NLTK 尝试 实验报告

宁晨然 17307130178

一、问题一

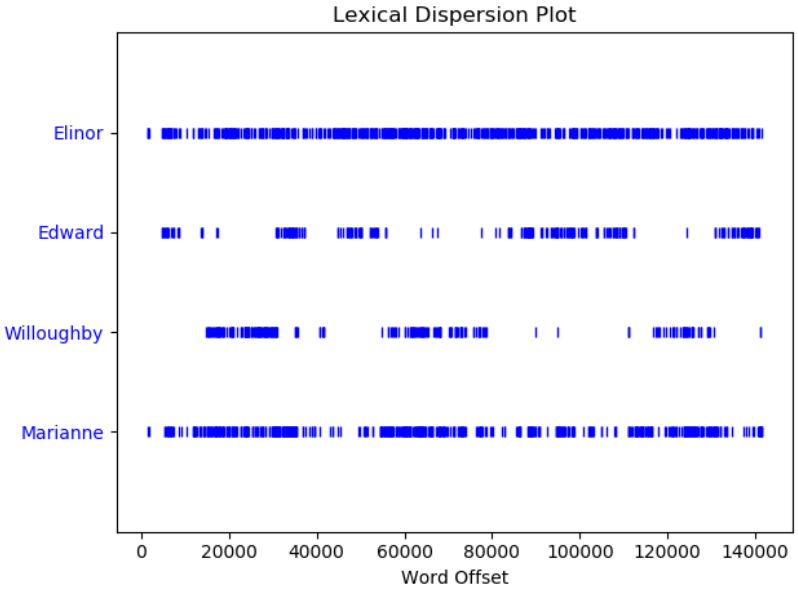
（1）问题

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？能找出一堆夫妻吗？

（2）解答

首先 pip install nltk 并根据 nltk.download 中内容加载 book 所需模块。根据以下代码画出分布图。

```
from nltk.book import text2
# main characters in <Sense and Sensibility>
text2.dispersion_plot(['Elinor', 'Edward', 'Willoughby', 'Marianne'])
```



（3）分析

根据名字的分布图可以看出 Elinor 是女主角，Marianne 是女配角，与原文符合。而 Edward/Willoughby 是男性角色，是女性角色的对象，但描写明显少很多，说明是一本以女性为主要线索的小说。根据 Willoughby 和 Marianne 的分布可以判断是一对夫妻。

二、问题二

（1）问题

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。

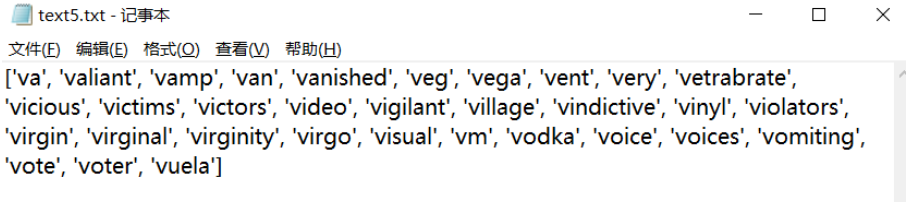
（2）解答

2.1 查找字母 v 开头的词

使用正则匹配即可 `'^v[A-Za-z]*$'`，排序去重后输出到 txt。

```
from nltk.book import text5
import re, pprint,nltk
# '^v'
vword = [v for v in text5 if re.search('^v[a-zA-Z]*$',v)]
vword = sorted(list(set(vword)))
data = open("text5.txt","w")
print(vword,file=data)
```

结果为：

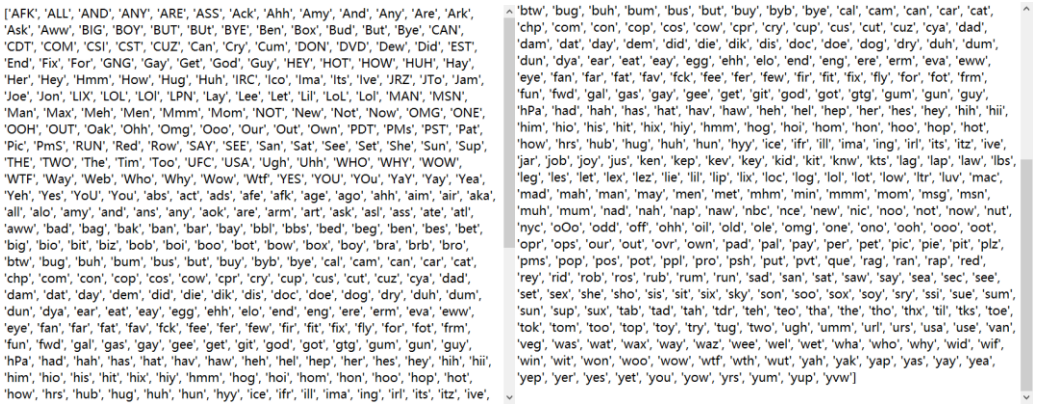


2.2 查找三个字母的词

同理使用正则匹配，'^[A-Za-z]{3}$'，输出到文件。改变部分代码。

```
vword = [v for v in text5 if re.search('^[a-zA-Z]{3}$',v)]
```

结果为：

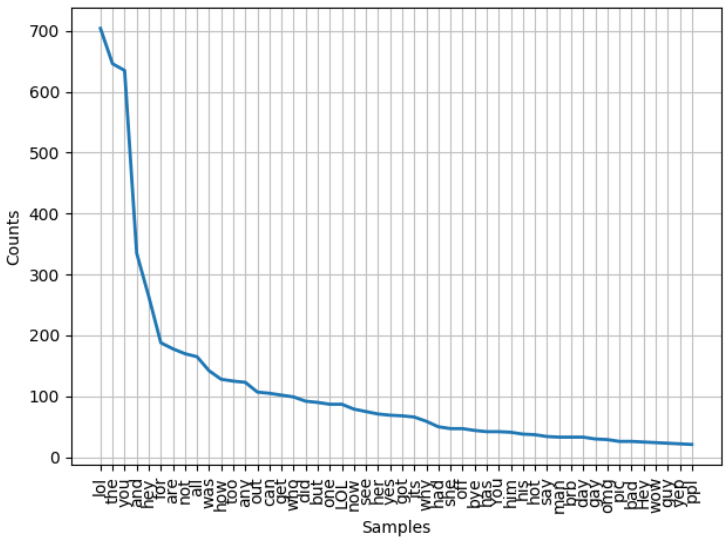


2.3 画出频率分布图

根据上面得到的 vword，利用 FreDist 函数可以画出频率分布图（由于横坐标单词过多，我简化画了频率前 50 的单词频率分布图），代码如下：

```
from nltk.book import text5
import re, pprint,nltk

vword = [v for v in text5 if
re.search('^[a-zA-
Z]{3}$',v)]
fdist = nltk.FreqDist(vword)
fdist.plot(50)
```



三、问题三

（1）问题

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',...]。 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）

（2）解答

根据以上要求改写不同的正则表达式，此处只写出正则表达式的代码。

```
word1 = [v for v in text2 if re.search('^[a-zA-Z]*or$',v)]
word2 = [v for v in text2 if re.search('^[a-zA-Z]*k[a-zA-Z]*$',v)]
word3 = [v for v in text2 if re.search('^[a-zA-Z]*wh[a-zA-Z]*$',v)]
word4 = [v for v in text2 if re.search('^[A-Z]{1}[a-z]*',v)]
```

结果为：

test1:

['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', 'proprietor', 'stupor', 'suitor', 'superior', 'terror', 'vigor', 'visitor']

test2:

['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', 'choked', 'chuckle', 'clerk', 'clock', 'coachmaker', 'colicky', 'crooked', 'dark', 'darker', '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', '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']

test3:

['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']

test4:

['A', 'ALL', 'ALMOST', 'ALWAYS', 'AM', 'ANY', 'ARE', 'Abbeyland', 'About', 'Absence', 'Abundance', 'Add', 'Affecting', 'After', 'Again', 'Against', 'Ah', 'All', 'Allenham', 'Allow', 'Almost', 'Altogether', 'Am', 'Among', 'Amongst', 'An', 'And', 'Annamaria', 'Anne', 'Another', 'Anxiety', 'April', 'Are', 'As', 'Ashamed', 'Astonished', 'Astonishment', 'At', 'Austen', 'Avignon', 'Ay', 'Aye', 'BEEN', 'BOTH', 'Bad', 'Bartlett', 'Barton', 'Bath', 'Beautifully', 'Because', 'Before', 'Being', 'Believe', 'Benevolent', 'Berkeley', 'Besides', 'Betty', 'Between', 'Beyond', 'Biddy', 'Bishop', 'Bond', 'Bonomi', 'Born', 'Brandon', 'Bristol', 'Brown', 'Building', 'Buildings', 'Burgess', 'Business', 'But', 'By', 'CAN', 'CATCHING', 'CHAPTER', 'COULD', 'Can', 'Careless', 'Careys', 'Cartwright', 'Casino', 'Cassino', 'Certainly', 'Chagrined', 'Charlotte', 'Choice', 'Christian', 'Christmas', 'Civil', 'Clarke', 'Cleveland', 'Cold', 'Colonel', 'Columella', 'Combe', 'Come', 'Common', 'Comparisons', 'Concealing', 'Concern', 'Conduit', 'Confess', 'Consider', 'Considering', 'Constantia', 'Continual', 'Conversation', 'Cottage', 'Could', 'Court', 'Courtland', 'Cowper', 'Cross', 'Cruel', 'D', 'DEAR', 'DID', 'DO', 'DOES', 'DRAW', 'Dartford', 'Dashwood', 'Dashwoods', 'Davies', 'Dawlish', 'Dear', 'Dearest', 'Delaford', 'Dennison', 'Depend', 'Design', 'Determined', 'Devonshire', 'Did', 'Disappointed', 'Disappointment', 'Do', 'Doctor', 'Does', 'Domestic', 'Don', 'Donavan', 'Dorsetshire', 'Down', 'Dr', 'Drury', 'Dullness', 'During', 'Duty', 'EDWARD', 'ELINOR', 'END', 'ESTEEM', 'Each', 'Eager', 'Early', 'East', 'Easter', 'Edward', 'Elinor', 'Eliza', 'Elliott', 'Ellison', 'Ellisons', 'Encouraged', 'Engaged', 'Engagement', 'England', 'Epicurism', 'Esq', 'Esteem', 'Even', 'Every', 'Everybody', 'Excellent', 'Exchange', 'Excuse', 'Exert', 'Exeter', 'Extend', 'Extravagance', 'F', 'FAITH', 'FERRARS', 'Fanny', 'Far', 'Farm', 'February', 'Ferrars', 'Few', 'Fifteen', 'Fifty', 'Five', 'Folly', 'For', 'Forgive', 'Fortunately', 'Four', 'Friday', 'From', 'Frosts', 'GAUCHERIE', 'Gardens', 'Gentleman', 'Get', 'Gibson', 'Gilberts', 'Go', 'God', 'Godby', 'Going', 'Gone', 'Good', 'Gracious', 'Grandeur', 'Gray', 'Greatness', 'Grecian', 'Grey', 'HAD', 'HAS', 'HE', 'HER', 'HERS', 'HIM', 'HIS', 'Had', 'Half', 'Hamlet', 'Hanger', 'Hanover', 'Happy', 'Harley', 'Harris', 'Harry', 'Has', 'Have', 'Having', 'He', 'Heaven', 'Henry', 'Henshawe', 'Her', 'Here', 'High', 'His', 'Hitherto', 'Holborn', 'Holburn', 'Hon', 'Honiton', 'Hope', 'Hour', 'House', 'How', 'However', 'Hum', 'Hunters', 'Hush', 'I', 'II', 'IN', 'INconvenience', 'IS', 'If', 'Imagine', 'Impatient', 'Impossible', 'Improve', 'Impudence', 'In', 'Indeed', 'Indies', 'Infirmity', 'Inn', 'Instead', 'Invited', 'Is', 'It', 'JOHN', 'James', 'Jane', 'January', 'Jenning', 'Jennings', 'John', 'Just', 'KNEW', 'Kensington', 'Kingham', 'Know', 'Knowing', 'L', 'LESS', 'LET', 'LONG', 'LOOK', 'LOOKED', 'LUCY', 'La', 'Lady', 'Ladyship', 'Lane', 'Last', 'Laughing', 'Law', 'Let', 'Life', 'Like', 'Little', 'Lodging', 'Lombardy', 'London', 'Long', 'Longstaple', 'Look', 'Lord', 'Luckily', 'Lucy', 'M', 'MADAM', 'MAY', 'ME', 'MIND', 'MONTH', 'MUST', 'MY', 'Ma', 'Mab', 'Madam', 'Magna', 'Mall', 'Mama', 'Mamma', 'Mansion', 'Many', 'March', 'Margaret', 'Marianne', 'Marlborough', 'Martha', 'Mary', 'Master', 'May', 'Me', 'Men', 'Michaelmas', 'Mid', 'Middleton', 'Middletons', 'Midsummer', 'Mind', 'Mine', 'Misery', 'Miss', 'Misses', 'Mistress', 'Monday', 'Months', 'More', 'Morton', 'Most',

'Mr', 'Mrs', 'Much', 'Music', 'Must', 'My', 'NOT', 'NOW', 'Nancy', 'Nay', 'Neither', 'Never', 'New', 'Newton', 'No', 'Nobody', 'None', 'Nor', 'Norfolk', 'Norland', 'Not', 'Nothing', 'November', 'Now', 'OCCASION', 'ONCE', 'ONE', 'OUGHT', 'OWN', 'October', 'Of', 'Offended', 'Oh', 'On', 'Once', 'One', 'Only', 'Opportunity', 'Opposition', 'Or', 'Other', 'Others', 'Our', 'Oxford', 'P', 'PARTIES', 'Pall', 'Palmer', 'Palmers', 'Pardon', 'Park', 'Parliament', 'Parrys', 'Parsonage', 'Perhaps', 'Pity', 'Please', 'Pleased', 'Plymouth', 'Poor', 'Pope', 'Portman', 'Pratt', 'Pray', 'Precious', 'Preparation', 'Prescriptions', 'Priory', 'Queen', 'Quite', 'REALLY', 'ROBERT', 'Rather', 'Reading', 'Really', 'Recollecting', 'Reflection', 'Regard', 'Relate', 'Remember', 'Reserved', 'Restless', 'Richard', 'Richardson', 'Richardsons', 'Robert', 'Rose', 'S', 'SHALL', 'SHE', 'SHOULD', 'SIR', 'SOMETIMES', 'STILL', 'Sackville', 'Sally', 'Sandersons', 'Saturday', 'Scarcely', 'Scotland', 'Scott', 'Secrecy', 'Selfish', 'Sense', 'Sensibility', 'September', 'Seven', 'Shakespeare', 'Shall', 'Sharpe', 'She', 'Short', 'Should', 'Shyness', 'Simpson', 'Since', 'Sincerely', 'Sir', 'Sit', 'Smith', 'So', 'Some', 'Somehow', 'Somerset', 'Somersetshire', 'Something', 'Sometimes', 'Soon', 'Sophia', 'Sparks', 'Square', 'St', 'Stanhill', 'Steele', 'Steeles', 'Still', 'Strange', 'Street', 'Streets', 'Such', 'Sunday', 'Supported', 'Supposing', 'Sure', 'Surely', 'Surprised', 'Sussex', 'THAT', 'THE', 'THEIR', 'THEM', 'THEN', 'THERE', 'THESE', 'THEY', 'THIS', 'THREE', 'TIME', 'TOLD', 'TRIED', 'TWICE', 'TWO', 'Take', 'Taylor', 'Tell', 'Temple', 'Thank', 'That', 'The', 'Their', 'Then', 'There', 'These', 'They', 'Think', 'This', 'Thomas', 'Thomson', 'Those', 'Though', 'Three', 'Thunderbolts', 'Thursday', 'Thus', 'Till', 'Time', 'Tis', 'To', 'Towards', 'Truth', 'Tuesday', 'Twice', 'Twill', 'Two', 'US', 'Unaccountable', 'Undoubtedly', 'Ungracious', 'Upon', 'Use', 'VERY', 'Valley', 'Vanity', 'Very', 'Volume', 'W', 'WAS', 'WE', 'WERE', 'WHAT', 'WHERE', 'WILL', 'WILLOUGHBY', 'WITHOUT', 'WORD', 'WOULD', 'Wait', 'Walker', 'Want', 'Was', 'Watched', 'We', 'Wednesday', 'Well', 'Were', 'Westminster', 'Westons', 'Weymouth', 'What', 'Whatever', 'When', 'Whenever', 'Where', 'Whereas', 'Wherever', 'Whether', 'Which', 'While', 'Whitakers', 'Whitwell', 'Who', 'Whoever', 'Whom', 'Why', 'Will', 'William', 'Williams', 'Willing', 'Willoughby', 'Willoughbys', 'With', 'Within', 'Without', 'Would', 'Writing', 'YOU', 'YOUR', 'Yes', 'Yet', 'You', 'Your']