suprem', 'court', 'decid'), 0.00017941817249775727)
(('alik', 'law', 'state'), 0.00015375700499807767)
(('h.', 'herndon', 'washington'), 0.00015375700499807767)
(('liberti', 'pursuit', 'happi'), 0.00015375700499807767)
(('life', 'liberti', 'pursuit'), 0.00015375700499807767)
(('peopl', 'exclud', 'slaveri'), 0.00015375700499807767)
(('proper', 'practic', 'relat'), 0.00015375700499807767)
(('repli', 'judg', 'dougla'), 0.00015375700499807767)
(('decis', 'suprem', 'court'), 0.00012815583749839805)
(('exclud', 'slaveri', 'limit'), 0.00012815583749839805)
(('exclud', 'slaveri', 'territori'), 0.00012815583749839805)
(('friend', 'judg', 'dougla'), 0.00012815583749839805)
(('fugit', 'slave', 'law'), 0.00012815583749839805)

(('govern', 'unit', 'state'), 0.00012815583749839805)
(('hous', 'divid', 'stand'), 0.00012815583749839805)
(('idea', 'anyth', 'wrong'), 0.00012815583749839805)
(('lincoln', 'repli', 'dougla'), 0.00012815583749839805)
(('north', 'well', 'south'), 0.00012815583749839805)
(('put', 'end', 'slaveri'), 0.00012815583749839805)
(('question', 'suprem', 'court'), 0.00012815583749839805)
(('unit', 'state', 'constitut'), 0.00012815583749839805)
(('unit', 'state', 'territori'), 0.00012815583749839805)
(('_from', 'lincoln', 'repli'), 0.00010252466999871845)
(('annual', 'messag', 'congress'), 0.00010252466999871845)
(('arrest', 'spread', 'place'), 0.00010252466999871845)
(('care', 'exclud', 'idea'), 0.00010252466999871845)
(('chief', 'justic', 'taney'), 0.00010252466999871845)
(('constitut', 'judg', 'dougla'), 0.00010252466999871845)
(('cranberri', 'law', 'indiana'), 0.00010252466999871845)
(('declar', 'men', 'creat'), 0.00010252466999871845)
(('dred', 'scott', 'case'), 0.00010252466999871845)
(('exclud', 'idea', 'anyth'), 0.00010252466999871845)

For Abraham Lincoln documents (PMI)-
(('h.', 'herndon', 'washington'), 21.163632556218396)
(('liberti', 'pursuit', 'happi'), 20.000982218573835)
(('cours', 'ultim', 'extinct'), 19.293279546774485)
(('life', 'liberti', 'pursuit'), 18.25555504565943)
(('rest', 'belief', 'cours'), 18.0689287912243)
(('belief', 'cours', 'ultim'), 17.90543005894142)
(('mind', 'rest', 'belief'), 17.830141931637183)
(('dred', 'scott', 'decis'), 17.623468632824252)
(('repeal', 'missouri', 'compromis'), 17.582571493179568)
(('proper', 'practic', 'relat'), 17.25614838107429)
(('public', 'mind', 'rest'), 16.88952047871611)
(('hous', 'divid', 'stand'), 16.724009418661282)
(('suprem', 'court', 'decid'), 16.447207740493024)
(('frame', 'govern', 'live'), 16.36431495322621)
(('father', 'frame', 'govern'), 16.04238685833885)
(('men', 'creat', 'equal'), 15.955886089341114)
(('half', 'slave', 'half'), 15.74423337763508)
(('idea', 'anyth', 'wrong'), 15.316377967683888)
(('north', 'well', 'south'), 15.208438310436417)
(('exclud', 'slaveri', 'limit'), 15.048155338798466)
(('end', 'slaveri', 'agit'), 14.989261649744893)
(('decis', 'suprem', 'court'), 14.856811353721248)
(('question', 'suprem', 'court'), 14.5298300311076)
(('fugit', 'slave', 'law'), 14.23350851628636)
(('lincoln', 'repli', 'dougla'), 14.141977361154986)
(('slave', 'half', 'free'), 13.610966846771614)
(('put', 'end', 'slaveri'), 13.372590289296404)
(('repli', 'judg', 'dougla'), 13.207474534338601)
(('care', 'whether', 'slaveri'), 13.168231790790209)

(('alik', 'law', 'state'), 12.92433360977956)
(('exclud', 'slaveri', 'territori'), 12.24080041674086)
(('whether', 'slaveri', 'vote'), 12.237940762601614)
(('presid', 'unit', 'state'), 11.91913657591429)
(('slaveri', 'vote', 'vote'), 11.709012296794786)
(('constitut', 'unit', 'state'), 11.540624952660558)
(('friend', 'judg', 'dougla'), 11.51559682970094)
(('peopl', 'exclud', 'slaveri'), 11.311189744632252)
(('unit', 'state', 'territori'), 10.603634750186366)
(('govern', 'unit', 'state'), 9.927173894979713)
(('unit', 'state', 'constitut'), 9.540624952660565)

For Mark Twain documents (raw frequency)-
(('introduc', 'mr', 'clemen'), 0.0002421487911187274)
(('rev', 'twichel', 'hartford'), 0.0002421487911187274)
(('twenti', 'five', 'year'), 0.0002421487911187274)
(('citi', 'new', 'york'), 0.00022352196103267146)
(('villa', 'di', 'quarto'), 0.00018626830086055954)
(('five', 'year', 'ago'), 0.00016764147077450358)
(('first', 'time', 'ever'), 0.00014901464068844763)
(('forti', 'two', 'year'), 0.0001303878106023917)
(('mani', 'year', 'ago'), 0.0001303878106023917)
(('new', 'york', 'citi'), 0.0001303878106023917)
(('seven', 'year', 'ago'), 0.0001303878106023917)
(('address', 'dinner', 'given'), 0.00011176098051633573)
(('di', 'quarto', 'florenc'), 0.00011176098051633573)
(('honor', 'mr', 'clemen'), 0.00011176098051633573)
(('mr', 'chang', 'riley'), 0.00011176098051633573)
(('new', 'england', 'weather'), 0.00011176098051633573)
(('seven', 'year', 'old'), 0.00011176098051633573)
(('thirti', 'six', 'year'), 0.00011176098051633573)
(('three', 'year', 'ago'), 0.00011176098051633573)
(('mr', 'clemen', 'appear'), 9.313415043027977e-05)
(('mr', 'clemen', 'introduc'), 9.313415043027977e-05)
(('mr', 'eng', 'nye'), 9.313415043027977e-05)
(('new', 'york', 'may'), 9.313415043027977e-05)
(('new', 'york', 'villa'), 9.313415043027977e-05)
(('oliv', 'wendel', 'holm'), 9.313415043027977e-05)
(('project', 'gutenberg', 'ebook'), 9.313415043027977e-05)
(('seventi', 'two', 'year'), 9.313415043027977e-05)
(('sir', 'mortim', 'durand'), 9.313415043027977e-05)
(('twenti', 'four', 'hour'), 9.313415043027977e-05)
(('york', 'villa', 'di'), 9.313415043027977e-05)
(('christian', 'privat', 'moral'), 7.450732034422381e-05)
(('club', 'new', 'york'), 7.450732034422381e-05)
(('dear', 'mr', 'clemen'), 7.450732034422381e-05)
(('dinner', 'given', 'honor'), 7.450732034422381e-05)
(('doubt', 'tell', 'truth'), 7.450732034422381e-05)
(('eight', 'year', 'ago'), 7.450732034422381e-05)
(('end', 'project', 'gutenberg'), 7.450732034422381e-05)

(('four', 'hundr', 'mile'), 7.450732034422381e-05)
(('fourth', 'juli', 'night'), 7.450732034422381e-05)
(('live', 'new', 'york'), 7.450732034422381e-05)
(('mr', 'clemen', 'guest'), 7.450732034422381e-05)
(('mr', 'clemen', 'mr'), 7.450732034422381e-05)
(('new', 'england', 'societi'), 7.450732034422381e-05)
(('new', 'york', 'novemb'), 7.450732034422381e-05)
(('novemb', 'mr', 'clemen'), 7.450732034422381e-05)
(('pudd', 'nhead', 'wilson'), 7.450732034422381e-05)
(('put', 'trust', 'god'), 7.450732034422381e-05)
(('san', 'francisco', 'earthquak'), 7.450732034422381e-05)
(('sixti', 'year', 'ago'), 7.450732034422381e-05)
(('st', 'clair', 'mckelway'), 7.450732034422381e-05)

For Mark Twain documents (PMI)-
(('project', 'gutenberg', 'ebook'), 25.809806750105942)
(('oliv', 'wendel', 'holm'), 24.71027107655503)
(('sir', 'mortim', 'durand'), 23.54799964765615)
(('villa', 'di', 'quarto'), 22.98780505208394)
(('di', 'quarto', 'florenc'), 22.321228785809136)
(('rev', 'twichel', 'hartford'), 19.746458236113234)
(('york', 'villa', 'di'), 18.872327834664002)
(('mr', 'eng', 'nye'), 18.82553362318506)
(('address', 'dinner', 'given'), 15.845200656641143)
(('mr', 'chang', 'riley'), 15.815549534612444)
(('twenti', 'four', 'hour'), 15.044989307983144)
(('thirti', 'six', 'year'), 14.76743079611203)
(('new', 'england', 'weather'), 14.653407605234367)
(('new', 'york', 'villa'), 14.480010411885242)
(('citi', 'new', 'york'), 14.277381245370222)
(('twenti', 'five', 'year'), 14.0033064468121)
(('new', 'york', 'citi'), 13.499773666706673)
(('introduc', 'mr', 'clemen'), 12.968683710076142)
(('seven', 'year', 'ago'), 12.775326588798716)
(('five', 'year', 'ago'), 12.633854162845836)
(('forti', 'two', 'year'), 12.529955936708852)
(('seventi', 'two', 'year'), 12.491988086509828)
(('mr', 'clemen', 'appear'), 11.912100181709771)
(('seven', 'year', 'old'), 11.793426004644974)
(('first', 'time', 'ever'), 11.730539674061017)
(('new', 'york', 'may'), 11.712456497885611)
(('three', 'year', 'ago'), 11.60637342642373)
(('mani', 'year', 'ago'), 11.590902017661286)
(('mr', 'clemen', 'introduc'), 11.590172086822406)
(('honor', 'mr', 'clemen'), 10.630814071319762)

Task 3 Comparison and Conclusion-

Part 1- Clearly describe the problem or question you are trying to address through the comparison between the two selected documents.

I will try to respond to following question based on my word/bigram frequency analysis:
Abraham Lincoln was an American politician and lawyer who served as the 16th President of the United States from March 1861 until his assassination in April 1865.
Samuel Langhorne Clemens better known by his pen name Mark Twain, was an American writer, humorist, entrepreneur, publisher, and lecturer. (source Wikipedia)

Does their designation/personality reflect their writing styles and what's the difference in their writing styles. In other words-how writing styles of Abraham Lincoln (president politician, lawyer) as an author differ from Mark Twain (well established writer)?

Part 2- Present and explain insights or conclusions based on the comparison to answer the question
Writing style is the way writing is dressed up to fit the specific context, purpose, or audience. Word choice also contribute to the style of a piece of writing.

Writing Type: It reflects his or her personality, unique voice, and way of approaching the audience and readers.
Abraham Lincoln seem to have persuasive type of writing. Persuasive writing's main purpose is to convince. It contains the opinions and biases of the author to convince others to agree with the author's point of view. This is clearly reflected from buzzwords like 'never', 'get', 'make' and phrases(bigrams) like 'come end', 'save union', 'let us', 'support constitution', 'question court', 'go back'. It is well known that President Abraham Lincoln had to face a civil war and it is clear from above phrases that he was trying to persuade people to end war.
Mark Twain seem to have narrative writing type. Narrative writing's main purpose is to tell a story. The author will create different characters and tell you what happens to them. It has definite and logical beginnings, intervals, and endings. This is somewhat indicated through trigrams like 'mani year ago', 'seven year old', 'five year ago', 'three year ago'. These phrases seem to be exemplary suitable for introduction part of story/novel

Audience: Mark Twain's addressing subjected to individuals indicated from bigrams like 'mr. roger', 'mr. carnegi', 'dear joe' or even to social gathering like dinner parties indicated from bigrams like 'ladi gentlemen' and phrases(trigrams) like 'address dinner given'. A phrase like ladies and gentlemen May I have your attention will be used in informal ceremonies rather than in poilitical party meeting.
Inspite of bigrams like 'public mind', 'public sentiment' and trigram like 'annual message congress', there is no clear indication to whom Abraham Lincoln's talk targeted but based on word frequency it seems to be connected to lawmakers and public as word frequency included lot of words like 'constitution', 'union', 'state', 'government', 'law', 'war', 'congress', 'nation'

Word Choice: It is an essential element of style that reveals the writer's personality. A writer's word choice can be impersonal or chatty, authoritative or reflective, objective or passionate, serious or funny.
Based on word frequency, Abraham Lincoln word choice indicate authoritative personality. Words like 'right', 'power', 'wrong', 'good' give us hint of his authoritative personality. Bigrams (raw frequency) like 'let us', 'question whether', 'go back' give us clearer picture of his leadership qualities. Even PMI bigrams like 'support constitution' and trigrams like 'end slavery agit', 'fugit slavery law' indicate the same. Affirmative phrases like these depict authoritative power of Abraham Lincoln on his supporters

In case of Mark Twain, word frequency does not give any kind of indications but bigrams like 'friend mine', 'old friend' reflect his chatty/carefree nature.

Apart from different writing style, these documents seem to belong to era of **different literary movement-**

Realism (1860-1914): Promoted facts over intellectual or emotional reasoning. Includes stories of everyday people and is written in natural language. Mark Twain writings also belonged to these time frame and seem to be connected to people with trigrams like ‘mani year ago’, ‘seven year old’, ‘san francisco earthquak’. Seven year old , many years ago phrase seem to be part of character description of a story. San Francisco earthquake seems to be related to real life disaster.

Transcendentalism (1830-1850): Everything in nature is meaningful, symbolic and important; Every hum an being is born inherently good. This idea seems to be connected to anti-slavery and is reflected in Abrah am Lincoln writing with trigrams like 'exclud slaveri' 'limit', 'end slaveri agit', 'slave half free', 'put end sla veri'

Conclusion: Abraham Lincoln’s persuasive writing type targeting public and lawmakers and authoritative personality matches his occupation of being a politician and lawyer.
Mark Twain’s narrative writing style targeting social gathering/individual with chatty nature matches at le ast his occupation of being a writer and lecturer. We can’t comment of him being a humorist or entrepren eur based on this document**.**