NAME : RETHINAGIRI G

ROLL NO : 225229130

COURSE TITLE : Natural Language Pre-processing Lab

COURSE TITLE : Natural Language Pre-processing Lab

LAB 02 -Computing Bigram Frequencies

EXERCISE-1 : Process simple bigram data file

Step 1: Open the file, count_2w.txt

In [1]:
```
file=open("C:\\Users\\user\\Downloads\\count_2w.txt")
lines=file.readlines()
```

In [2]:
```
lines
```

Out[2]:
```
['0Uplink verified\t523545\n',
 '0km to\t116103\n',
 '1000s of\t939476\n',
 '100s of\t539389\n',
 '100th anniversary\t158621\n',
 '10am to\t376141\n',
 '10th and\t183715\n',
 '10th anniversary\t242830\n',
 '10th century\t117755\n',
 '10th grade\t174046\n',
 '10th in\t107194\n',
 '10th of\t277970\n',
 '11am to\t127624\n',
 '11th and\t178884\n',
 '11th century\t168601\n',
 '11th grade\t126301\n',
 '11th of\t189501\n',
 '125Mbps w\t108645\n',
 '12th and\t136706\n',
```

Step2 : Build goog2w_list

In [3]:
```
mini=lines[:10]
mini
```

Out[3]:
```
['0Uplink verified\t523545\n',
 '0km to\t116103\n',
 '1000s of\t939476\n',
 '100s of\t539389\n',
 '100th anniversary\t158621\n',
 '10am to\t376141\n',
 '10th and\t183715\n',
 '10th anniversary\t242830\n',
 '10th century\t117755\n',
 '10th grade\t174046\n']
```

In [4]:
```
mini[0]
```

Out[4]:
```
'0Uplink verified\t523545\n'
```

In [5]:
```
mini[0].split()
```

Out[5]:
```
['0Uplink', 'verified', '523545']
```

In [6]:
```
mini_list=[]
for m in mini:
    (w1,w2,count)=m.split()
    count=int(count)
    mini_list.append(((w1,w2),count))
```

In [7]:
```
mini_list
```

Out[7]:
```
[(('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 [8]:
```
mini_list[0]
```

Out[8]:
```
(('0Uplink', 'verified'), 523545)
```

Step 3 : Build goog2w_fd

In [9]:
```
import nltk
```

In [10]:
```
goog2w_fd=nltk.FreqDist()
```

In [11]:
```
for y in lines:
    w1,w2,count=y.split()
    goog2w_fd[(w1,w2)]=count
```

In [12]:
```
goog2w_fd[('of','the')]
```

Out[12]:
```
'2766332391'
```

In [13]:
```
goog2w_fd[('so','beautiful')]
```

Out[13]:
```
'612472'
```

In [14]:
```
goog2w_fd[10]
```

Out[14]:
```
0
```

Step 4 : Explore

Top 10 bigrams

In [15]:
```
goog2w_fd.most_common(10)
```

Out[15]:
```
[(('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')]
```

Top so-initial bigrams

In [16]:
```
g=[]
for h in lines:
    (w1,w2,count)=h.split()
    count=int(count)
    if w1=='so':
        g.append(((w1,w2),count))
```

In [17]:
```
g
```

Out[17]:
```
[(('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),
```

Step 5: pickle the data

In [18]:
```
import pickle
with open('goog2w_list.pkl','wb') as handle:
    pickle.dump(goog2w_fd,handle,protocol=pickle.HIGHEST_PROTOCOL)
```

Excercise-2 : Frequency distribution from Jane Austen Novels

In [19]:
```
with open("C:\\Users\\user\\Downloads\\gutenberg\\gutenberg\\austen-emma.txt") as a:
    ase=a.read()
```

In [20]:
```
with open("C:\\Users\\user\\Downloads\\gutenberg\\gutenberg\\austen-persuasion.txt")as b:
    aspe=b.read()
```

In [21]:
```
c=open("C:/Users/user/Downloads/gutenberg/gutenberg/austen-sense.txt")
asse=c.read()
```

In [22]:
```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\user\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[22]:
```
True
```

In [23]:
```
from nltk.tokenize import sent_tokenize as snt
```

In [24]:
```
a=snt(ase)
a
```

```
ally reach, and many a long October and November evening must\nbe struggled through at Hartfield, before Christmas brought the next\nvisit f
rom Isabella and her husband, and their little children,\nto fill the house, and give her pleasant society again.',
 'Highbury, the large and populous village, almost amounting to a town,\nto which Hartfield, in spite of its separate lawn, and shrubberie
s,\nand name, did really belong, afforded her no equals.',
 'The Woodhouses\nwere first in consequence there.',
 'All looked up to them.',
 'She had\nmany acquaintance in the place, for her father was universally civil,\nbut not one among them who could be accepted in lieu of Mi
ss\nTaylor for even half a day.',
 'It was a melancholy change; and Emma\ncould not but sigh over it, and wish for impossible things,\ntill her father awoke, and made it nece
ssary to be cheerful.',
 'His spirits required support.',
 'He was a nervous man, easily depressed;\nfond of every body that he was used to, and hating to part with them;\nhating change of every kin
d.',
 "Matrimony, as the origin of change,\nwas always disagreeable; and he was by no means yet reconciled\nto his own daughter's marrying, nor c
ould ever speak of her but\nwith compassion, though it had been entirely a match of affection,\nwhen he was now obliged to part with Miss Ta
ylor too; and from\nhis habits of gentle selfishness, and of being never able to\nsuppose that other people could feel differently from hims
elf,\nhe was very much disposed to think Miss Taylor had done as sad\na thing for herself as for them, and would have been a great deal\nhap
pier if she had spent all the rest of her life at Hartfield.",
 'Emma smiled and chatted as cheerfully as she could, to keep him\nfrom such thoughts; but when tea came, it was impossible for him\nnot to
say exactly as he had said at dinner,\n\n"Poor Miss Taylor!--I wish she were here again.',
```

In [25]:
```
b=snt(aspe)
b
```

```
n of\nhimself and his family, these words, after the date of Mary\'s birth--\n"Married, December 16, 1810, Charles, son and heir of Charles \nMusgrove, Esq.',
 'of Uppercross, in the county of Somerset,"\nand by inserting most accurately the day of the month on which\nhe had lost his wife.',
 'Then followed the history and rise of the ancient and respectable family,\nin the usual terms; how it had been first settled in Cheshir e;\nhow mentioned in Dugdale, serving the office of high sheriff,\nrepresenting a borough in three successive parliaments,\nexertions of loy alty, and dignity of baronet, in the first year\nof Charles II, with all the Marys and Elizabeths they had married;\nforming altogether two handsome duodecimo pages, and concluding with\nthe arms and motto:--"Principal seat, Kellynch Hall, in the county\nof Somerset," and Sir Wal ter\'s handwriting again in this finale:--\n\n"Heir presumptive, William Walter Elliot, Esq., great grandson of\nthe second Sir Walter."',
 "Vanity was the beginning and the end of Sir Walter Elliot's character;\nvanity of person and of situation.",
 'He had been remarkably handsome\nin his youth; and, at fifty-four, was still a very fine man.',
 'Few women could think more of their personal appearance than he did,\nnor could the valet of any new made lord be more delighted with\nthe place he held in society.',
 'He considered the blessing of beauty\nas inferior only to the blessing of a baronetcy; and the Sir Walter Elliot,\nwho united these gifts, was the constant object of his warmest respect\nand devotion.',
 'His good looks and his rank had one fair claim on his attachment;\nsince to them he must have owed a wife of very superior character\nto a ny thing deserved by his own.',
 'Lady Elliot had been an excellent woman,\nsensible and amiable; whose judgement and conduct, if they might be\npardoned the youthful infat uation which made her Lady Elliot,\nhad never required indulgence afterwards.--She had humoured,\nor softened, or concealed his failings, an d promoted his real\nrespectability for seventeen years; and though not the very happiest\nbeing in the world herself, had found enough in h er duties, her friends,\nand her children, to attach her to life, and make it no matter of\nindifference to her when she was called on to qu
```

In [26]:
```
c=snt(asse)
c
```

```
lived in so respectable a manner as to engage\nthe general good opinion of their surrounding acquaintance.',
 'The late owner of this estate was a single man, who lived\nto a very advanced age, and who for many years of his life,\nhad a constant com panion and housekeeper in his sister.',
 'But her death, which happened ten years before his own,\nproduced a great alteration 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 Gentleman's days were\ncomfortably spent.",
 'His attachment to them all increased.',
 'The constant attention of Mr. and Mrs. Henry Dashwood\nto his wishes, which proceeded not merely from interest,\nbut from goodness of hear t, gave him every degree of solid\ncomfort which his age could receive; and the cheerfulness\nof the children added a relish to his existenc e.',
 'By a former marriage, Mr. Henry Dashwood had one\nson: by his present lady, three daughters.',
 'The son,\na steady respectable young man, was amply provided\nfor by the fortune of his mother, which had been large,\nand half of which d evolved on him on his coming of age.',
 'By his own marriage, likewise, which happened soon afterwards,\nhe added to his wealth.',
 "To him therefore the succession\nto the Norland estate was not so really important as to\nhis sisters; for their fortune, independent of w hat might\narise to them from their father's inheriting that property,\ncould be but small.",
 'Their mother had nothing, and their\nfather only seven thousand pounds in his own disposal;\nfor the remaining moiety of his first wife's fortune was\nalso secured to her child, and he had only a life-interest\nin it.',
 'The old gentleman died; his will was read, and\nlike almost every other will, gave as much disappointment\nas pleasure.',
```

In [27]:
```
len(a)
```

Out[27]:
```
7456
```

In [28]:
```
len(b)
```

Out[28]:
```
3654
```

In [29]:
```
len(c)
```

Out[29]:
```
4833
```

In [30]:
```
from nltk.tokenize import word_tokenize as wt
```

In [31]:
```
w1=wt(ase)
w1
```

Out[31]:
```
['[',
 'Emma',
 'by',
 'Jane',
 'Austen',
 '1816',
 ']',
 'VOLUME',
 'I',
 'CHAPTER',
 'I',
 'Emma',
 'Woodhouse',
 ',',
 'handsome',
 ',',
 'clever',
 ',',
 'and',
```

In [32]:
```
w2=wt(aspe)
w2
```

Out[32]:
```
['[',
 'Persuasion',
 'by',
 'Jane',
 'Austen',
 '1818',
 ']',
 'Chapter',
 '1',
 'Sir',
 'Walter',
 'Elliot',
 ',',
 'of',
 'Kellynch',
 'Hall',
 ',',
 'in',
 'Somersetshire',
```

In [33]:
```
w3=wt(asse)
w3
```

Out[33]:
```
['[',
 'Sense',
 'and',
 'Sensibility',
 'by',
 'Jane',
 'Austen',
 '1811',
 ']',
 'CHAPTER',
 '1',
 'The',
 'family',
 'of',
 'Dashwood',
 'had',
 'long',
 'been',
 'settled',
```

In [34]:
```
from nltk import *
```

In [35]:
```
c1=FreqDist(w1)
c1
```

Out[35]:
```
FreqDist({',': 12016, '.': 6346, 'to': 5125, 'the': 4844, 'and': 4653, 'of': 4272, 'I': 3177, '--': 3100, 'a': 3001, "'": 2454, ...})
```

In [36]:
```
c2=FreqDist(w2)
c2
```

Out[36]:
```
FreqDist({',': 7024, 'the': 3119, '.': 3119, 'to': 2751, 'and': 2724, 'of': 2562, 'a': 1528, 'in': 1340, 'was': 1330, ';': 1319, ...})
```

In [37]:
```
c3=FreqDist(w1)
c3
```

Out[37]:
```
FreqDist({',': 12016, '.': 6346, 'to': 5125, 'the': 4844, 'and': 4653, 'of': 4272, 'I': 3177, '--': 3100, 'a': 3001, "'": 2454, ...})
```

In [38]:
```
tk1 = nltk.word_tokenize(ase.lower())
tk1
```

Out[38]:
```
['[',
 'emma',
 'by',
 'jane',
 'austen',
 '1816',
 ']',
 'volume',
 'i',
 'chapter',
 'i',
 'emma',
 'woodhouse',
 ',',
 'handsome',
 ',',
 'clever',
 ',',
 'and',
```

In [39]:
```
tk2 = nltk.word_tokenize(aspe.lower())
tk2
```

Out[39]:
```
['[',
 'persuasion',
 'by',
 'jane',
 'austen',
 '1818',
 ']',
 'chapter',
 '1',
 'sir',
 'walter',
 'elliot',
 ',',
 'of',
 'kellynch',
 'hall',
 ',',
 'in',
 'somersetshire',
```

In [40]:
```
tk3 = nltk.word_tokenize(asse.lower())
tk3
```

```
 'generations',
 ',',
 'they',
 'had',
 'lived',
 'in',
 'so',
 'respectable',
 'a',
 'manner',
 'as',
 'to',
 'engage',
 'the',
 'general',
 'good',
 'opinion',
 'of',
 'their',
 'surrounding',
```

In [41]:
```
s1 = sorted(set(tk1))
s1
```

Out[41]:
```
['!',
 '&',
 "'",
 "''",
 "'d",
 "'s",
 "'t",
 "'ye",
 '(',
 ')',
 ',',
 '--',
 '.',
 '10,000',
 '1816',
 '23rd',
 '24th',
 '26th',
 '28th',
```

In [42]:
```
s2 = sorted(set(tk2))
s2
```

Out[42]:
```
['!',
 '&',
 "'",
 "''",
 "'ll",
 "'s",
 "'squire",
 "'ye",
 '(',
 ')',
 ',',
 '-',
 '--',
 '-5',
 '.',
 '1',
 '10',
 '11',
 '12',
```

In [43]:
```
s3 = sorted(set(tk3))
s3
```

Out[43]:
```
['!',
 '&',
 "'",
 "''",
 "'but",
 "'consider",
 "'d",
 "'do",
 "'em",
 "'for",
 "'is",
 "'it",
 "'la",
 "'ll",
 "'lord",
 "'m",
 "'making",
 "'my",
 "'prenticed",
```

In [44]:
```
c1.most_common(50)
```

Out[44]:
```
[(',', 12016),
 ('.', 6346),
 ('to', 5125),
 ('the', 4844),
 ('and', 4653),
 ('of', 4272),
 ('I', 3177),
 ('--', 3100),
 ('a', 3001),
 ("'", 2454),
 ('was', 2383),
 ('her', 2360),
 (';', 2353),
 ('not', 2242),
 ('in', 2103),
 ('it', 2103),
 ('be', 1965),
 ('she', 1774),
 ('``', 1733),
 ('that', 1729),
 ('you', 1664),
 ('had', 1605),
 ('as', 1387),
 ('he', 1365),
 ('for', 1320),
 ('have', 1301),
 ('is', 1221),
 ('with', 1185),
 ('very', 1148),
 ('but', 1148),
 ('Mr.', 1091),
 ('his', 1084),
 ('!', 1063),
 ('at', 996),
 ('so', 918),
 ("'s", 866),
 ('Emma', 855),
 ('all', 831),
 ('could', 823),
 ('would', 813),
 ('been', 755),
 ('him', 748),
 ('on', 674),
 ('Mrs.', 668),
 ('any', 651),
 ('?', 621),
 ('my', 619),
 ('no', 616),
 ('Miss', 592),
 ('were', 590)]
```

In [45]:
```
c2.most_common(50)
```

Out[45]:
```
[(',', 7024),
 ('the', 3119),
 ('.', 3119),
 ('to', 2751),
 ('and', 2724),
 ('of', 2562),
 ('a', 1528),
 ('in', 1340),
 ('was', 1330),
 (';', 1319),
 ('had', 1177),
 ('her', 1158),
 ('I', 1123),
 ('not', 968),
 ('be', 949),
 ("'", 912),
 ('it', 857),
 ('that', 853),
 ('she', 819),
 ('as', 787),
 ('he', 736),
 ('for', 695),
 ('``', 652),
 ('with', 643),
 ('his', 625),
 ('have', 583),
 ('but', 553),
 ('you', 548),
 ('at', 519),
 ('all', 517),
 ('Anne', 496),
 ('been', 496),
 ('him', 467),
 ("'s", 464),
 ('could', 444),
 ('were', 426),
 ('very', 425),
 ('which', 415),
 ('by', 409),
 ('is', 393),
 ('on', 386),
 ('would', 351),
 ('so', 338),
 ('She', 327),
 ('they', 323),
 ('!', 318),
 ('no', 309),
 ('Captain', 297),
 ('Mrs', 291),
 ('from', 290)]
```

In [46]:
```
c3.most_common(50)
```

Out[46]:
```
[(',', 12016),
 ('.', 6346),
 ('to', 5125),
 ('the', 4844),
 ('and', 4653),
 ('of', 4272),
 ('I', 3177),
 ('--', 3100),
 ('a', 3001),
 ("'", 2454),
 ('was', 2383),
 ('her', 2360),
 (';', 2353),
 ('not', 2242),
 ('in', 2103),
 ('it', 2103),
 ('be', 1965),
 ('she', 1774),
 ('``', 1733),
 ('that', 1729),
 ('you', 1664),
 ('had', 1605),
 ('as', 1387),
 ('he', 1365),
 ('for', 1320),
 ('have', 1301),
 ('is', 1221),
 ('with', 1185),
 ('very', 1148),
 ('but', 1148),
 ('Mr.', 1091),
 ('his', 1084),
 ('!', 1063),
 ('at', 996),
 ('so', 918),
 ("'s", 866),
 ('Emma', 855),
 ('all', 831),
 ('could', 823),
 ('would', 813),
 ('been', 755),
 ('him', 748),
 ('on', 674),
 ('Mrs.', 668),
 ('any', 651),
 ('?', 621),
 ('my', 619),
 ('no', 616),
 ('Miss', 592),
 ('were', 590)]
```

Exercise 3

In [47]:
```
print(ase)
```

```
[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 too long ago for her to have more than an indistinct
remembrance of her caresses; and her place had been supplied
by an excellent woman as governess, who had fallen little short
of a mother in affection.
```

In [48]:
```
tokenizer=nltk.tokenize.WhitespaceTokenizer()
toke=tokenizer.tokenize(ase)
print(toke)
```

```
end', 'and', 'companion', 'such', 'as', 'few', 'possessed:', 'intelligent,', 'well-informed,', 'useful,', 'gentle,', 'knowing', 'all', 'th
e', 'ways', 'of', 'the', 'family,', 'interested', 'in', 'all', 'its', 'concerns,', 'and', 'peculiarly', 'interested', 'in', 'herself,', 'i
n', 'every', 'pleasure,', 'every', 'scheme', 'of', 'hers--one', 'to', 'whom', 'she', 'could', 'speak', 'every', 'thought', 'as', 'it', 'aros
e,', 'and', 'who', 'had', 'such', 'an', 'affection', 'for', 'her', 'as', 'could', 'never', 'find', 'fault.', 'How', 'was', 'she', 'to', 'bea
r', 'the', 'change?--It', 'was', 'true', 'that', 'her', 'friend', 'was', 'going', 'only', 'half', 'a', 'mile', 'from', 'them;', 'but', 'Emm
a', 'was', 'aware', 'that', 'great', 'must', 'be', 'the', 'difference', 'between', 'a', 'Mrs.', 'Weston,', 'only', 'half', 'a', 'mile', 'fro
m', 'them,', 'and', 'a', 'Miss', 'Taylor', 'in', 'the', 'house;', 'and', 'with', 'all', 'her', 'advantages,', 'natural', 'and', 'domestic,',
'she', 'was', 'now', 'in', 'great', 'danger', 'of', 'suffering', 'from', 'intellectual', 'solitude.', 'She', 'dearly', 'loved', 'her', 'fath
er,', 'but', 'he', 'was', 'no', 'companion', 'for', 'her.', 'He', 'could', 'not', 'meet', 'her', 'in', 'conversation,', 'rational', 'or', 'p
layful.', 'The', 'evil', 'of', 'the', 'actual', 'disparity', 'in', 'their', 'ages', '(and', 'Mr.', 'Woodhouse', 'had', 'not', 'married', 'ea
rly)', 'was', 'much', 'increased', 'by', 'his', 'constitution', 'and', 'habits;', 'for', 'having', 'been', 'a', 'valetudinarian', 'all', 'hi
s', 'life,', 'without', 'activity', 'of', 'mind', 'or', 'body,', 'he', 'was', 'a', 'much', 'older', 'man', 'in', 'ways', 'than', 'in', 'year
s;', 'and', 'though', 'everywhere', 'beloved', 'for', 'the', 'friendliness', 'of', 'his', 'heart', 'and', 'his', 'amiable', 'temper,', 'hi
s', 'talents', 'could', 'not', 'have', 'recommended', 'him', 'at', 'any', 'time.', 'Her', 'sister,', 'though', 'comparatively', 'but', 'litt
le', 'removed', 'by', 'matrimony,', 'being', 'settled', 'in', 'London,', 'only', 'sixteen', 'miles', 'off,', 'was', 'much', 'beyond', 'her',
'daily', 'reach;', 'and', 'many', 'a', 'long', 'October', 'and', 'November', 'evening', 'must', 'be', 'struggled', 'through', 'at', 'Hartfie
ld,', 'before', 'Christmas', 'brought', 'the', 'next', 'visit', 'from', 'Isabella', 'and', 'her', 'husband,', 'and', 'their', 'little', 'chi
ldren,', 'to', 'fill', 'the', 'house,', 'and', 'give', 'her', 'pleasant', 'society', 'again.', 'Highbury,', 'the', 'large', 'and', 'populou
s', 'village,', 'almost', 'amounting', 'to', 'a', 'town,', 'to', 'which', 'Hartfield,', 'in', 'spite', 'of', 'its', 'separate', 'lawn,', 'an
```

In [49]:
```
tokenizer=nltk.tokenize.WhitespaceTokenizer()
tok=tokenizer.tokenize(ase.lower())
print(tok)
```

```
['[emma', 'by', 'jane', 'austen', '1816]', 'volume', 'i', 'chapter', 'i', 'emma', 'woodhouse,', 'handsome,', 'clever,', 'and', 'rich,', 'wit
h', 'a', 'comfortable', 'home', 'and', 'happy', 'disposition,', 'seemed', 'to', 'unite', 'some', 'of', 'the', 'best', 'blessings', 'of', 'ex
istence;', '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', 'ea
rly', 'period.', 'her', 'mother', 'had', 'died', 'too', 'long', 'ago', 'for', 'her', 'to', 'have', 'more', 'than', 'an', 'indistinct', 'reme
mbrance', 'of', 'her', 'caresses;', 'and', 'her', 'place', 'had', 'been', 'supplied', 'by', 'an', 'excellent', 'woman', 'as', 'governess,',
'who', 'had', 'fallen', 'little', 'short', 'of', 'a', 'mother', 'in', 'affection.', 'sixteen', 'years', 'had', 'miss', 'taylor', 'been', 'i
n', 'mr.', "woodhouse's", 'family,', 'less', 'as', 'a', 'governess', 'than', 'a', 'friend,', 'very', 'fond', 'of', 'both', 'daughters,', 'bu
t', 'particularly', 'of', 'emma.', 'between', '_them_', 'it', 'was', 'more', 'the', 'intimacy', 'of', 'sisters.', 'even', 'before', 'miss',
'taylor', 'had', 'ceased', 'to', 'hold', 'the', 'nominal', 'office', 'of', 'governess,', 'the', 'mildness', 'of', 'her', 'temper', 'had', 'h
ardly', 'allowed', 'her', 'to', 'impose', 'any', 'restraint;', 'and', 'the', 'shadow', 'of', 'authority', 'being', 'now', 'long', 'passed',
'away,', 'they', 'had', 'been', 'living', 'together', 'as', 'friend', 'and', 'friend', 'very', 'mutually', 'attached,', 'and', 'emma', 'doin
g', 'just', 'what', 'she', 'liked;', 'highly', 'esteeming', 'miss', "taylor's", 'judgment,', 'but', 'directed', 'chiefly', 'by', 'her', 'ow
n.', 'the', 'real', 'evils,', 'indeed,', 'of', "emma's", 'situation', 'were', 'the', 'power', 'of', 'having', 'rather', 'too', 'much', 'he
r', 'own', 'way,', 'and', 'a', 'disposition', 'to', 'think', 'a', 'little', 'too', 'well', 'of', 'herself;', 'these', 'were', 'the', 'disadv
antages', 'which', 'threatened', 'alloy', 'to', 'her', 'many', 'enjoyments.', 'the', 'danger,', 'however,', 'was', 'at', 'present', 'so', 'u
nperceived,', 'that', 'they', 'did', 'not', 'by', 'any', 'means', 'rank', 'as', 'misfortunes', 'with', 'her.', 'sorrow', 'came--a', 'gentl
e', 'sorrow--but', 'not', 'at', 'all', 'in', 'the', 'shape', 'of', 'any', 'disagreeable', 'consciousness.--miss', 'taylor', 'married.', 'i
```

In [50]:
```
z = FreqDist(tok)
z
```

Out[50]:
```
FreqDist({'the': 5120, 'to': 5079, 'and': 4445, 'of': 4196, 'a': 3055, 'i': 2602, 'was': 2302, 'she': 2169, 'in': 2091, 'not': 2028, ...})
```

In [51]:
```
b2=list(nltk.bigrams(tok))
tb1=nltk.FreqDist(b2)
tb1
```

Out[51]:
```
FreqDist({('to', 'be'): 566, ('of', 'the'): 558, ('in', 'the'): 441, ('it', 'was'): 387, ('she', 'had'): 313, ('i', 'am'): 302, ('she', 'wa
s'): 301, ('had', 'been'): 299, ('could', 'not'): 271, ('to', 'the'): 236, ...})
```

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

In [53]:
```
w=z.most_common(20)
m=dict(w)
m
```

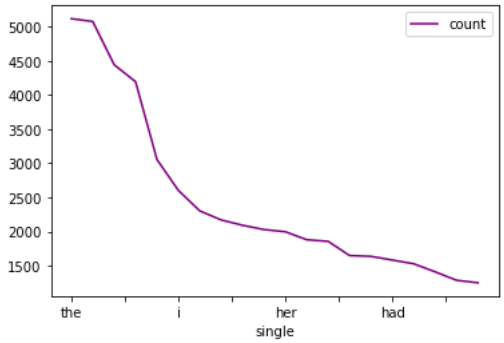
Out[53]:
```
{'the': 5120,
 'to': 5079,
 'and': 4445,
 'of': 4196,
 'a': 3055,
 'i': 2602,
 'was': 2302,
 'she': 2169,
 'in': 2091,
 'not': 2028,
 'her': 1996,
 'be': 1879,
 'it': 1855,
 'that': 1647,
 'he': 1635,
 'had': 1581,
 'you': 1526,
 'as': 1408,
 'have': 1284,
 'for': 1248}
```

In [54]:
```
import pandas as pd
df=pd.DataFrame(list(m.items()))
df.columns=['single','count']
df
```

Out[54]:

	single	count
0	the	5120
1	to	5079
2	and	4445
3	of	4196
4	a	3055
5	i	2602
6	was	2302
7	she	2169
8	in	2091
9	not	2028
10	her	1996
11	be	1879
12	it	1855
13	that	1647
14	he	1635
15	had	1581
16	you	1526
17	as	1408
18	have	1284
19	for	1248

In [55]:
```
import matplotlib.pyplot as plt
df.plot(kind='line',x='single',y='count',color='purple')
plt.show()
```



In [56]:
```
t=tb1.most_common(20)
n=dict(t)
n
```

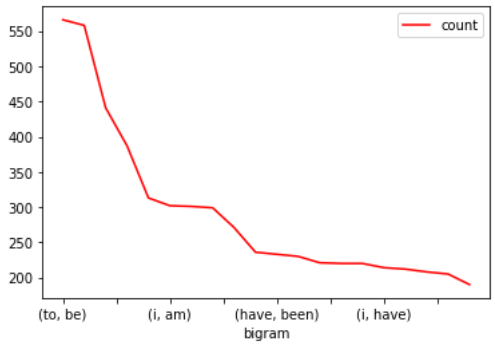
Out[56]:
```
{('to', 'be'): 566,
 ('of', 'the'): 558,
 ('in', 'the'): 441,
 ('it', 'was'): 387,
 ('she', 'had'): 313,
 ('i', 'am'): 302,
 ('she', 'was'): 301,
 ('had', 'been'): 299,
 ('could', 'not'): 271,
 ('to', 'the'): 236,
 ('have', 'been'): 233,
 ('of', 'her'): 230,
 ('it', 'is'): 221,
 ('he', 'had'): 220,
 ('do', 'not'): 220,
 ('i', 'have'): 214,
 ('and', 'the'): 212,
 ('would', 'be'): 208,
 ('he', 'was'): 205,
 ('such', 'a'): 190}
```

In [57]:
```
df2=pd.DataFrame(list(n.items()))
df2.columns=['bigram','count']
df2
```

Out[57]:

	bigram	count
0	(to, be)	566
1	(of, the)	558
2	(in, the)	441
3	(it, was)	387
4	(she, had)	313
5	(i, am)	302
6	(she, was)	301
7	(had, been)	299
8	(could, not)	271
9	(to, the)	236
10	(have, been)	233
11	(of, her)	230
12	(it, is)	221
13	(he, had)	220
14	(do, not)	220
15	(i, have)	214
16	(and, the)	212
17	(would, be)	208
18	(he, was)	205
19	(such, a)	190

In [58]:
```
df2.plot(kind='line',x='bigram',y='count',color='red')
plt.show()
```



In [59]:
```
so_count=z['so']
print(so_count)
tot=len(z)
print(tot)
rel_freq=so_count/tot
rel_freq
```

```
843
16945
```

Out[59]:
```
0.04974918855119504
```

In [60]:
```
words=re.findall(r'so+ \w+',open("C:\\Users\\user\\Downloads\\gutenberg\\gutenberg\\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 sh
e',): 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 fe
w',): 4, ('so that',): 4, ('so strong',): 4, ('so liberal',): 4, ('so miserable',): 4, ('so happily',): 3, ('so proper',): 3, ('so pleasantl
y',): 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 obli
ged',): 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 surprized',): 2, ('so early',): 2, ('so too',): 2, ('so easy',): 2, ('so decidedly',): 2, ('s
o absolutely',): 2, ('so particular',): 2, ('so deceived',): 2, ('so palpably',): 2, ('so clever',): 2, ('so short',): 2, ('so cold',): 2, ('s
o high',): 2, ('so happened',): 2, ('so full',): 2, ('so thoroughly',): 2, ('so equal',): 2, ('so off',): 2, ('so naturally',): 2, ('so afrai
d',): 2, ('so deep',): 2, ('so kindly',): 2, ('so pale',): 2, ('so noble',): 2, ('so lovely',): 2, ('so mad',): 2, ('so nearly',): 2, ('so sor
ry',): 2, ('so cheerful',): 2, ('so unfeeling',): 2, ('so ready',): 2, ('so unperceived',): 1, ('so mild',): 1, ('so constantly',): 1, ('so co
mfortably',): 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 slig
htingly',): 1, ('so by',): 1, ('so loudly',): 1, ('so materially',): 1, ('so hard',): 1, ('so delightful',): 1, ('so pointed',): 1, ('so equal
led',): 1, ('so evidently',): 1, ('so immediately',): 1, ('so sought',): 1, ('so excellent',): 1, ('so prettily',): 1, ('so extreme',): 1, ('s
o 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 overcharged',): 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 misinterpre
ted',): 1, ('so active',): 1, ('so pointedly',): 1, ('so striking',): 1, ('so sudden',): 1, ('so industriously',): 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 deserving',): 1, ('so disappointed',): 1, ('so ended',): 1, ('so sluggis
h',): 1, ('so amiable',): 1, ('so quiet',): 1, ('so idolized',): 1, ('so cried',): 1, ('so acceptable',): 1, ('so properly',): 1, ('so reasona
ble',): 1, ('so delightfully',): 1, ('so rich',): 1, ('so warm',): 1, ('so large',): 1, ('so handsomely',): 1, ('so abundant',): 1, ('so outre
e',): 1, ('so thoughtful',): 1, ('so must',): 1, ('so effectually',): 1, ('so beautiful',): 1, ('so Patty',): 1, ('so honoured',): 1, ('so clo
se',): 1, ('so imprudent',): 1, ('so limited',): 1, ('so from',): 1, ('so amusing',): 1, ('so indifferent',): 1, ('so indignant',): 1, ('so sa
id',): 1, ('so right',): 1, ('so wretched',): 1, ('so now',): 1, ('so occupied',): 1, ('so unhappy',): 1, ('so highly',): 1, ('so generall
y',): 1, ('so exactly',): 1, ('so double',): 1, ('so secluded',): 1, ('so regular',): 1, ('so determined',): 1, ('so motherly',): 1, ('so th
e',): 1, ('so glibly',): 1, ('so calculated',): 1, ('so thrown',): 1, ('so exclusively',): 1, ('so disgustingly',): 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 movin
g',): 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 unea
sy',): 1, ('so wonderful',): 1, ('so _very_',): 1, ('so expressly',): 1, ('so angry',): 1, ('so anxiously',): 1, ('so strange',): 1, ('so stou
tly',): 1, ('so mistake',): 1, ('so mistaken',): 1, ('so dreadfully',): 1, ('so voluntarily',): 1, ('so satisfactory',): 1, ('so disintereste
d',): 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 su
ffered',): 1})
```

In [61]:
```
ab_dict=dict(ab)
ab_dict
```

Out[61]:
```
{('so unperceived',): 1,
 ('so far',): 17,
 ('so obliged',): 2,
 ('so mild',): 1,
 ('so much',): 95,
 ('so to',): 7,
 ('so well',): 30,
 ('so happily',): 3,
 ('so many',): 27,
 ('so long',): 27,
 ('so perfectly',): 2,
 ('so constantly',): 1,
 ('so entirely',): 4,
 ('so comfortably',): 1,
 ('so very',): 76,
 ('so kind',): 13,
 ('so avowed',): 1,
 ('so dear',): 2,
 ('so deservedly',): 1,
```

In [62]:
```
total=len(ab_dict)
total
```

Out[62]:
```
326
```

In [63]:
```
for i,j in ab_dict.items():
    if i==('so much',):
        print(i,j)
        print(j/total)
```

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

In [64]:
```
for i,j in ab_dict.items():
    if i==('so will',):
        print(i,j)
        print(j/total)
```

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