

# Homework 1: nltk尝试

学号：16307130194 姓名：陈中钰

```
In [1]: %matplotlib inline
```

导入nltk库和nltk.book语料：

```
In [2]: import nltk
        #nltk.download()
        from nltk.book import *

*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

导入正则表达式模块：

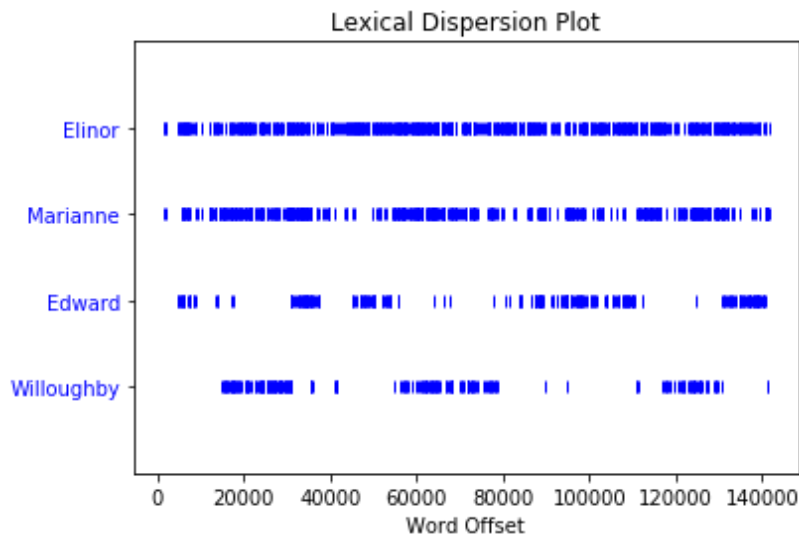
```
In [3]: import re
```

## 问题1

制作text2（《理智与情感》）中四个主角：Elinor，Marianne，Edward和Willoughby的分布图。在这部小说中关于男性和女性所扮演的不同角色，你能观察到什么？能找出一对夫妻吗？

a. 使用dispersion\_plot()，制作text2中Elinor、Marianne、Edward和Willoughby的分布图：

```
In [4]: text2.dispersion_plot(['Elinor', 'Marianne', 'Edward', 'Willoughby'])
```



使用FreqDist()对上述4个词进行词频统计：

```
In [5]: fdist2 = FreqDist(text2)
```

```
In [6]: fdist2['Elinor']
```

```
Out[6]: 684
```

```
In [7]: fdist2['Marianne']
```

```
Out[7]: 566
```

```
In [8]: fdist2['Edward']
```

```
Out[8]: 262
```

```
In [9]: fdist2['Willoughby']
```

```
Out[9]: 215
```

b. 能观察到：通过词频统计，可以猜测Elinor是女一号、Marianne是女二号，而Edward为男一号、Willoughby为男二号。而且两位女性的词频比两位男性的词频高出许多，可以认为这部小说主要描述的是这两位女性。

c. 找到夫妻：通过观察4个词的分布图，可以发现，Marianne的密度分布有明显的稀疏、稠密部分，而她的稠密部分与Willoughby的稠密部分恰好对应，可以推断Marianne与Willoughby是一对；而Elinor的稠密部分与Edward的稠密部分也大致吻合，可以猜测Elinor和Edward是一对。

输出Elinor的词语索引视图：

```
In [10]: text2.concordance('Elinor')
```

Displaying 25 of 685 matches:

avoid a breach with their brother . Elinor , this eldest daughter , whose advi  
, in many respects , quite equal to Elinor ' s . She was sensible and clever ;  
d her mother was strikingly great . Elinor saw , with concern , the excess of  
r admitting consolation in future . Elinor , too , was deeply afflicted ; but  
at he loved her daughter , and that Elinor returned the partiality . It was co  
semblance of disposition ; and that Elinor ' s merit should not be acknowledge  
him farther , by a reflection which Elinor chanced one day to make on the diff  
" I think you will like him , " said Elinor , " when you know more of him . " "  
; the persuasion of his regard for Elinor perhaps assisted her penetration ;  
symptom of love in his behaviour to Elinor , than she considered their serious  
, my dear Marianne . " said she , " Elinor will , in all probability be settle  
attract him , and though he admires Elinor ' s drawings very much , it is not  
ust allow for difference of taste . Elinor has not my feelings , and therefore  
!" CHAPTER 4 " What a pity it is , Elinor , " said Marianne , " that Edward sh  
" " No taste for drawing !" replied Elinor , " why should you think so ? He do  
; but the kind of approbation which Elinor described as excited in him by the  
. " I hope , Marianne , " continued Elinor , " you do not consider him as defi  
he replied : " Do not be offended , Elinor , if my praise of him is not in eve  
d amiable . " " I am sure , " replied Elinor , with a smile , " that his dearest  
sense and his goodness , " continued Elinor , " no one can , I think , be in do  
hall very soon think him handsome , Elinor , if I do not now . When you tell m  
ace , than I now do in his heart . " Elinor started at this declaration , and w  
eem him ! Like him ! Cold - hearted Elinor ! Oh ! worse than cold - hearted !  
will leave the room this moment . " Elinor could not help laughing . " Excuse  
elf , how delightful it would be !" Elinor had given her real opinion to her s

## 问题2

在聊天语料库 ( text5 ) 中查找所有以字母v开头的词，按字母顺序显示出来。找出text5中所有3个字母的词。使用频率分布函数 ( FreqDist ) ，以频率从高到低显示这些词。

用set()对text5中的词做去重处理，并用sorted()把词按照字母顺序排序，建立text5的词库：

```
In [11]: tx5 = sorted(set(text5))
```

a. 在text5的词库中，用正则表达式找出以v开头的词，并按字母顺序显示：

```
In [12]: print([w for w in tx5 if re.search('^v', w)])
```

```
['va', 'valiant', 'vamp', 'van', 'vanished', 'veg', 'vega', 'vent', 'very', 'vetrab  
rate', 'vicious', 'victims', 'victors', 'video', 'vigilant', 'village', 'vindictiv  
e', 'vinyl', 'violators', 'virgin', 'virginal', 'virginity', 'virgo', 'visual', 'v  
m', 'vodka', 'voice', 'voices', 'vomiting', 'vote', 'voter', 'vs.', 'vuela']
```

b. 在text5中，用正则表达式找出以v开头的词，并用FreqDist()统计词频，再调用most\_common()以频率从高到低显示这些词：

```
In [13]: print(FreqDist([w for w in text5 if re.search('^v', w)]).most_common())
```

```
[('very', 19), ('virgin', 17), ('video', 7), ('vote', 3), ('vuela', 3), ('virgina  
l', 2), ('vm', 1), ('vindictive', 1), ('vetrbrate', 1), ('village', 1), ('veg',  
1), ('virginity', 1), ('voices', 1), ('va', 1), ('voice', 1), ('victors', 1), ('val  
iant', 1), ('voter', 1), ('violators', 1), ('van', 1), ('vicious', 1), ('vinyl',  
1), ('vodka', 1), ('victims', 1), ('vomiting', 1), ('virgo', 1), ('vigilant', 1),  
( 'vamp', 1), ('vent', 1), ('vanished', 1), ('visual', 1), ('vega', 1), ('vs.', 1)]
```

c. 在text5的词库中，用正则表达式找出是3个字母的词，并按字母顺序显示：

```
In [14]: print([w for w in tx5 if re.search('^[A-Za-z]{3}$', w)])
```

```
['AFK', 'ALL', 'AND', 'ANY', 'ARE', 'ASS', 'Ack', 'Ahh', 'Amy', 'And', 'Any', 'Ar  
e', 'Ark', 'Ask', 'Aww', 'BIG', 'BOY', 'BUT', 'BUT', 'BYE', 'Ben', 'Box', 'Bud', 'B  
ut', 'Bye', 'CAN', 'CDT', 'COM', 'CSI', 'CST', 'CUZ', 'Can', 'Cry', 'Cum', 'DON',  
'DVD', 'Dew', 'Did', 'EST', 'End', 'Fix', 'For', 'GNG', 'Gay', 'Get', 'God', 'Guy',  
'HEY', 'HOT', 'HOW', 'HUH', 'Hay', 'Her', 'Hey', 'Hmm', 'How', 'Hug', 'Huh', 'IRC',  
'Ico', 'Ima', 'Its', 'Ive', 'JRZ', 'JTo', 'Jam', 'Joe', 'Jon', 'LIX', 'LOL', 'LOI',  
'LPN', 'Lay', 'Lee', 'Let', 'Lil', 'LoL', 'Lol', 'MAN', 'MSN', 'Man', 'Max', 'Meh',  
'Men', 'Mmm', 'Mom', 'NOT', 'New', 'Not', 'Now', 'OMG', 'ONE', 'OOH', 'OUT', 'Oak',  
'Ohh', 'Omg', 'Ooo', 'Our', 'Out', 'Own', 'PDT', 'PMS', 'PST', 'Pat', 'Pic', 'PmS',  
'RUN', 'Red', 'Row', 'SAY', 'SEE', 'San', 'Sat', 'See', 'Set', 'She', 'Sun', 'Sup',  
'THE', 'TWO', 'The', 'Tim', 'Too', 'UFC', 'USA', 'Ugh', 'Uhh', 'WHO', 'WHY', 'WOW',  
'WTF', 'Way', 'Web', 'Who', 'Why', 'Wow', 'Wtf', 'YES', 'YOU', 'YoU', 'YaY', 'Yay',  
'Yea', 'Yeh', 'Yes', 'YoU', 'You', 'abs', 'act', 'ads', 'afe', 'afk', 'age', 'ago',  
'ahh', 'aim', 'air', 'aka', 'all', 'alo', 'amy', 'and', 'ans', 'any', 'aok', 'are',  
'arm', 'art', 'ask', 'asl', 'ass', 'ate', 'atl', 'aww', 'bad', 'bag', 'bak', 'ban',  
'bar', 'bay', 'bbl', 'bbs', 'bed', 'beg', 'ben', 'bes', 'bet', 'big', 'bio', 'bit',  
'biz', 'bob', 'boi', 'boo', 'bot', 'bow', 'box', 'boy', 'bra', 'brb', 'bro', 'btw',  
'bug', 'buh', 'bum', 'bus', 'but', 'buy', 'byb', 'bye', 'cal', 'cam', 'can', 'car',  
'cat', 'chp', 'com', 'con', 'cop', 'cos', 'cow', 'cpr', 'cry', 'cup', 'cus', 'cut',  
'cuz', 'cya', 'dad', 'dam', 'dat', 'day', 'dem', 'did', 'die', 'dik', 'dis', 'doc',  
'doe', 'dog', 'dry', 'duh', 'dum', 'dun', 'dya', 'ear', 'eat', 'eay', 'egg', 'ehh',  
'elo', 'end', 'eng', 'ere', 'erm', 'eva', 'eww', 'eye', 'fan', 'far', 'fat', 'fav',  
'fck', 'fee', 'fer', 'few', 'fir', 'fit', 'fix', 'fly', 'for', 'fot', 'frm', 'fun',  
'fwd', 'gal', 'gas', 'gay', 'gee', 'get', 'git', 'god', 'got', 'gtg', 'gum', 'gun',  
'guy', 'hPa', 'had', 'hah', 'has', 'hat', 'hav', 'haw', 'heh', 'hel', 'hep', 'her',  
'hes', 'hey', 'hih', 'hii', 'him', 'hio', 'his', 'hit', 'hix', 'hiy', 'hmm', 'hog',  
'hoi', 'hom', 'hon', 'hoo', 'hop', 'hot', 'how', 'hrs', 'hub', 'hug', 'huh', 'hun',  
'hyy', 'ice', 'ifr', 'ill', 'ima', 'ing', 'irl', 'its', 'itz', 'ive', 'jar', 'job',  
'joy', 'jus', 'ken', 'kep', 'kev', 'key', 'kid', 'kit', 'knw', 'kts', 'lag', 'lap',  
'law', 'lbs', 'leg', 'les', 'let', 'lex', 'lez', 'lie', 'lil', 'lip', 'lix', 'loc',  
'log', 'lol', 'lot', 'low', 'ltr', 'luv', 'mac', 'mad', 'mah', 'man', 'may', 'men',  
'met', 'mhm', 'min', 'mmm', 'mom', 'msg', 'msn', 'muh', 'mum', 'nad', 'nah', 'nap',  
'naw', 'nbc', 'nce', 'new', 'nic', 'noo', 'not', 'now', 'nut', 'nyc', 'oOo', 'odd',  
'off', 'ohh', 'oil', 'old', 'ole', 'omg', 'one', 'ono', 'ooh', 'ooo', 'oot', 'opr',  
'ops', 'our', 'out', 'ovr', 'own', 'pad', 'pal', 'pay', 'per', 'pet', 'pic', 'pie',  
'pit', 'plz', 'pms', 'pop', 'pos', 'pot', 'ppl', 'pro', 'psh', 'put', 'pvt', 'que',  
'rag', 'ran', 'rap', 'red', 'rey', 'rid', 'rob', 'ros', 'rub', 'rum', 'run', 'sad',  
'san', 'sat', 'saw', 'say', 'sea', 'sec', 'see', 'set', 'sex', 'she', 'sho', 'sis',  
'sit', 'six', 'sky', 'son', 'soo', 'sox', 'soy', 'sry', 'ssi', 'sue', 'sum', 'sun',  
'sup', 'sux', 'tab', 'tad', 'tah', 'tdr', 'teh', 'teo', 'tha', 'the', 'tho', 'thx',  
'til', 'tks', 'toe', 'tok', 'tom', 'too', 'top', 'toy', 'try', 'tug', 'two', 'ugh',  
'umm', 'url', 'urs', 'usa', 'use', 'van', 'veg', 'was', 'wat', 'wax', 'way', 'waz',  
'wee', 'wel', 'wet', 'wha', 'who', 'why', 'wid', 'wif', 'win', 'wit', 'won', 'woo',  
'wow', 'wtf', 'wth', 'wut', 'yah', 'yak', 'yap', 'yas', 'yay', 'yea', 'yep', 'yer',  
'yes', 'yet', 'you', 'yow', 'yrs', 'yum', 'yup', 'yvw']
```

d. 在text5中，用正则表达式找出是3个字母的词，并用FreqDist()统计词频，再调用most\_common()以频率从高到低显示这些词：

```
In [15]: print(FreqDist([w for w in text5 if re.search('[A-Za-z]{3}$', w)]).most_common())
```

['lol', 704), ('the', 646), ('you', 635), ('and', 335), ('hey', 264), ('for', 18  
 8), ('are', 178), ('not', 170), ('all', 165), ('was', 142), ('how', 128), ('too', 1  
 25), ('any', 123), ('out', 107), ('can', 105), ('get', 102), ('who', 99), ('did', 9  
 2), ('but', 90), ('one', 87), ('LOL', 87), ('now', 79), ('see', 75), ('her', 71),  
 ('yes', 69), ('got', 68), ('its', 66), ('why', 59), ('had', 50), ('she', 47), ('of  
 f', 47), ('bye', 44), ('has', 42), ('You', 42), ('him', 41), ('his', 38), ('hot', 3  
 7), ('say', 34), ('man', 33), ('brb', 33), ('day', 33), ('gay', 30), ('omg', 29),  
 ('pic', 26), ('bad', 26), ('Hey', 25), ('wow', 24), ('guy', 23), ('yep', 22), ('pp  
 l', 21), ('old', 21), ('ask', 21), ('new', 20), ('way', 20), ('LoL', 20), ('huh', 1  
 9), ('hug', 19), ('put', 18), ('let', 17), ('And', 17), ('own', 17), ('yea', 17),  
 ('fun', 16), ('try', 16), ('ass', 15), ('red', 15), ('eat', 15), ('ago', 15), ('Ye  
 s', 15), ('yer', 15), ('may', 14), ('boy', 13), ('The', 13), ('yet', 12), ('bit', 1  
 2), ('cat', 12), ('big', 12), ('few', 12), ('hmm', 12), ('aww', 11), ('LoL', 10),  
 ('two', 10), ('saw', 10), ('sex', 10), ('bot', 10), ('buy', 9), ('Now', 9), ('asl',  
 9), ('sup', 9), ('nic', 9), ('ill', 9), ('yup', 9), ('bed', 8), ('bet', 8), ('use',  
 8), ('sit', 8), ('our', 8), ('How', 8), ('hit', 8), ('men', 8), ('die', 8), ('wtf',  
 7), ('lot', 7), ('far', 7), ('YOU', 7), ('kid', 7), ('lil', 7), ('cry', 7), ('car',  
 7), ('yay', 6), ('ima', 6), ('bbl', 6), ('cuz', 6), ('ice', 6), ('hun', 6), ('fly',  
 6), ('hon', 6), ('She', 6), ('NOT', 6), ('god', 6), ('cut', 6), ('dum', 6), ('ohh',  
 6), ('OMG', 6), ('bar', 6), ('sad', 6), ('nah', 6), ('Who', 6), ('cam', 6), ('pie',  
 6), ('boo', 5), ('mom', 5), ('dad', 5), ('cya', 5), ('yah', 5), ('But', 5), ('min',  
 5), ('pay', 5), ('dat', 5), ('job', 5), ('bus', 5), ('son', 5), ('run', 5), ('pop',  
 5), ('msn', 5), ('Why', 5), ('end', 5), ('afk', 5), ('top', 5), ('AND', 5), ('fir',  
 5), ('hii', 5), ('git', 5), ('Man', 5), ('lez', 5), ('ate', 5), ('fat', 5), ('Any',  
 4), ('box', 4), ('hes', 4), ('lie', 4), ('umm', 4), ('mmm', 4), ('mad', 4), ('com',  
 4), ('ONE', 4), ('cup', 4), ('Wow', 4), ('tho', 4), ('jus', 4), ('sea', 4), ('dam',  
 4), ('wut', 4), ('ugh', 4), ('pms', 4), ('dis', 4), ('ahh', 4), ('EST', 4), ('rum',  
 4), ('sho', 3), ('ALL', 3), ('ear', 3), ('yum', 3), ('eye', 3), ('psh', 3), ('Amy',  
 3), ('gal', 3), ('sis', 3), ('Ugh', 3), ('PDT', 3), ('Dew', 3), ('Set', 3), ('CDT',  
 3), ('hah', 3), ('age', 3), ('muh', 3), ('Out', 3), ('dog', 3), ('AFK', 3), ('plz',  
 3), ('win', 3), ('bbs', 3), ('fer', 3), ('bag', 3), ('air', 3), ('HEY', 3), ('pro',  
 3), ('ive', 3), ('bro', 3), ('ban', 3), ('frm', 3), ('Ooo', 3), ('wha', 3), ('set',  
 3), ('per', 3), ('Not', 3), ('YES', 3), ('lbs', 3), ('duh', 3), ('ere', 3), ('wax',  
 3), ('yow', 3), ('won', 3), ('Its', 3), ('lap', 2), ('Hmm', 2), ('bug', 2), ('lip',  
 2), ('Way', 2), ('lix', 2), ('ooo', 2), ('wat', 2), ('ads', 2), ('Ahh', 2), ('biz',  
 2), ('Too', 2), ('cal', 2), ('sun', 2), ('nut', 2), ('hiy', 2), ('til', 2), ('wth',  
 2), ('met', 2), ('WTF', 2), ('ARE', 2), ('oOo', 2), ('naw', 2), ('Oak', 2), ('hPa',  
 2), ('cus', 2), ('Ima', 2), ('wif', 2), ('yrs', 2), ('Ben', 2), ('nap', 2), ('tha',  
 2), ('toe', 2), ('nyc', 2), ('heh', 2), ('que', 2), ('HOW', 2), ('bum', 2), ('eww',  
 2), ('sue', 2), ('san', 2), ('Get', 2), ('sum', 2), ('buh', 2), ('cop', 2), ('BIG',  
 2), ('wid', 2), ('ran', 2), ('tks', 2), ('thx', 2), ('wit', 2), ('fck', 2), ('fit',  
 2), ('Wtf', 2), ('hat', 2), ('Lil', 2), ('Let', 2), ('bob', 2), ('ooh', 2), ('low',  
 2), ('UFC', 2), ('elo', 2), ('SEE', 2), ('bay', 2), ('ANY', 2), ('San', 2), ('pet',  
 2), ('usa', 2), ('tug', 2), ('gas', 2), ('les', 2), ('SAY', 2), ('haw', 1), ('kts',  
 1), ('Box', 1), ('Red', 1), ('For', 1), ('noo', 1), ('kev', 1), ('rap', 1), ('Cum',  
 1), ('nad', 1), ('erm', 1), ('Row', 1), ('Mmm', 1), ('hep', 1), ('Uhh', 1), ('Men',  
 1), ('byb', 1), ('fav', 1), ('nce', 1), ('WOW', 1), ('six', 1), ('bio', 1), ('Omg',  
 1), ('sky', 1), ('log', 1), ('joy', 1), ('con', 1), ('amy', 1), ('bra', 1), ('rid',  
 1), ('doc', 1), ('dya', 1), ('hix', 1), ('aok', 1), ('art', 1), ('odd', 1), ('ssi',  
 1), ('opr', 1), ('pos', 1), ('teh', 1), ('ops', 1), ('hio', 1), ('knw', 1), ('key',  
 1), ('beg', 1), ('Huh', 1), ('sat', 1), ('Guy', 1), ('Fix', 1), ('OOH', 1), ('LIX',  
 1), ('ole', 1), ('aka', 1), ('afe', 1), ('dry', 1), ('alo', 1), ('Pat', 1), ('sry',  
 1), ('rob', 1), ('fan', 1), ('pvt', 1), ('pit', 1), ('tom', 1), ('Ask', 1), ('BUT',  
 1), ('ans', 1), ('dun', 1), ('soo', 1), ('ben', 1), ('OUT', 1), ('arm', 1), ('gtg',  
 1), ('fwd', 1), ('dem', 1), ('yas', 1), ('pot', 1), ('Tim', 1), ('pad', 1), ('ing',  
 1), ('kep', 1), ('veg', 1), ('Lay', 1), ('ros', 1), ('doe', 1), ('DVD', 1), ('Ico',  
 1), ('End', 1), ('Our', 1), ('ehh', 1), ('hom', 1), ('God', 1), ('hoi', 1), ('JTo',  
 1), ('Own', 1), ('Cry', 1), ('bak', 1), ('wel', 1), ('Ack', 1), ('kit', 1), ('urs',  
 1), ('Can', 1), ('loc', 1), ('DON', 1), ('gee', 1), ('YaY', 1), ('oot', 1), ('btw',  
 1), ('New', 1), ('tdr', 1), ('cpr', 1), ('YoU', 1), ('PmS', 1), ('rag', 1), ('HUH',  
 1), ('van', 1), ('YOu', 1), ('lex', 1), ('Bud', 1), ('CAN', 1), ('BOY', 1), ('ifr',

```
1), ('act', 1), ('ken', 1), ('hih', 1), ('Hay', 1), ('Gay', 1), ('Ohh', 1), ('Pic',
1), ('mac', 1), ('LOl', 1), ('hoo', 1), ('ono', 1), ('egg', 1), ('wet', 1), ('mah',
1), ('HOT', 1), ('yak', 1), ('gum', 1), ('wee', 1), ('jar', 1), ('PMS', 1), ('itz',
1), ('fee', 1), ('tad', 1), ('yvw', 1), ('boi', 1), ('Are', 1), ('CST', 1), ('tok',
1), ('abs', 1), ('IRC', 1), ('hog', 1), ('mhm', 1), ('eva', 1), ('Jam', 1), ('eay',
1), ('COM', 1), ('ur1', 1), ('law', 1), ('rey', 1), ('Yay', 1), ('woo', 1), ('gun',
1), ('Max', 1), ('Aww', 1), ('Jon', 1), ('JRZ', 1), ('mum', 1), ('See', 1), ('Did',
1), ('Hug', 1), ('hel', 1), ('eng', 1), ('Ark', 1), ('Yea', 1), ('waz', 1), ('hyy',
1), ('BYE', 1), ('GNG', 1), ('cos', 1), ('CSI', 1), ('dik', 1), ('bow', 1), ('PST',
1), ('BUT', 1), ('oil', 1), ('sox', 1), ('leg', 1), ('toy', 1), ('Bye', 1), ('pal',
1), ('Ive', 1), ('hop', 1), ('ltr', 1), ('tah', 1), ('cow', 1), ('hub', 1), ('chp',
1), ('teo', 1), ('hrs', 1), ('Sat', 1), ('Sun', 1), ('sux', 1), ('LPN', 1), ('Mom',
1), ('rub', 1), ('soy', 1), ('sec', 1), ('nbc', 1), ('WHY', 1), ('ASS', 1), ('TWO',
1), ('irl', 1), ('Sup', 1), ('Joe', 1), ('Web', 1), ('Lee', 1), ('fix', 1), ('luv',
1), ('tab', 1), ('Yeh', 1), ('atl', 1), ('aim', 1), ('WHO', 1), ('RUN', 1), ('CUZ',
1), ('MAN', 1), ('bes', 1), ('hav', 1), ('yap', 1), ('lag', 1), ('Her', 1), ('THE',
1), ('msg', 1), ('MSN', 1), ('Meh', 1), ('USA', 1), ('ovr', 1), ('fot', 1)]
```

### 问题3

对以下要求，找出text2 中所有符合下列条件的词，每种要求写一个表达式。结果是词链表的形式：['word 1', 'word2', ...]。 a. 以or 结尾； b. 包含字母k； c. 包含字母序列wh； d. 除了首字母外是全部小写字母的词（即 titlecase）。

用set()对text2中的词做去重处理，并用sorted()把词按照字母顺序排序，建立text2的词库：

```
In [16]: tx2 = sorted(set(text2))
```

a. 在text2的词库中，用正则表达式找出以or结尾的词，并按字母顺序显示：

```
In [17]: print([w for w in tx2 if re.search('or$', w)])

['Doctor', 'Elinor', 'For', 'Nor', 'Poor', 'Taylor', 'abhor', 'author', 'bachelor',
'counsellor', 'demeanor', 'door', 'endeavor', 'error', 'exterior', 'floor', 'for',
'honor', 'horror', 'inferior', 'inheritor', 'labor', 'liquor', 'manor', 'metaphor',
'minor', 'narrator', 'nor', 'or', 'orator', 'poor', 'possessor', 'prior', 'propri-
or', 'stupor', 'suitor', 'superior', 'terror', 'vigor', 'visitor']
```

b. 在text2的词库中，用正则表达式找出包含字母k的词，并按字母顺序显示：



```
In [18]: print([w for w in tx2 if re.search('k', w)])
```

```
['Berkeley', 'Clarke', 'Like', 'Look', 'Luckily', 'Norfolk', 'Park', 'Sackville',  
'Shakespeare', 'Sparks', 'Take', 'Thank', 'Think', 'Walker', 'Whitakers', 'acknowledge',  
'acknowledged', 'acknowledging', 'acknowledgment', 'acknowledgments', 'alike',  
'ankle', 'ankles', 'ask', 'asked', 'asking', 'attack', 'attacked', 'attacks',  
'awake', 'awaken', 'awakened', 'awakening', 'awaking', 'awkward', 'awkwardness', 'awoke',  
'back', 'backwardness', 'backwards', 'bank', 'banker', 'basket', 'bespeak',  
'bespoke', 'black', 'blackest', 'blackguard', 'blank', 'blockhead', 'book', 'books',  
'booksellers', 'break', 'breakfast', 'breakfasting', 'breaking', 'broke', 'broken',  
'bulk', 'check', 'checked', 'checking', 'cheek', 'cheeks', 'chicken', 'choke',  
'd', 'chuckle', 'clerk', 'clock', 'coachmaker', 'colicky', 'crooked', 'dark', 'darken',  
'dislike', 'disliked', 'disliking', 'drawback', 'drink', 'drinking', 'drunk',  
'fickle', 'forsaking', 'frank', 'frankness', 'gentlemanlike', 'hackneyed', 'handkerchief',  
'honeysuckles', 'horseback', 'housekeeper', 'housekeeping', 'irksome', 'jacket',  
'joke', 'joked', 'jokes', 'joking', 'keen', 'keep', 'keeping', 'keeps', 'kept',  
'keys', 'kicked', 'kill', 'kind', 'kinder', 'kindest', 'kindly', 'kindness', 'kingdom',  
'kiss', 'kissed', 'kisses', 'kitchen', 'knack', 'knave', 'kneeling', 'knees',  
'knew', 'knives', 'knocking', 'knoll', 'know', 'knowing', 'knowledge', 'known',  
'knows', 'like', 'liked', 'likelihood', 'likely', 'likeness', 'likes', 'likewise',  
'liking', 'lock', 'look', 'looked', 'looking', 'looks', 'luck', 'luckily', 'lucky',  
'lurking', 'make', 'makes', 'making', 'mankind', 'mark', 'marked', 'marking', 'mistake',  
'mistaken', 'mistakes', 'monkey', 'neck', 'overlook', 'overlooked', 'packages',  
'packed', 'park', 'partook', 'pink', 'pocket', 'pocketbook', 'poking', 'provoke',  
'provoked', 'provoking', 'quick', 'quickened', 'quicker', 'quickest', 'quickly',  
'rank', 'ranked', 'reckoned', 'reckons', 'remark', 'remarkable', 'remarkably',  
'remarks', 'risk', 'risking', 'rocks', 'sake', 'sakes', 'seek', 'seeking', 'shake',  
'shaken', 'shock', 'shocked', 'shocking', 'shook', 'sick', 'sickly', 'sickness', 'silk',  
'silks', 'sink', 'sketch', 'skin', 'sky', 'smirked', 'smokes', 'sparkling', 'speak',  
'speaking', 'speaks', 'spoke', 'spoken', 'stake', 'stock', 'stockings', 'stocks',  
'strike', 'strikes', 'striking', 'strikingly', 'stroke', 'struck', 'sunk', 'take',  
'taken', 'takes', 'taking', 'talk', 'talked', 'talker', 'talking', 'talks', 'task',  
'thank', 'thanked', 'thankful', 'thankfully', 'thanks', 'thick', 'thickly', 'think',  
'thinking', 'thinks', 'took', 'toothpick', 'trick', 'tricked', 'tricking', 'tricks',  
'turnpike', 'unacknowledged', 'unbroken', 'undertake', 'undertaking', 'unkind',  
'unkindly', 'unkindness', 'unknowingly', 'unknown', 'unlike', 'unlikely', 'unlock',  
'unlocked', 'unluckily', 'unlucky', 'unpacked', 'unshaken', 'unspeakable', 'walk',  
'walked', 'walking', 'walks', 'weak', 'weaken', 'weakened', 'weakening', 'weakness',  
'weaknesses', 'week', 'weeks', 'wicked', 'wicket', 'winks', 'work', 'worked', 'working',  
'workmen', 'works']
```

c. 在text2的词库中，用正则表达式找出包含字母序列wh的词，并按字母顺序显示：

```
In [19]: print([w for w in tx2 if re.search('wh', w)])
```

```
['awhile', 'elsewhere', 'meanwhile', 'somewhat', 'somewhere', 'what', 'whatever',  
'when', 'whence', 'whenever', 'where', 'wherever', 'whether', 'which', 'while', 'whiled',  
'whims', 'whip', 'whisper', 'whispered', 'whispering', 'whist', 'white', 'whiter',  
'whither', 'who', 'whoever', 'whole', 'wholly', 'whom', 'whomsoever', 'whose', 'why']
```

d. 在text2的词库中，用正则表达式找出符合titlecase的词，也就是以1个大写字母开头、其他都是小写字母的词，并按字母顺序显示：

```
In [20]: print([w for w in tx2 if re.search('^[A-Z][a-z]*$', w)])
```

['A', 'Abbeyland', 'About', 'Absence', 'Abundance', 'Add', 'Affecting', 'After', 'A gain', 'Against', 'Ah', 'All', 'Allenham', 'Allow', 'Almost', 'Altogether', 'Am', 'Among', 'Amongst', 'An', 'And', 'Annamaria', 'Anne', 'Another', 'Anxiety', 'Apri l', 'Are', 'As', 'Ashamed', 'Astonished', 'Astonishment', 'At', 'Austen', 'Avigno n', 'Ay', 'Aye', 'Bad', 'Bartlett', 'Barton', 'Bath', 'Beautifully', 'Because', 'Be fore', 'Being', 'Believe', 'Benevolent', 'Berkeley', 'Besides', 'Betty', 'Between', 'Beyond', 'Biddy', 'Bishop', 'Bond', 'Bonomi', 'Born', 'Brandon', 'Bristol', 'Brow n', 'Building', 'Buildings', 'Burgess', 'Business', 'But', 'By', 'Can', 'Careless', 'Careys', 'Cartwright', 'Casino', 'Cassino', 'Certainly', 'Chagrined', 'Charlotte', 'Choice', 'Christian', 'Christmas', 'Civil', 'Clarke', 'Cleveland', 'Cold', 'Colone l', 'Columella', 'Combe', 'Come', 'Common', 'Comparisons', 'Concealing', 'Concern', 'Conduit', 'Confess', 'Consider', 'Considering', 'Constantia', 'Continual', 'Conver sation', 'Cottage', 'Could', 'Court', 'Courtland', 'Cowper', 'Cross', 'Cruel', 'D', 'Dartford', 'Dashwood', 'Dashwoods', 'Davies', 'Dawlish', 'Dear', 'Dearest', 'Delaf ord', 'Dennison', 'Depend', 'Design', 'Determined', 'Devonshire', 'Did', 'Disappoin ted', 'Disappointment', 'Do', 'Doctor', 'Does', 'Domestic', 'Don', 'Donavan', 'Dors etshire', 'Down', 'Dr', 'Drury', 'Dullness', 'During', 'Duty', 'Each', 'Eager', 'Ea rly', 'East', 'Easter', 'Edward', 'Elinor', 'Eliza', 'Elliott', 'Ellison', 'Ellison s', 'Encouraged', 'Engaged', 'Engagement', 'England', 'Epicurism', 'Esq', 'Esteem', 'Even', 'Every', 'Everybody', 'Excellent', 'Exchange', 'Excuse', 'Exert', 'Exeter', 'Extend', 'Extravagance', 'F', 'Fanny', 'Far', 'Farm', 'February', 'Ferrars', 'Fe w', 'Fifteen', 'Fifty', 'Five', 'Folly', 'For', 'Forgive', 'Fortunately', 'Four', 'Friday', 'From', 'Frosts', 'Gardens', 'Gentleman', 'Get', 'Gibson', 'Gilberts', 'G o', 'God', 'Godby', 'Going', 'Gone', 'Good', 'Gracious', 'Grandeur', 'Gray', 'Great ness', 'Grecian', 'Grey', 'Had', 'Half', 'Hamlet', 'Hanger', 'Hanover', 'Happy', 'H arley', 'Harris', 'Harry', 'Has', 'Have', 'Having', 'He', 'Heaven', 'Henry', 'Hensh awe', 'Her', 'Here', 'High', 'His', 'Hitherto', 'Holborn', 'Holburn', 'Hon', 'Honit on', 'Hope', 'Hour', 'House', 'How', 'However', 'Hum', 'Hunters', 'Hush', 'I', 'I f', 'Imagine', 'Impatient', 'Impossible', 'Improve', 'Impudence', 'In', 'Indeed', 'Indies', 'Infirmary', 'Inn', 'Instead', 'Invited', 'Is', 'It', 'James', 'Jane', 'J anuary', 'Jenning', 'Jennings', 'John', 'Just', 'Kensington', 'Kingham', 'Know', 'K nowing', 'L', 'La', 'Lady', 'Ladyship', 'Lane', 'Last', 'Laughing', 'Law', 'Let', 'Life', 'Like', 'Little', 'Lodging', 'Lombardy', 'London', 'Long', 'Longstaple', 'L ook', 'Lord', 'Luckily', 'Lucy', 'M', 'Ma', 'Mab', 'Madam', 'Magna', 'Mall', 'Mam a', 'Mamma', 'Mansion', 'Many', 'March', 'Margaret', 'Marianne', 'Marlborough', 'Ma rtha', 'Mary', 'Master', 'May', 'Me', 'Men', 'Michaelmas', 'Mid', 'Middleton', 'Mid dletons', 'Midsummer', 'Mind', 'Mine', 'Misery', 'Miss', 'Misses', 'Mistress', 'Mon day', 'Months', 'More', 'Morton', 'Most', 'Mr', 'Mrs', 'Much', 'Music', 'Must', 'M y', 'Nancy', 'Nay', 'Neither', 'Never', 'New', 'Newton', 'No', 'Nobody', 'None', 'N or', 'Norfolk', 'Norland', 'Not', 'Nothing', 'November', 'Now', 'October', 'Of', 'O ffended', 'Oh', 'On', 'Once', 'One', 'Only', 'Opportunity', 'Opposition', 'Or', 'Ot her', 'Others', 'Our', 'Oxford', 'P', 'Pall', 'Palmer', 'Palmers', 'Pardon', 'Par k', 'Parliament', 'Parrys', 'Parsonage', 'Perhaps', 'Pity', 'Please', 'Pleased', 'P lymouth', 'Poor', 'Pope', 'Portman', 'Pratt', 'Pray', 'Precious', 'Preparation', 'P rescriptions', 'Priory', 'Queen', 'Quite', 'Rather', 'Reading', 'Really', 'Recollec ting', 'Reflection', 'Regard', 'Relate', 'Remember', 'Reserved', 'Restless', 'Richa rd', 'Richardson', 'Richardsons', 'Robert', 'Rose', 'S', 'Sackville', 'Sally', 'San dersons', 'Saturday', 'Scarcely', 'Scotland', 'Scott', 'Secrecy', 'Selfish', 'Sens e', 'Sensibility', 'September', 'Seven', 'Shakespeare', 'Shall', 'Sharpe', 'She', 'Short', 'Should', 'Shyness', 'Simpson', 'Since', 'Sincerely', 'Sir', 'Sit', 'Smit h', 'So', 'Some', 'Somehow', 'Somerset', 'Somersetshire', 'Something', 'Sometimes', 'Soon', 'Sophia', 'Sparks', 'Square', 'St', 'Stanhill', 'Steele', 'Steeles', 'Stil l', 'Strange', 'Street', 'Streets', 'Such', 'Sunday', 'Supported', 'Supposing', 'Su re', 'Surely', 'Surprised', 'Sussex', 'Take', 'Taylor', 'Tell', 'Temple', 'Thank', 'That', 'The', 'Their', 'Then', 'There', 'These', 'They', 'Think', 'This', 'Thoma s', 'Thomson', 'Those', 'Though', 'Three', 'Thunderbolts', 'Thursday', 'Thus', 'Til l', 'Time', 'Tis', 'To', 'Towards', 'Truth', 'Tuesday', 'Twice', 'Twill', 'Two', 'U naccountable', 'Undoubtedly', 'Ungracious', 'Upon', 'Use', 'Valley', 'Vanity', 'Ver y', 'Volume', 'W', 'Wait', 'Walker', 'Want', 'Was', 'Watched', 'We', 'Wednesday', 'Well', 'Were', 'Westminster', 'Westons', 'Weymouth', 'What', 'Whatever', 'When', 'Whenever', 'Where', 'Whereas', 'Wherever', 'Whether', 'Which', 'While', 'Whitaker

s', 'Whitwell', 'Who', 'Whoever', 'Whom', 'Why', 'Will', 'William', 'Williams', 'Willing', 'Willoughby', 'Willoughbys', 'With', 'Within', 'Without', 'Would', 'Writing', 'Yes', 'Yet', 'You', 'Your']

In [ ]: