DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 2: Basic text processing

Answered to problems 2.1, 2.2, 2.3

2.1

Code in *E2_1.py.*

In part 1 I done the following:

Original version:

```
pagestocrawl=pagestocrawl[1:len(pagestocrawl)]
```

My version:

```
pagestocrawl.remove(pagetocrawl_url)
```

```
pagetocrawl_urls = list(set(pagetocrawl_urls) - set(crawled_urls))
```

In my version I used set operations to verify that already processed urls don't get crawled again simply by removing already crawled ones.

In part 2 I done the following:

Original version:

```
pagetocrawl_url = pagestocrawl[0]
```

My version:

```
pagetocrawl_url = pagestocrawl[0]
        if random.random() < 0.7:
            pagetocrawl_url = random.choice(pagestocrawl)
```

In my version I select the topmost remaining page by probability of 0.7, and random page otherwise.

2.2

Code in *E2_2.py*.

Output from b)

```
Downloading book: Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft
Shelley
From source: https://www.gutenberg.org/84/84-0.txt
Downloading book: Pride and Prejudice by Jane Austen
From source: https://www.gutenberg.org/1342/1342-0.txt
Downloading book: A Christmas Carol in Prose; Being a Ghost Story of Christmas by
Charles Dickens
From source: https://www.gutenberg.org/46/46-0.txt
Downloading book: The Scarlet Letter by Nathaniel Hawthorne
From source: https://www.gutenberg.org/25344/25344-0.txt
Downloading book: The Yellow Wallpaper by Charlotte Perkins Gilman
From source: https://www.gutenberg.org/1952/1952-0.txt
Downloading book: Et dukkehjem. English by Henrik Ibsen
From source: https://www.gutenberg.org/2542/2542-0.txt
Downloading book: Alice's Adventures in Wonderland by Lewis Carroll
From source: https://www.gutenberg.org/11/11-0.txt
Downloading book: A Modest Proposal by Jonathan Swift
From source: https://www.gutenberg.org/1080/1080-0.txt
Downloading book: Metamorphosis by Franz Kafka
From source: https://www.gutenberg.org/5200/5200.txt
Downloading book: Beowulf: An Anglo-Saxon Epic Poem
From source: https://www.gutenberg.org/16328/16328-8.txt
Downloading book: Anthem by Ayn Rand
From source: https://www.gutenberg.org/1250/1250-0.txt
Downloading book: A Tale of Two Cities by Charles Dickens
From source: https://www.gutenberg.org/98/98-0.txt
Downloading book: Il Principe. English by Niccolò Machiavelli
From source: https://www.gutenberg.org/1232/1232.txt
Downloading book: The Strange Case of Dr. Jekyll and Mr. Hyde by Robert Louis
Stevenson
From source: https://www.gutenberg.org/43/43-0.txt
Downloading book: Moby Dick; Or, The Whale by Herman Melville
From source: https://www.gutenberg.org/2701/2701-0.txt
Downloading book: The Adventures of Sherlock Holmes by Arthur Conan Doyle
From source: https://www.gutenberg.org/1661/1661-0.txt
Downloading book: The Legend of Sleepy Hollow by Washington Irving
From source: https://www.gutenberg.org/41/41-0.txt
Downloading book: The Importance of Being Earnest: A Trivial Comedy for Serious People
by Oscar Wilde
From source: https://www.gutenberg.org/844/844.txt
Downloading book: Dracula by Bram Stoker
From source: https://www.gutenberg.org/345/345-8.txt
Downloading book: Adventures of Huckleberry Finn by Mark Twain
From source: https://www.gutenberg.org/76/76-0.txt
```

Output from d)

```
top-100 words
['the' '.' 'be' 'and' 'of' 'to' 'a' 'I' 'in' 'have' 'that' 'it' ';' '“'
 'his' 'he' '”' 'with' 'as' 'you' 'for' 'not' 'her' '!' 'at' 'on' 'my'
 'him' '’' 'do' 'say' 'all' 'me' 'but' '?' 'so' 'by' 'which' 'this' 'from'
 'The' 'she' 'they' 'we' 'one' 's' '--' 'go' 'would' 'there' 'no' 'or'
 'them' 'out' 'when' 'come' 'up' 'could' 'know' 'see' 'an' 'if' 'their'
 "'" "'s" 'He' 'It' 'make' 'what' 'will' ':' 'then' 'But' 'more' 'time'
 'who' 'into' "n't" 'your' 'look' 'take' 'some' 'now' 'very' 'man' 'can'
 'get' 'think' 'upon' 'like' 'And' 'Mr.' 'little' 'down' 'than' 'hand'
 'any' '``' 'good']
```

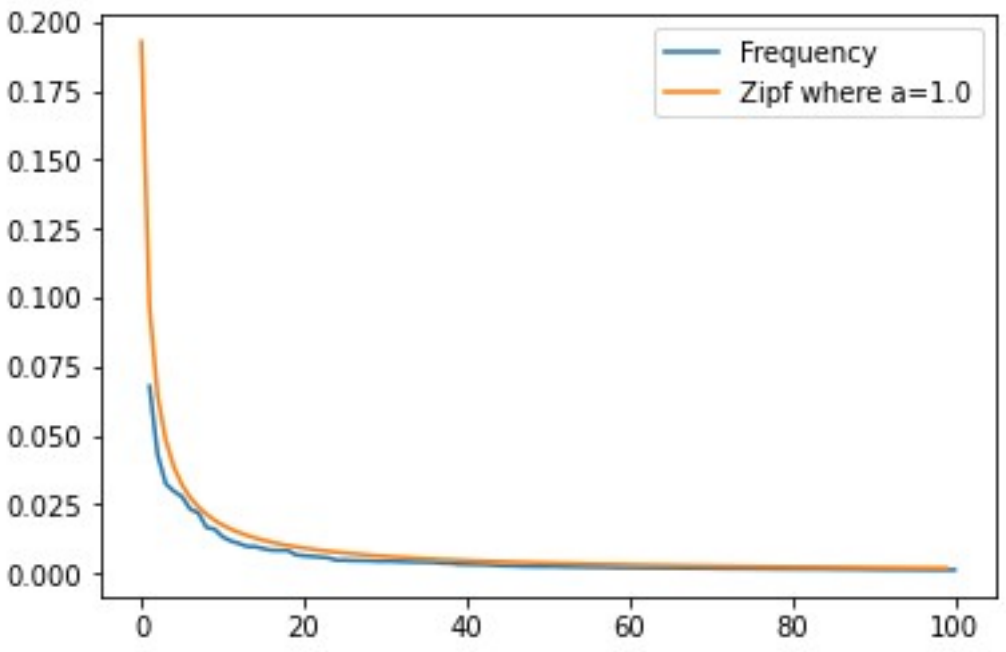
Output from e)

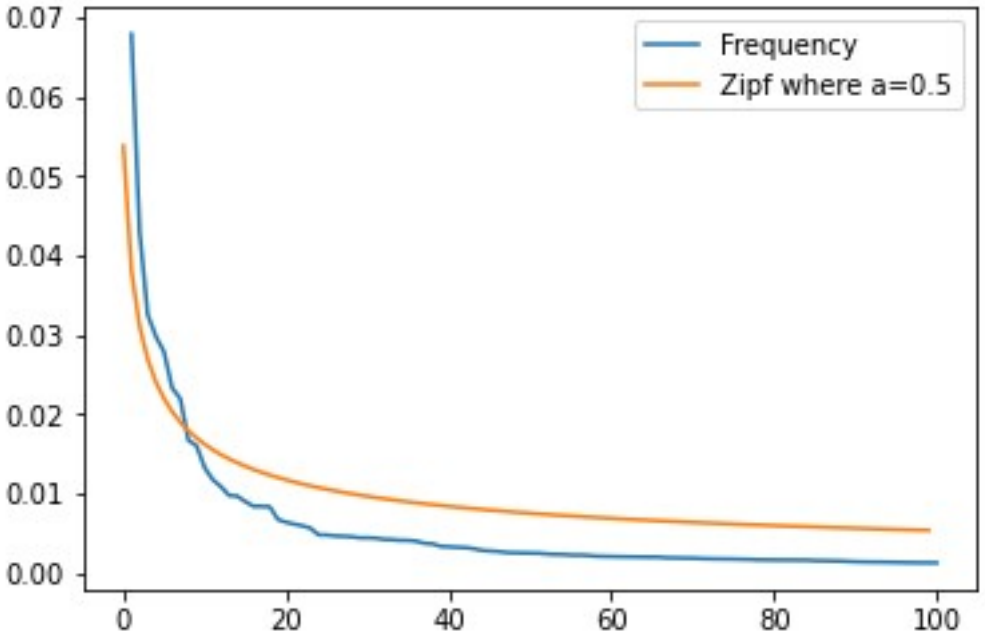
```
top-100 words
['wife' 'really' 'suppose' 'lie' 'already' 'Van' 'hard' 'With' 'either'
 'doubt' 'least' 'care' 'strong' 'All' 'true' 'Jane' 'thus' 'husband'
 'line' 'receive' 'touch' 'de' 'feeling' 'else' 'black' 'One' 'Bennet'
 'person' 'therefore' 'thee' 'happy' 'king' 'street' 'wonder' 'deep'
 'able' 'trouble' 'remember' 'ready' 'Helsing' 'everything' 'raise' 'save'
 'sometimes' 'certain' 'daughter' 'On' 'nature' 'Bingley' 'laugh'
 'consider' 'thousand' 'Defarge' 'pretty' 'short' 'table' 'fell' 'HELMER'
 'tear' 'ai' 'seek' 'girl' 'hardly' 'Whale' 'lead' 'learn' 'drive'
 'Doctor' 'Gregor' 'earth' 'wind' 'Not' 'sun' 'Lucy' 'four' 'wall'
 'observe' 'money' 'subject' 'wild' 'chance' 'Here' 'ground' 'blood'
 'second' 'shake' 'force' 'clear' 'add' 'land' 'expect' 'book' 'prince'
 'boy' 'play' 'drop' 'fellow' 'creature' 'forward']
```

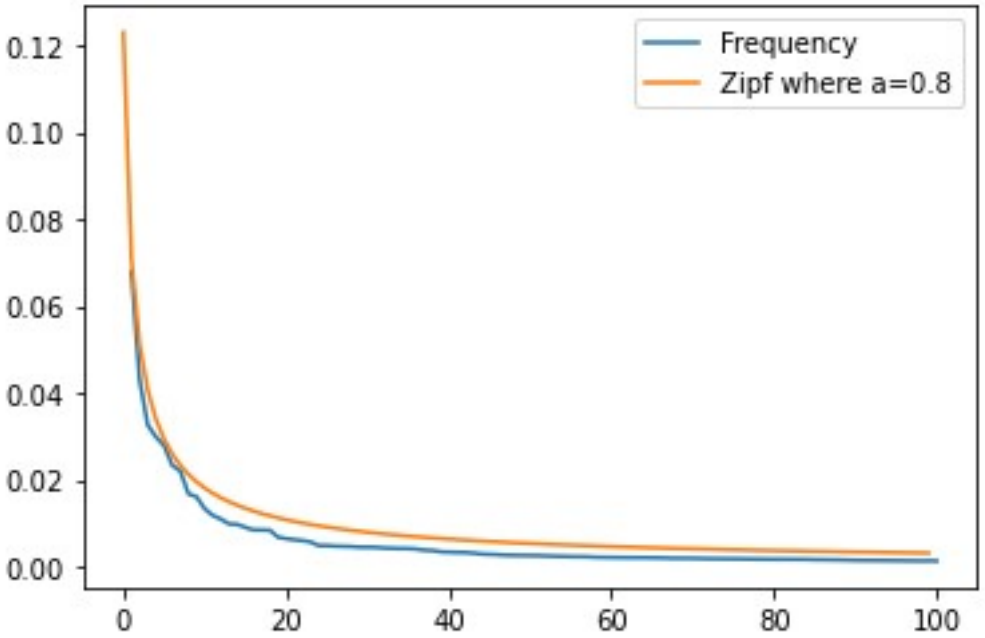
2.3

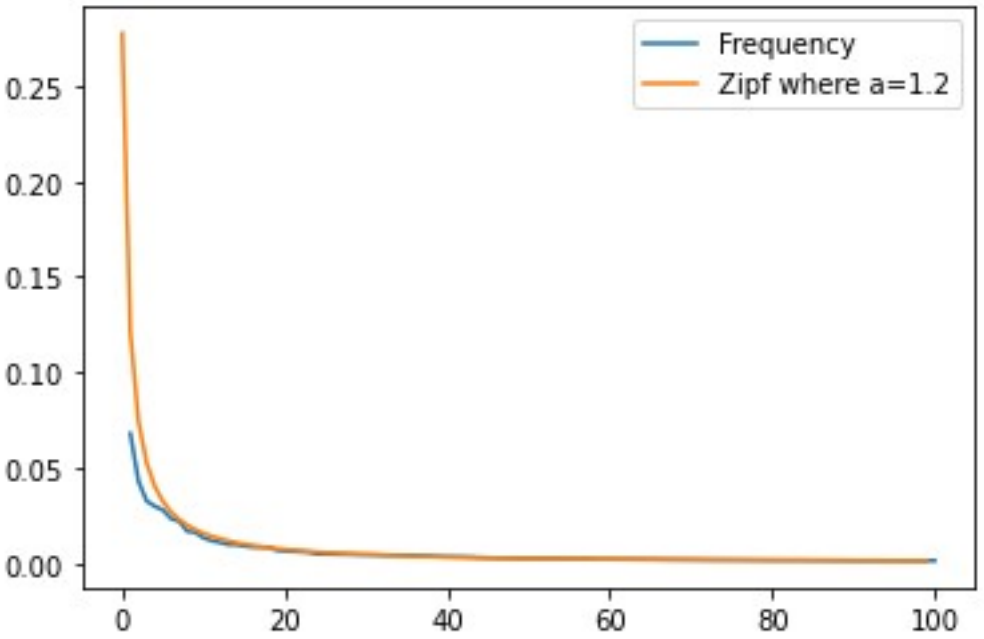
Code in *E2_3.py*.

Output from a) & b)




Frequency
Zipf where a=0.5
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100


Frequency
Zipf where a=0.8
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100


Frequency
Zipf where a=1.2
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100