

SURIYA S

225229140

Exercise-1:Process simple bigram datafile

```
In [25]: #STEP 1: Open the file, count_2w.txt
myfile=open('count_2w.txt','r',encoding='utf8')
myfile.read()
\n18th centuries\t114624\n18th century\t1347520\n18th OT\t1205474\n1920s and\t
201458\n1930s and\t201470\n1940s and\t170389\n1950s and\t393926\n1960s and\t5
16205\n1970s and\t549181\n1970s to\t102387\n1980s and\t599473\n1980s to\t1165
16\n1990s and\t239890\n19th and\t397448\n19th centuries\t135465\n19th century
\t3028799\n19th of\t211562\n1px solid\t353472\n1st and\t675070\n1st century\t
114479\n1st class\t336408\n1st day\t224931\n1st ed\t564270\n1st edition\t3853
67\n1st floor\t318319\n1st for\t100767\n1st grade\t124571\n1st half\t126095\n
1st in\t204580\n1st item\t164506\n1st month\t145269\n1st of\t572039\n1st plac
e\t336631\n1st quarter\t140622\n1st round\t161251\n1st tcm\t705500\n1st time
\t347093\n1st to\t227604\n1st year\t290655\n20s and\t111681\n20six does\t1831
32\n20th and\t144331\n20th anniversary\t215022\n20th centuries\t175216\n20th
century\t3262785\n20th of\t268218\n21st birthday\t124863\n21st century\t23494
97\n21st of\t231405\n22nd of\t214870\n23rd of\t269027\n24th of\t327460\n25th
anniversary\t261023\n25th of\t397735\n26th of\t271707\n27th of\t276619\n28th
of\t268705\n29th of\t247411\n2d at\t268493\n2d ed\t191384\n2nd and\t522559\n2
nd century\t101359\n2nd day\t243291\n2nd ed\t828486\n2nd edition\t590646\n2nd
floor\t455461\n2nd generation\t111890\n2nd grade\t142340\n2nd half\t164011\n2
nd hand\t145568\n2nd in\t223137\n2nd mortgage\t140302\n2nd of\t225805\n2nd p1
ace\t319158\n2nd quarter\t129953\n2nd round\t137201\n2nd time\t166337\n2nd to
\t117206\n2nd year\t275672\n2pac bush\t145934\n2pac shakira\t113587\n30s and
```

```
In [26]: #STEP 2:  
with open('count_2w.txt') as f:  
    lines=[line.rstrip() for line in f]  
lines
```

```
Out[26]: ['0Uplink verified\t523545',  
          '0km to\t116103',  
          '1000s of\t939476',  
          '100s of\t539389',  
          '100th anniversary\t158621',  
          '10am to\t376141',  
          '10th and\t183715',  
          '10th anniversary\t242830',  
          '10th century\t117755',  
          '10th grade\t174046',  
          '10th in\t107194',  
          '10th of\t277970',  
          '11am to\t127624',  
          '11th and\t178884',  
          '11th century\t168601',  
          '11th grade\t126301',  
          '11th of\t189501',  
          '125Mbps w\t108645',  
          '12th and\t136706',  
          '12th century\t117755',  
          '12th grade\t174046']
```

```
In [27]: mini=lines[:10]
```

```
In [28]: mini
```

```
Out[28]: ['0Uplink verified\t523545',  
          '0km to\t116103',  
          '1000s of\t939476',  
          '100s of\t539389',  
          '100th anniversary\t158621',  
          '10am to\t376141',  
          '10th and\t183715',  
          '10th anniversary\t242830',  
          '10th century\t117755',  
          '10th grade\t174046']
```

```
In [29]: mini[0]
```

```
Out[29]: '0Uplink verified\t523545'
```

```
In [30]: mini[0].split()
```

```
Out[30]: ['0Uplink', 'verified', '523545']
```

```
In [31]: mini_list=[]
```

```
In [32]: for m in mini:  
    (w1,w2,count)=m.split()  
    count=int(count)  
    mini_list.append(((w1,w2),count))
```

In [33]: mini_list

```
Out[33]: [(_('0Uplink', 'verified'), 523545),
          (_('0km', 'to'), 116103),
          (_('1000s', 'of'), 939476),
          (_('100s', 'of'), 539389),
          (_('100th', 'anniversary'), 158621),
          (_('10am', 'to'), 376141),
          (_('10th', 'and'), 183715),
          (_('10th', 'anniversary'), 242830),
          (_('10th', 'century'), 117755),
          (_('10th', 'grade'), 174046)]
```

In [34]: mini_list[0]

```
Out[34]: (_('0Uplink', 'verified'), 523545)
```

In [35]: #STEP 3: build goog2w_fd

```
import nltk
goog2w_fd=nltk.FreqDist()
for i in lines:
    w1,w2,count=i.split()
    goog2w_fd[w1,w2]=count
```

In [36]: goog2w_fd[(_('of', 'the'))]

```
Out[36]: '2766332391'
```

In [37]: goog2w_fd[(_('so', 'beautiful'))]

```
Out[37]: '612472'
```

In [38]: #STEP 4:Explore

#1. Top-10 bigrams

```
goog2w_fd.most_common(10)
```

```
Out[38]: [(_('You', 'think'), '999988'),
          (_('a', 'middle'), '999987'),
          (_('his', 'wife'), '9999448'),
          (_('traditional', 'and'), '999927'),
          (_('Ask', 'your'), '999907'),
          (_('towards', 'the'), '9998989'),
          (_('<S>', 'central'), '999848'),
          (_('no', 'man'), '999833'),
          (_('committee', 'members'), '999819'),
          (_('each', 'country'), '999818')]
```

```
In [39]: #2. Top-so bigrams
text=[]
for t in lines:
    (w1,w2,count)=t.split()
    count=int(count)
    if w1=='so':
        text.append(((w1,w2),count))
text
```

```
Out[39]: [ (('so', 'a'), 1565933),
  (('so', 'afraid'), 181401),
  (('so', 'after'), 400665),
  (('so', 'again'), 197409),
  (('so', 'ago'), 226156),
  (('so', 'all'), 894606),
  (('so', 'already'), 101152),
  (('so', 'also'), 233562),
  (('so', 'am'), 206896),
  (('so', 'amazing'), 165724),
  (('so', 'an'), 229478),
  (('so', 'and'), 905653),
  (('so', 'angry'), 217654),
  (('so', 'any'), 429057),
  (('so', 'anyone'), 118496),
  (('so', 'are'), 949912),
  (('so', 'as'), 6866078),
  (('so', 'at'), 1044557),
  (('so', 'awesome'), 193277),
  ('so', 'be'), 1007610)
```

```
In [40]: #Pickle
import pickle
with open('goog2w_list.pkl','wb') as handle:
    pickle.dump(goog2w_fd,handle,protocol=pickle.HIGHEST_PROTOCOL)
```

```
In [41]: handle
```

```
Out[41]: <_io.BufferedWriter name='goog2w_list.pkl'>
```

Exercise 2: Frequency distribution from Jane Austen Novels

```
In [42]: with open('austen-emma.txt','r')as fe:
    ae=fe.read(600)
```

```
In [43]: with open('austen-persuasion.txt','r')as fp:
    ap=fp.read(900)
```

```
In [44]: with open('austen-sense.txt','r')as fe:
    ase=fe.read(1000)
```

```
In [45]: from nltk.tokenize import sent_tokenize as st
```

In [46]: st(ae)

Out[46]: `['[Emma by Jane Austen 1816]\n\nVOLUME I\n\nCHAPTER I\n\nEmma Woodhouse, hand
some, clever, and rich, with a comfortable home\nand happy disposition, seemed
to unite some of the best blessings\nof existence; and had lived nearly twenty-
one years in the world\nwith very little to distress or vex her.]`

`"She was the youngest of the two daughters of a most affectionate,\nindulgent
father; and had, in consequence of her sister's marriage,\nbeen mistress of his
house from a very early period." ,`

`'Her mother\nhad died too long ago for her to have more than an indistinct\nre
membrance of her caresses; and her place had b']`

In [47]: st(ap)

Out[47]: `['[Persuasion by Jane Austen 1818]\n\nChapter 1\n\nSir Walter Elliot, of Ke
llynch Hall, in Somersetshire, was a man who,\nfor his own amusement, never too
k up any book but the Baronetage;\nthere he found occupation for an idle hour,
and consolation in a\ndistressed one; there his faculties were roused into admi
ration and\nrespect, by contemplating the limited remnant of the earliest paten
ts;\nthere any unwelcome sensations, arising from domestic affairs\nchanged nat
urally into pity and contempt as he turned over\nthe almost endless creations o
f the last century; and there,\nif every other leaf were powerless, he could re
ad his own history\nwith an interest which never failed.]`

`'This was the page at which\nthe favourite volume always opened:\n\n"ELLIOT OF KELLYNCH HALL." ,`

`'"Walter Elliot, born March 1, 1760, married, July 15, 1784, Elizabeth,\ndaugh
ter of James Stevenson, Esq.' ,
'of South Park, in th']`

In [48]: st(ase)

Out[48]: `['[Sense and Sensibility by Jane Austen 1811]\n\nCHAPTER 1\n\nThe family of D
ashwood had long been settled in Sussex.]`

`'Their estate was large, and their residence was at Norland Park,\nin the cent
re of their property, where, for many generations,\nthey had lived in so respec
table a manner as to engage\nthe general good opinion of their surrounding acqu
aintance.]`

`'The late owner of this estate was a single man, who lived\ninto a very advanced
age, and who for many years of his life,\nhad a constant companion and housekee
per in his sister.]`

`'But her death, which happened ten years before his own,\nproduced a great alt
eration in his home; for to supply\nher loss, he invited and received into his
house the family\nof his nephew Mr. Henry Dashwood, the legal inheritor\nof the
Norland estate, and the person to whom he intended\nto bequeath it.]`

`"In the society of his nephew and niece,\nand their children, the old Gentlema
n's days were\ncomfortably spent." ,`

`'His attachment to them all increased.]`

`'The constant attention']`

In [49]: `len(st(ae))`

Out[49]: 3

In [50]: `len(st(ap))`

Out[50]: 4

In [51]: `len(st(ase))`

Out[51]: 7

In [52]: `from nltk.tokenize import word_tokenize`

In [53]: `t1=word_tokenize(ae)
print(t1)`

```
[[], 'Emma', 'by', 'Jane', 'Austen', '1816', ']', 'VOLUME', 'I', 'CHAPTER',
'I', 'Emma', 'Woodhouse', ',', 'handsome', ',', 'clever', ',', 'and', 'rich',
',', 'with', 'a', 'comfortable', 'home', 'and', 'happy', 'disposition', ',', 's
eemed', 'to', 'unite', 'some', 'of', 'the', 'best', 'blessings', 'of', 'existen
ce', ';', 'and', 'had', 'lived', 'nearly', 'twenty-one', 'years', 'in', 'the',
'world', 'with', 'very', 'little', 'to', 'distress', 'or', 'vex', 'her', '.',
'She', 'was', 'the', 'youngest', 'of', 'the', 'two', 'daughters', 'of', 'a', 'm
ost', 'affectionate', ',', 'indulgent', 'father', ';', 'and', 'had', ',', 'in',
'consequence', 'of', 'her', 'sister', "'s", 'marriage', ',', 'been', 'mistres
s', 'of', 'his', 'house', 'from', 'a', 'very', 'early', 'period', '.', 'Her',
'mother', 'had', 'died', 'too', 'long', 'ago', 'for', 'her', 'to', 'have', 'mor
e', 'than', 'an', 'indistinct', 'remembrance', 'of', 'her', 'caresses', ';', 'a
nd', 'her', 'place', 'had', 'b']
```

In [54]: `t2=word_tokenize(ap)
print(t2)`

```
[[], 'Persuasion', 'by', 'Jane', 'Austen', '1818', ']', 'Chapter', '1', 'Sir',
'Walter', 'Elliot', ',', 'of', 'Kellynch', 'Hall', ',', 'in', 'Somersetshire',
',', 'was', 'a', 'man', 'who', ',', 'for', 'his', 'own', 'amusement', ',', 'nev
er', 'took', 'up', 'any', 'book', 'but', 'the', 'Baronetage', ';', 'there', 'h
e', 'found', 'occupation', 'for', 'an', 'idle', 'hour', ',', 'and', 'consolatio
n', 'in', 'a', 'distressed', 'one', ';', 'there', 'his', 'faculties', 'were',
'roused', 'into', 'admiration', 'and', 'respect', ',', 'by', 'contemplating',
'the', 'limited', 'remnant', 'of', 'the', 'earliest', 'patents', ';', 'there',
'any', 'unwelcome', 'sensations', ',', 'arising', 'from', 'domestic', 'affair
s', 'changed', 'naturally', 'into', 'pity', 'and', 'contempt', 'as', 'he', 'tur
ned', 'over', 'the', 'almost', 'endless', 'creations', 'of', 'the', 'last', 'ce
ntury', ';', 'and', 'there', ',', 'if', 'every', 'other', 'leaf', 'were', 'pow
erless', ',', 'he', 'could', 'read', 'his', 'own', 'history', 'with', 'an', 'int
erest', 'which', 'never', 'failed', '.', 'This', 'was', 'the', 'page', 'at', 'w
hich', 'the', 'favourite', 'volume', 'always', 'opened', ':', ``', 'ELLIOT',
'OF', 'KELLYNCH', 'HALL', '.', ``', 'Walter', 'Elliot', ',', 'born', 'March',
'1', ',', '1760', ',', 'married', ',', 'July', '15', ',', '1784', ',', 'Elizabe
th', ',', 'daughter', 'of', 'James', 'Stevenson', ',', 'Esq', '.', 'of', 'Sout
h', 'Park', ',', 'in', 'th']
```

```
In [55]: t3=word_tokenize(ase)
print(t3)
```

```
['[', 'Sense', 'and', 'Sensibility', 'by', 'Jane', 'Austen', '1811', ']', 'CHAP
TER', '1', 'The', 'family', 'of', 'Dashwood', 'had', 'long', 'been', 'settled',
'in', 'Sussex', '.', 'Their', 'estate', 'was', 'large', ',', 'and', 'their', 'r
esidence', 'was', 'at', 'Norland', 'Park', ',', 'in', 'the', 'centre', 'of', 't
heir', 'property', ',', 'where', ',', 'for', 'many', 'generations', ',', 'the
y', 'had', 'lived', 'in', 'so', 'respectable', 'a', 'manner', 'as', 'to', 'enga
ge', 'the', 'general', 'good', 'opinion', 'of', 'their', 'surrounding', 'acquai
ntance', '.', 'The', 'late', 'owner', 'of', 'this', 'estate', 'was', 'a', 'sing
le', 'man', ',', 'who', 'lived', 'to', 'a', 'very', 'advanced', 'age', ',', 'an
d', 'who', 'for', 'many', 'years', 'of', 'his', 'life', ',', 'had', 'a', 'const
ant', 'companion', 'and', 'housekeeper', 'in', 'his', 'sister', '.', 'But', 'he
r', 'death', ',', 'which', 'happened', 'ten', 'years', 'before', 'his', 'own',
', 'produced', 'a', 'great', 'alteration', 'in', 'his', 'home', ';', 'for',
'to', 'supply', 'her', 'loss', ',', 'he', 'invited', 'and', 'received', 'into',
'his', 'house', 'the', 'family', 'of', 'his', 'nephew', 'Mr.', 'Henry', 'Dashwo
od', ',', 'the', 'legal', 'inheritor', 'of', 'the', 'Norland', 'estate', ',',
'and', 'the', 'person', 'to', 'whom', 'he', 'intended', 'to', 'bequeath', 'it',
'.', 'In', 'the', 'society', 'of', 'his', 'nephew', 'and', 'niece', ',', 'and',
'their', 'children', ',', 'the', 'old', 'Gentleman', "'s", 'days', 'were', 'com
fortably', 'spent', '.', 'His', 'attachment', 'to', 'them', 'all', 'increased',
'.', 'The', 'constant', 'attention']
```

```
In [56]: from nltk import *
```

```
In [57]: d1=FreqDist(t1)
d1
```

```
Out[57]: FreqDist({'s': 1,
',': 8,
'.': 2,
'1816': 1,
';': 3,
'Austen': 1,
'CHAPTER': 1,
'Emma': 2,
'Her': 1,
'I': 2,
'Jane': 1,
'She': 1,
'VOLUME': 1,
'Woodhouse': 1,
 '[': 1,
 ']': 1,
 'a': 3,
 'affectionate': 1,
 'ago': 1,
 'an': 1,
 'and': 5,
 'b': 1,
 'been': 1,
 'best': 1,
 'blessings': 1,
 'by': 1,
 'caresses': 1,
 'clever': 1,
 'comfortable': 1,
 'consequence': 1,
 'daughters': 1,
 'died': 1,
 'disposition': 1,
 'distress': 1,
 'early': 1,
 'existence': 1,
 'father': 1,
 'for': 1,
 'from': 1,
 'had': 4,
 'handsome': 1,
 'happy': 1,
 'have': 1,
 'her': 5,
 'his': 1,
 'home': 1,
 'house': 1,
 'in': 2,
 'indistinct': 1,
 'indulgent': 1,
 'little': 1,
 'lived': 1,
 'long': 1,
 'marriage': 1,
```

```
'mistress': 1,
'more': 1,
'most': 1,
'mother': 1,
'nearly': 1,
'of': 7,
'or': 1,
'period': 1,
'place': 1,
'remembrance': 1,
'rich': 1,
'seemed': 1,
'sister': 1,
'some': 1,
'than': 1,
'the': 4,
'to': 3,
'too': 1,
'twenty-one': 1,
'two': 1,
'unite': 1,
'very': 2,
'vex': 1,
'was': 1,
'with': 2,
'world': 1,
'years': 1,
'youngest': 1})
```

In [58]: d2=FreqDist(t2)
d2

Out[58]: FreqDist({' ': 19,
'.': 3,
'1': 2,
'15': 1,
'1760': 1,
'1784': 1,
'1818': 1,
':': 1,
';': 4,
'Austen': 1,
'Baronetage': 1,
'Chapter': 1,
'ELLIOT': 1,
'Elizabeth': 1,
'Elliot': 2,
'Esq': 1,
'HALL': 1,
'Hall': 1,
'James': 1,
'...': 1})

```
In [59]: d3=FreqDist(t3)
d3
```

```
Out[59]: FreqDist({'s': 1,
',': 15,
'.': 6,
'1': 1,
'1811': 1,
';': 1,
'Austen': 1,
'But': 1,
'CHAPTER': 1,
'Dashwood': 2,
'Gentleman': 1,
'Henry': 1,
'His': 1,
'In': 1,
'Jane': 1,
'Mr.': 1,
'Norland': 2,
'Park': 1,
'Sense': 1,
'`': 1,
'`.': 1,
'`1816': 1,
'`CHAPTER': 1,
'`Dashwood': 1,
'`Gentleman': 1,
'`Henry': 1,
'`His': 1,
'`In': 1,
'`Jane': 1,
'`Mr.': 1,
'`Norland': 1,
'`Park': 1,
'`Sense': 1,
'`VOLUME': 1,
'`Woodhouse': 1,
'`handsome': 1,
'`lever': 1,
'`rich': 1,
'`comfortable': 1,
'`home': 1,
'`happy': 1,
'`disposition': 1,
'`seemed': 1,
'`unite': 1,
'`some': 1,
'`best': 1,
'`blessing': 1,
'`existence': 1,
'`lived': 1,
'`nearly': 1,
'`twenty-one': 1,
'`years': 1,
'`world': 1,
'`little': 1,
'`distress': 1,
'`or': 1,
'`vex': 1,
`She': 1,
`was': 1,
`youngest': 1,
`two': 1}]
```

```
In [60]: print(d1.most_common(50))
```

```
[(',', 8), ('of', 7), ('and', 5), ('her', 5), ('the', 4), ('had', 4), ('a', 3),
('to', 3), (';', 3), ('Emma', 2), ('I', 2), ('with', 2), ('in', 2), ('very',
2), ('.', 2), ('[', 1), ('by', 1), ('Jane', 1), ('Austen', 1), ('1816', 1),
(']', 1), ('VOLUME', 1), ('CHAPTER', 1), ('Woodhouse', 1), ('handsome', 1), ('c
lever', 1), ('rich', 1), ('comfortable', 1), ('home', 1), ('happy', 1), ('dispo
sition', 1), ('seemed', 1), ('unite', 1), ('some', 1), ('best', 1), ('blessing
s', 1), ('existence', 1), ('lived', 1), ('nearly', 1), ('twenty-one', 1), ('yea
rs', 1), ('world', 1), ('little', 1), ('distress', 1), ('or', 1), ('vex', 1),
('She', 1), ('was', 1), ('youngest', 1), ('two', 1)]
```

```
In [61]: print(d2.most_common(50))
```

```
[(',', 19), ('the', 7), ('of', 5), (';', 4), ('there', 4), ('and', 4), ('in',
3), ('his', 3), ('he', 3), ('.', 3), ('by', 2), ('1', 2), ('Walter', 2), ('Elli
ot', 2), ('was', 2), ('a', 2), ('for', 2), ('own', 2), ('never', 2), ('any',
2), ('an', 2), ('were', 2), ('into', 2), ('which', 2), ('``', 2), ('[', 1), ('P
ersuasion', 1), ('Jane', 1), ('Austen', 1), ('1818', 1), (']', 1), ('Chapter',
1), ('Sir', 1), ('Kellynch', 1), ('Hall', 1), ('Somersetshire', 1), ('man', 1),
('who', 1), ('amusement', 1), ('took', 1), ('up', 1), ('book', 1), ('but', 1),
('Baronetage', 1), ('found', 1), ('occupation', 1), ('idle', 1), ('hour', 1),
('consolation', 1), ('distressed', 1)]
```

```
In [62]: print(d3.most_common(50))
```

```
[(',', 15), ('and', 8), ('of', 8), ('the', 8), ('his', 7), ('.', 6), ('to', 6),
('in', 5), ('a', 5), ('their', 4), ('The', 3), ('had', 3), ('estate', 3), ('wa
s', 3), ('for', 3), ('family', 2), ('Dashwood', 2), ('Norland', 2), ('many',
2), ('lived', 2), ('who', 2), ('years', 2), ('constant', 2), ('her', 2), ('he',
2), ('nephew', 2), ('[', 1), ('Sense', 1), ('Sensibility', 1), ('by', 1), ('Jan
e', 1), ('Austen', 1), ('1811', 1), (']', 1), ('CHAPTER', 1), ('1', 1), ('lon
g', 1), ('been', 1), ('settled', 1), ('Sussex', 1), ('Their', 1), ('large', 1),
('residence', 1), ('at', 1), ('Park', 1), ('centre', 1), ('property', 1), ('whe
re', 1), ('generations', 1), ('they', 1)]
```

Excercise 3: Bigram frequencies of Jane Austen Novels

```
In [63]: with open("austen-emma.txt") as fn:
    nov=fn.read(500)
    print(nov)
```

[Emma by Jane Austen 1816]

VOLUME I

CHAPTER I

Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence; and had lived nearly twenty-one years in the world with very little to distress or vex her.

She was the youngest of the two daughters of a most affectionate, indulgent father; and had, in consequence of her sister's marriage, been mistress of his house from a very early period. Her mother had died t

```
In [64]: tokenizer=nltk.tokenize.WhitespaceTokenizer()
tok=tokenizer.tokenize(nov)
print(tok)
```

```
['[Emma', 'by', 'Jane', 'Austen', '1816]', 'VOLUME', 'I', 'CHAPTER', 'I', 'Emm
a', 'Woodhouse,', 'handsome,', 'clever,', 'and', 'rich,', 'with', 'a', 'comfort
able', 'home', 'and', 'happy', 'disposition,', 'seemed', 'to', 'unite', 'some',
'of', 'the', 'best', 'blessings', 'of', 'existence;', 'and', 'had', 'lived', 'n
early', 'twenty-one', 'years', 'in', 'the', 'world', 'with', 'very', 'little',
'to', 'distress', 'or', 'vex', 'her.', 'She', 'was', 'the', 'youngest', 'of',
'the', 'two', 'daughters', 'of', 'a', 'most', 'affectionate,', 'indulgent', 'fa
ther;', 'and', 'had', 'in', 'consequence', 'of', 'her', 'sister's', 'marriag
e,', 'been', 'mistress', 'of', 'his', 'house', 'from', 'a', 'very', 'early', 'p
eriod.', 'Her', 'mother', 'had', 'died', 't']
```

```
In [65]: tokenizer=nltk.tokenize.WhitespaceTokenizer()
a_tok=tokenizer.tokenize(nov.lower())
print(a_tok)
```

```
['[emma', 'by', 'jane', 'austen', '1816]', 'volume', 'i', 'chapter', 'i', 'emm
a', 'woodhouse,', 'handsome,', 'clever,', 'and', 'rich,', 'with', 'a', 'comfort
able', 'home', 'and', 'happy', 'disposition,', 'seemed', 'to', 'unite', 'some',
'of', 'the', 'best', 'blessings', 'of', 'existence;', 'and', 'had', 'lived', 'n
early', 'twenty-one', 'years', 'in', 'the', 'world', 'with', 'very', 'little',
'to', 'distress', 'or', 'vex', 'her.', 'she', 'was', 'the', 'youngest', 'of',
'the', 'two', 'daughters', 'of', 'a', 'most', 'affectionate', 'indulgent', 'fa
ther;', 'and', 'had', 'in', 'consequence', 'of', 'her', "sister's", 'marriag
e,', 'been', 'mistress', 'of', 'his', 'house', 'from', 'a', 'very', 'early', 'p
eriod.', 'her', 'mother', 'had', 'died', 't']
```

```
In [66]: b2=list(nltk.bigrams(tok))
b2fd=nltk.FreqDist(b2)
b2fd
```

```
Out[66]: FreqDist({('1816]', 'VOLUME'): 1,
                    ('Austen', '1816']): 1,
                    ('CHAPTER', 'I'): 1,
                    ('Emma', 'Woodhouse,'): 1,
                    ('Her', 'mother'): 1,
                    ('I', 'CHAPTER'): 1,
                    ('I', 'Emma'): 1,
                    ('Jane', 'Austen'): 1,
                    ('She', 'was'): 1,
                    ('VOLUME', 'I'): 1,
                    ('Woodhouse,', 'handsome,'): 1,
                    ('[Emma', 'by'): 1,
                    ('a', 'comfortable'): 1,
                    ('a', 'most'): 1,
                    ('a', 'very'): 1,
                    ('affectionate,', 'indulgent'): 1,
                    ('and', 'had'): 1,
                    ('and', 'had,'): 1,
                    ('and', 'happy'): 1,
                    ('and', 'rich,'): 1,
                    ('been', 'mistress'): 1,
                    ('best', 'blessings'): 1,
                    ('blessings', 'of'): 1,
                    ('by', 'Jane'): 1,
                    ('clever,', 'and'): 1,
                    ('comfortable', 'home'): 1,
                    ('consequence', 'of'): 1,
                    ('daughters', 'of'): 1,
                    ('died', 't'): 1,
                    ('disposition,', 'seemed'): 1,
                    ('distress', 'or'): 1,
                    ('early', 'period.'): 1,
                    ('existence;', 'and'): 1,
                    ('father;', 'and'): 1,
                    ('from', 'a'): 1,
                    ('had', 'died'): 1,
                    ('had', 'lived'): 1,
                    ('had,', 'in'): 1,
                    ('handsome,', 'clever,'): 1,
                    ('happy', 'disposition,'): 1,
                    ('her', "sister's"): 1,
                    ('her.', 'She'): 1,
                    ('his', 'house'): 1,
                    ('home', 'and'): 1,
                    ('house', 'from'): 1,
                    ('in', 'consequence'): 1,
                    ('in', 'the'): 1,
                    ('indulgent', 'father;'): 1,
                    ('little', 'to'): 1,
                    ('lived', 'nearly'): 1,
                    ('marriage,', 'been'): 1,
                    ('mistress', 'of'): 1,
                    ('most', 'affectionate,'): 1,
```

```
('mother', 'had'): 1,
('nearly', 'twenty-one'): 1,
('of', 'a'): 1,
('of', 'existence;'): 1,
('of', 'her'): 1,
('of', 'his'): 1,
('of', 'the'): 2,
('or', 'vex'): 1,
('period.', 'Her'): 1,
('rich,', 'with'): 1,
('seemed', 'to'): 1,
("sister's", 'marriage,'): 1,
('some', 'of'): 1,
('the', 'best'): 1,
('the', 'two'): 1,
('the', 'world'): 1,
('the', 'youngest'): 1,
('to', 'distress'): 1,
('to', 'unite'): 1,
('twenty-one', 'years'): 1,
('two', 'daughters'): 1,
('unite', 'some'): 1,
('very', 'early'): 1,
('very', 'little'): 1,
('vex', 'her.'): 1,
('was', 'the'): 1,
('with', 'a'): 1,
('with', 'very'): 1,
('world', 'with'): 1,
('years', 'in'): 1,
('youngest', 'of'): 1})
```

```
In [67]: import re
from collections import Counter
```

```
In [68]: words=re.findall(r'so+ \w+',open('austen-emma.txt').read())
ab=Counter(zip(words))
print(ab)
```

Counter({('so much',): 95, ('so very',): 76, ('so well',): 30, ('so many',): 27, ('so long',): 27, ('so little',): 20, ('so far',): 17, ('so I',): 14, ('so kind',): 13, ('so good',): 12, ('so often',): 10, ('so soon',): 9, ('so great',): 8, ('so to',): 7, ('so fond',): 7, ('so she',): 7, ('so it',): 6, ('so anxious',): 6, ('so as',): 6, ('so you',): 6, ('so truly',): 6, ('so completely',): 5, ('so obliging',): 5, ('so extremely',): 5, ('so entirely',): 4, ('so happy',): 4, ('so interesting',): 4, ('so fast',): 4, ('so near',): 4, ('so pleased',): 4, ('so few',): 4, ('so that',): 4, ('so strong',): 4, ('so liberal',): 4, ('so miserable',): 4, ('so happily',): 3, ('so proper',): 3, ('so pleasantly',): 3, ('so superior',): 3, ('so warmly',): 3, ('so bad',): 3, ('so odd',): 3, ('so ill',): 3, ('so delighted',): 3, ('so particularly',): 3, ('so easily',): 3, ('so on',): 3, ('so attentive',): 3, ('so fortunate',): 3, ('so glad',): 3, ('so shocked',): 3, ('so at',): 3, ('so obliged',): 2, ('so perfectly',): 2, ('so dear',): 2, ('so busy',): 2, ('so did',): 2, ('so forth',): 2, ('so totally',): 2, ('so remarkably',): 2, ('so plainly',): 2, ('so charming',): 2, ('so surprised',): 2, ('so early',): 2, ('so too',): 2, ('so easy',): 2, ('so decidedly',): 2, ('so absolutely',): 2, ('so particular',): 2, ('so deceived',): 2, ('so palpably',): 2, ('so clever',): 2, ('so short',): 2, ('so cold',): 2, ('so high',): 2, ('so happened',): 2, ('so full',): 2, ('so thoroughly',): 2, ('so equal',): 2, ('so off',): 2, ('so naturally',): 2, ('so afraid',): 2, ('so deep',): 2, ('so kindly',): 2, ('so pale',): 2, ('so noble',): 2, ('so lovely',): 2, ('so mad',): 2, ('so nearly',): 2, ('so sorry',): 2, ('so cheerful',): 2, ('so unfeeling',): 2, ('so ready',): 2, ('so unperceived',): 1, ('so mild',): 1, ('so constantly',): 1, ('so comfortably',): 1, ('so avowed',): 1, ('so deservedly',): 1, ('so convenient',): 1, ('so just',): 1, ('so apparent',): 1, ('so sorrowful',): 1, ('so spent',): 1, ('so artlessly',): 1, ('so plain',): 1, ('so firmly',): 1, ('so genteel',): 1, ('so_then',): 1, ('so brilliant',): 1, ('so seldom',): 1, ('so nervous',): 1, ('so indeed',): 1, ('so pack',): 1, ('so doubtful',): 1, ('so with',): 1, ('so contemptible',): 1, ('so slightly',): 1, ('so by',): 1, ('so loudly',): 1, ('so materially',): 1, ('so hard',): 1, ('so delightful',): 1, ('so pointed',): 1, ('so equalled',): 1, ('so evidently',): 1, ('so immediately',): 1, ('so sought',): 1, ('so excellent',): 1, ('so prettily',): 1, ('so extreme',): 1, ('so wonder',): 1, ('so always',): 1, ('so silly',): 1, ('so satisfied',): 1, ('so smiling',): 1, ('so prosing',): 1, ('so undistinguishing',): 1, ('so apt',): 1, ('so dreadful',): 1, ('so respected',): 1, ('so tenderly',): 1, ('so grieved',): 1, ('so shocking',): 1, ('so conceited',): 1, ('so before',): 1, ('so prevalent',): 1, ('so heavy',): 1, ('so swiftly',): 1, ('so spoken',): 1, ('so or',): 1, ('so overcharge',): 1, ('so pleasant',): 1, ('so fenced',): 1, ('so hospitable',): 1, ('so interested',): 1, ('so sanguine',): 1, ('so sure',): 1, ('so careless',): 1, ('so rapidly',): 1, ('so frequent',): 1, ('so sensible',): 1, ('so misled',): 1, ('so blind',): 1, ('so complaisant',): 1, ('so misinterpreted',): 1, ('so active',): 1, ('so pointedly',): 1, ('so striking',): 1, ('so sudden',): 1, ('so indistinctly',): 1, ('so partial',): 1, ('so natural',): 1, ('so inevitable',): 1, ('so lately',): 1, ('so beautifully',): 1, ('so distinct',): 1, ('so considerate',): 1, ('so light',): 1, ('so intimate',): 1, ('so magnified',): 1, ('so cautious',): 1, ('so confined',): 1, ('so wish',): 1, ('so he',): 1, ('so glorious',): 1, ('so quick',): 1, ('so sweetly',): 1, ('so inseparably',): 1, ('so serving',): 1, ('so disappointed',): 1, ('so ended',): 1, ('so sluggish',): 1, ('so amiable',): 1, ('so quiet',): 1, ('so idolized',): 1, ('so cried',): 1, ('so acceptable',): 1, ('so properly',): 1, ('so reasonable',): 1, ('so delightfully',): 1, ('so rich',): 1, ('so warm',): 1, ('so large',): 1, ('so handsome',): 1, ('so abundant',): 1, ('so outree',): 1, ('so thoughtful',): 1, ('so mu

st'): 1, ('so effectually'): 1, ('so beautiful'): 1, ('so Patty'): 1, ('so honoured'): 1, ('so close'): 1, ('so imprudent'): 1, ('so limited'): 1, ('so from'): 1, ('so amusing'): 1, ('so indifferent'): 1, ('so indignant'): 1, ('so said'): 1, ('so right'): 1, ('so wretched'): 1, ('so now'): 1, ('so occupied'): 1, ('so unhappy'): 1, ('so highly'): 1, ('so generally'): 1, ('so exactly'): 1, ('so double'): 1, ('so secluded'): 1, ('so regular'): 1, ('so determined'): 1, ('so motherly'): 1, ('so the'): 1, ('so glibly'): 1, ('so calculated'): 1, ('so thrown'): 1, ('so exclusively'): 1, ('so disgusting'): 1, ('so needlessly'): 1, ('so does'): 1, ('so resolutely'): 1, ('so would'): 1, ('so infinitely'): 1, ('so fluently'): 1, ('so they'): 1, ('so impatient'): 1, ('so briskly'): 1, ('so vigorously'): 1, ('so young'): 1, ('so hardened'): 1, ('so gratified'): 1, ('so received'): 1, ('so then'): 1, ('so and'): 1, ('so gratefully'): 1, ('so found'): 1, ('so placed'): 1, ('so lain'): 1, ('so his'): 1, ('so arranged'): 1, ('so moving'): 1, ('so walking'): 1, ('so when'): 1, ('so favourable'): 1, ('so late'): 1, ('so silent'): 1, ('so dull'): 1, ('so irksome'): 1, ('so agitated'): 1, ('so brutal'): 1, ('so cruel'): 1, ('so depressed'): 1, ('so no'): 1, ('so justly'): 1, ('so astonished'): 1, ('so will'): 1, ('so simple'): 1, ('so dignified'): 1, ('so suddenly'): 1, ('so a'): 1, ('so herself'): 1, ('so peremptorily'): 1, ('so uneasy'): 1, ('so wonderful'): 1, ('so _very_'): 1, ('so expressly'): 1, ('so angry'): 1, ('so anxiously'): 1, ('so strange'): 1, ('so stoutly'): 1, ('so mistake'): 1, ('so mistaken'): 1, ('so dreadfully'): 1, ('so voluntarily'): 1, ('so satisfactory'): 1, ('so disinterested'): 1, ('so foolishly'): 1, ('so ingeniously'): 1, ('so entreated'): 1, ('so like'): 1, ('so cordially'): 1, ('so essential'): 1, ('so designedly'): 1, ('so hasty'): 1, ('so richly'): 1, ('so grateful'): 1, ('so tenaciously'): 1, ('so feeling'): 1, ('so engaging'): 1, ('so engaged'): 1, ('so hot'): 1, ('so useful'): 1, ('so attached'): 1, ('so peculiarly'): 1, ('so singularly'): 1, ('so taken'): 1, ('so recently'): 1, ('so fresh'): 1, ('so hateful'): 1, ('so heartily'): 1, ('so steady'): 1, ('so complete'): 1, ('so in'): 1, ('so suffered'): 1})

```
In [69]: a_tokfd=FreqDist(a_tok)
a_tokfd
```

```
Out[69]: FreqDist({'1816': 1,
                    '[emma': 1,
                    'a': 3,
                    'affectionate,' : 1,
                    'and': 4,
                    'austen': 1,
                    'been': 1,
                    'best': 1,
                    'blessings': 1,
                    'by': 1,
                    'chapter': 1,
                    'clever,' : 1,
                    'comfortable': 1,
                    'consequence': 1,
                    'daughters': 1,
                    'died': 1,
                    'disposition,' : 1,
                    'distress': 1,
                    'early': 1,
                    'emma': 1,
                    'existence;': 1,
                    'father;': 1,
                    'from': 1,
                    'had': 2,
                    'had,' : 1,
                    'handsome,' : 1,
                    'happy': 1,
                    'her': 2,
                    'her.': 1,
                    'his': 1,
                    'home': 1,
                    'house': 1,
                    'i': 2,
                    'in': 2,
                    'indulgent': 1,
                    'jane': 1,
                    'little': 1,
                    'lived': 1,
                    'marriage,' : 1,
                    'mistress': 1,
                    'most': 1,
                    'mother': 1,
                    'nearly': 1,
                    'of': 6,
                    'or': 1,
                    'period.': 1,
                    'rich,' : 1,
                    'seemed': 1,
                    'she': 1,
                    "sister's": 1,
                    'some': 1,
                    't': 1,
                    'the': 4,
                    'to': 2,
```

```
'twenty-one': 1,  
'two': 1,  
'unite': 1,  
'very': 2,  
'vex': 1,  
'volume': 1,  
'was': 1,  
'with': 2,  
'woodhouse, ': 1,  
'world': 1,  
'years': 1,  
'youngest': 1})
```

```
In [70]: a_bigrams=list(nltk.bigrams(a_tok))
a_bigrams
```

```
Out[70]: [('emma', 'by'),
('by', 'jane'),
('jane', 'austen'),
('austen', '1816'],
('1816]', 'volume'),
('volume', 'i'),
('i', 'chapter'),
('chapter', 'i'),
('i', 'emma'),
('emma', 'woodhouse,'),
('woodhouse,', 'handsome,'),
('handsome,', 'clever,'),
('clever,', 'and'),
('and', 'rich,'),
('rich,', 'with'),
('with', 'a'),
('a', 'comfortable'),
('comfortable', 'home'),
('home', 'and'),
('and', 'happy'),
('happy', 'disposition,'),
('disposition,', 'seemed'),
('seemed', 'to'),
('to', 'unite'),
('unite', 'some'),
('some', 'of'),
('of', 'the'),
('the', 'best'),
('best', 'blessings'),
('blessings', 'of'),
('of', 'existence;'),
('existence;', 'and'),
('and', 'had'),
('had', 'lived'),
('lived', 'nearly'),
('nearly', 'twenty-one'),
('twenty-one', 'years'),
('years', 'in'),
('in', 'the'),
('the', 'world'),
('world', 'with'),
('with', 'very'),
('very', 'little'),
('little', 'to'),
('to', 'distress'),
('distress', 'or'),
('or', 'vex'),
('vex', 'her.'),
('her.', 'she'),
('she', 'was'),
('was', 'the'),
('the', 'youngest'),
('youngest', 'of'),
('of', 'the'),
```

```
('the', 'two'),
('two', 'daughters'),
('daughters', 'of'),
('of', 'a'),
('a', 'most'),
('most', 'affectionate,'),
('affectionate,', 'indulgent'),
('indulgent', 'father;'),
('father;', 'and'),
('and', 'had,'),
('had,', 'in'),
('in', 'consequence'),
('consequence', 'of'),
('of', 'her'),
('her', "sister's"),
("sister's", 'marriage,'),
('marriage,', 'been'),
('been', 'mistress'),
('mistress', 'of'),
('of', 'his'),
('his', 'house'),
('house', 'from'),
('from', 'a'),
('a', 'very'),
('very', 'early'),
('early', 'period.'),
('period.', 'her'),
('her', 'mother'),
('mother', 'had'),
('had', 'died'),
('died', 't')]
```

```
In [71]: a_bigramfd=nltk.FreqDist(a_bigrams)
a_bigramfd
```

```
Out[71]: FreqDist({('1816]', 'volume'): 1,
                   ('[emma', 'by'): 1,
                   ('a', 'comfortable'): 1,
                   ('a', 'most'): 1,
                   ('a', 'very'): 1,
                   ('affectionate,', 'indulgent'): 1,
                   ('and', 'had'): 1,
                   ('and', 'had,'): 1,
                   ('and', 'happy'): 1,
                   ('and', 'rich,'): 1,
                   ('austen', '1816']): 1,
                   ('been', 'mistress'): 1,
                   ('best', 'blessings'): 1,
                   ('blessings', 'of'): 1,
                   ('by', 'jane'): 1,
                   ('chapter', 'i'): 1,
                   ('clever,', 'and'): 1,
                   ('comfortable', 'home'): 1,
                   ('consequence', 'of'): 1,
                   ('daughters', 'of'): 1,
                   ('died', 't'): 1,
                   ('disposition,', 'seemed'): 1,
                   ('distress', 'or'): 1,
                   ('early', 'period.'): 1,
                   ('emma', 'woodhouse,'): 1,
                   ('existence;', 'and'): 1,
                   ('father;', 'and'): 1,
                   ('from', 'a'): 1,
                   ('had', 'died'): 1,
                   ('had', 'lived'): 1,
                   ('had,', 'in'): 1,
                   ('handsome,', 'clever,'): 1,
                   ('happy', 'disposition,'): 1,
                   ('her', 'mother'): 1,
                   ('her', "sister's"): 1,
                   ('her.', 'she'): 1,
                   ('his', 'house'): 1,
                   ('home', 'and'): 1,
                   ('house', 'from'): 1,
                   ('i', 'chapter'): 1,
                   ('i', 'emma'): 1,
                   ('in', 'consequence'): 1,
                   ('in', 'the'): 1,
                   ('indulgent', 'father;'): 1,
                   ('jane', 'austen'): 1,
                   ('little', 'to'): 1,
                   ('lived', 'nearly'): 1,
                   ('marriage,', 'been'): 1,
                   ('mistress', 'of'): 1,
                   ('most', 'affectionate,'): 1,
                   ('mother', 'had'): 1,
                   ('nearly', 'twenty-one'): 1,
                   ('of', 'a'): 1,
                   ('of', 'existence;'): 1,
```

```
('of', 'her'): 1,
('of', 'his'): 1,
('of', 'the'): 2,
('or', 'vex'): 1,
('period.', 'her'): 1,
('rich,', 'with'): 1,
('seemed', 'to'): 1,
('she', 'was'): 1,
("sister's", 'marriage,'): 1,
('some', 'of'): 1,
('the', 'best'): 1,
('the', 'two'): 1,
('the', 'world'): 1,
('the', 'youngest'): 1,
('to', 'distress'): 1,
('to', 'unite'): 1,
('twenty-one', 'years'): 1,
('two', 'daughters'): 1,
('unite', 'some'): 1,
('very', 'early'): 1,
('very', 'little'): 1,
('vex', 'her.'): 1,
('volume', 'i'): 1,
('was', 'the'): 1,
('with', 'a'): 1,
('with', 'very'): 1,
('woodhouse,', 'handsome,'): 1,
('world', 'with'): 1,
('years', 'in'): 1,
('youngest', 'of'): 1})
```

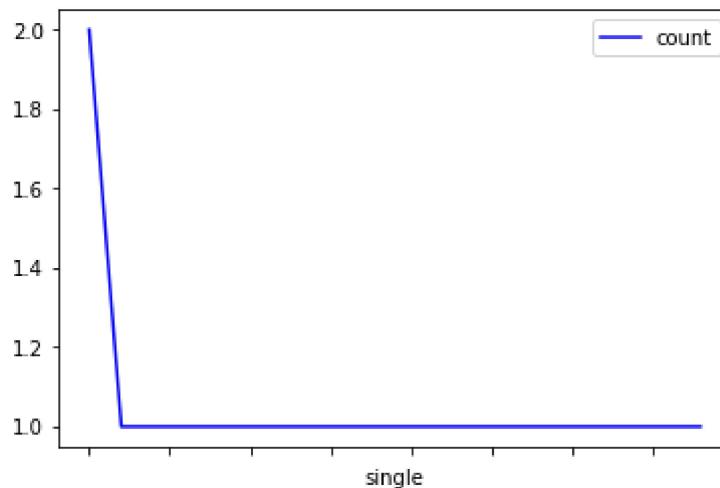
```
In [72]: from nltk.probability import ConditionalFreqDist
from nltk.tokenize import word_tokenize
```

```
In [76]: import pandas as pd  
df=pd.DataFrame(list(m.items()))  
df.columns=['single','count']  
df
```

Out[76]:

	single	count
0	(of, the)	2
1	([emma, by)	1
2	(by, jane)	1
3	(jane, austen)	1
4	(austen, 1816])	1
5	(1816], volume)	1
6	(volume, i)	1
7	(i, chapter)	1
8	(chapter, i)	1
9	(i, emma)	1
10	(emma, woodhouse,)	1
11	(woodhouse,, handsome,)	1
12	(handsome,, clever,)	1
13	(clever,, and)	1
14	(and, rich,)	1
15	(rich,, with)	1
16	(with, a)	1
17	(a, comfortable)	1
18	(comfortable, home)	1
19	(home, and)	1

```
In [77]: import matplotlib.pyplot as plt  
df.plot(kind='line',x='single',y='count',color='blue')  
plt.show()
```



```
In [78]: v=a_bigramfd.most_common(20)  
m=dict(v)  
m
```

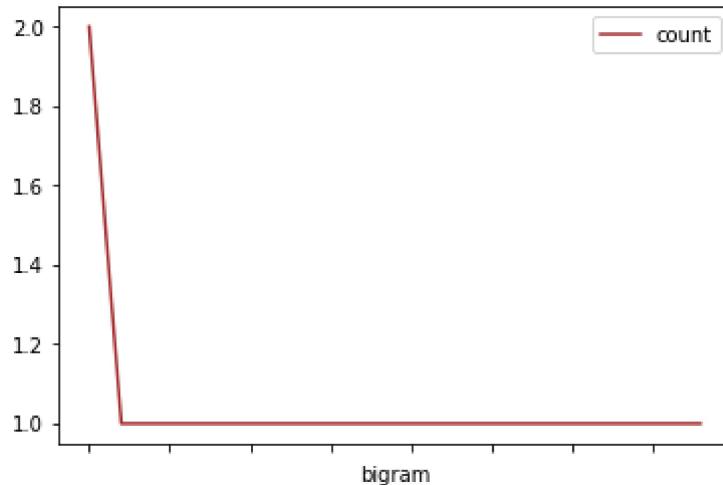
```
Out[78]: {('1816]', 'volume'): 1,  
          ('[emma', 'by'): 1,  
          ('a', 'comfortable'): 1,  
          ('and', 'rich,'): 1,  
          ('austen', '1816']): 1,  
          ('by', 'jane'): 1,  
          ('chapter', 'i'): 1,  
          ('clever,', 'and'): 1,  
          ('comfortable', 'home'): 1,  
          ('emma', 'woodhouse,'): 1,  
          ('handsome,', 'clever,'): 1,  
          ('home', 'and'): 1,  
          ('i', 'chapter'): 1,  
          ('i', 'emma'): 1,  
          ('jane', 'austen'): 1,  
          ('of', 'the'): 2,  
          ('rich,', 'with'): 1,  
          ('volume', 'i'): 1,  
          ('with', 'a'): 1,  
          ('woodhouse,', 'handsome,'): 1}
```

```
In [79]: df2=pd.DataFrame(list(m.items()))
df2.columns=['bigram','count']
df2
```

Out[79]:

	bigram	count
0	(of, the)	2
1	(emma, by)	1
2	(by, jane)	1
3	(jane, austen)	1
4	(austen, 1816])	1
5	(1816], volume)	1
6	(volume, i)	1
7	(i, chapter)	1
8	(chapter, i)	1
9	(i, emma)	1
10	(emma, woodhouse,)	1
11	(woodhouse,, handsome,)	1
12	(handsome,, clever,)	1
13	(clever,, and)	1
14	(and, rich,)	1
15	(rich,, with)	1
16	(with, a)	1
17	(a, comfortable)	1
18	(comfortable, home)	1
19	(home, and)	1

```
In [80]: df2.plot(kind='line',x='bigram',y='count',color='brown')
plt.show()
```



```
In [81]: so_count=a_tokfd['so']
print(so_count)
tot=len(a_tokfd)
print(tot)
rel_freq=so_count/tot
rel_freq
```

0

66

Out[81]: 0.0

```
In [82]: ab.most_common(20)
```

```
Out[82]: [(['so much',), 95),
          ('so very',), 76),
          ('so well',), 30),
          ('so many',), 27),
          ('so long',), 27),
          ('so little',), 20),
          ('so far',), 17),
          ('so I',), 14),
          ('so kind',), 13),
          ('so good',), 12),
          ('so often',), 10),
          ('so soon',), 9),
          ('so great',), 8),
          ('so to',), 7),
          ('so fond',), 7),
          ('so she',), 7),
          ('so it',), 6),
          ('so anxious',), 6),
          ('so as',), 6),
          ('so you',), 6)]
```

```
In [83]: ab_dict=dict(ab)
ab_dict
('so active',): 1,
('so afraid',): 2,
('so agitated',): 1,
('so always',): 1,
('so amiable',): 1,
('so amusing',): 1,
('so and',): 1,
('so angry',): 1,
('so anxious',): 6,
('so anxiously',): 1,
('so apparent',): 1,
('so apt',): 1,
('so arranged',): 1,
('so artlessly',): 1,
('so as',): 6,
('so astonished',): 1,
('so at',): 3,
('so attached',): 1,
('so attentive',): 3,
('so avowed',): 1,
```

```
In [84]:
```

```
tot_occ=len(ab_dict)
tot_occ
```

```
Out[84]: 326
```

```
In [85]: for i,j in ab_dict.items():
    if i=='so much':
        print(i,j)
        print(j/tot_occ)

('so much',) 95
0.29141104294478526
```

```
In [86]: for i,j in ab_dict.items():
    if i=='so will':
        print(i,j)
        print(j/tot_occ)

('so will',) 1
0.003067484662576687
```

```
In [ ]:
```

```
In [ ]:
```

