Taller 9

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 28-oct-2016 11:59 PM

[David Valles]

[david.valles@urosario.edu.co]

Instrucciones:

- Guarde una copia de este *Jupyter Notebook* en su computador, idealmente en una carpeta destinada al material del curso.
- Modifique el nombre del archivo del *notebook*, agregando al final un guión inferior y su nombre y apellido, separados estos últimos por otro guión inferior. Por ejemplo, mi *notebook* se llamaría: mcpp_taller9_santiago_matallana
- Marque el *notebook* con su nombre y e-mail en el bloque verde arriba. Reemplace el texto "[Su nombre acá]" con su nombre y apellido. Similar para su e-mail.
- Desarrolle la totalidad del taller sobre este *notebook*, insertando las celdas que sea necesario debajo de cada pregunta. Haga buen uso de las celdas para código y de las celdas tipo *markdown* según el caso.
- Recuerde salvar periódicamente sus avances.
- Cuando termine el taller:
    1. Descárguelo en PDF. Si tiene algún problema con la conversión, descárguelo en HTML.
    2. Suba todos los archivos a su repositorio en GitHub, en una carpeta destinada exclusivamente para este taller, antes de la fecha y hora límites.

NLTK Book (http://www.nltk.org/book/ (http://www.nltk.org/book/)), Exercises:

- Chapter 1: 22, 26, 28
- Chapter 2: 2, 4, 11

In [1]:

```
import nltk
```

In [2]:

```
dler = nltk.downloader.Downloader()
dler._update_index()
dler._status_cache['panlex_lite'] = 'installed' # Trick the index to treat panlex_lit
e as it's already installed.
dler.download('all')
```

```
[nltk_data] Downloading collection 'all'
[nltk_data]    |
[nltk_data]    | Downloading package abc to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package alpino to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package alpino is already up-to-date!
[nltk_data]    | Downloading package biocreative_ppi to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package biocreative_ppi is already up-to-date!
[nltk_data]    | Downloading package brown to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package brown is already up-to-date!
[nltk_data]    | Downloading package brown_tei to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package brown_tei is already up-to-date!
[nltk_data]    | Downloading package cess_cat to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cess_cat is already up-to-date!
[nltk_data]    | Downloading package cess_esp to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cess_esp is already up-to-date!
[nltk_data]    | Downloading package chat80 to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package chat80 is already up-to-date!
[nltk_data]    | Downloading package city_database to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package city_database is already up-to-date!
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package comparative_sentences to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package comparative_sentences is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package comtrans to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package comtrans is already up-to-date!
[nltk_data]    | Downloading package conll2000 to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2000 is already up-to-date!
[nltk_data]    | Downloading package conll2002 to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2002 is already up-to-date!
[nltk_data]    | Downloading package conll2007 to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2007 is already up-to-date!
[nltk_data]    | Downloading package crubadan to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package crubadan is already up-to-date!
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package dependency_treebank is already up-to-date!
[nltk_data]    | Downloading package dolch to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    | Downloading package europarl_raw to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
```

```
[nltk_data]     |   Package europarl_raw is already up-to-date!
[nltk_data]     | Downloading package floresta to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package floresta is already up-to-date!
[nltk_data]     | Downloading package framenet_v15 to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package framenet_v15 is already up-to-date!
[nltk_data]     | Downloading package framenet_v17 to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package framenet_v17 is already up-to-date!
[nltk_data]     | Downloading package gazetteers to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package gazetteers is already up-to-date!
[nltk_data]     | Downloading package genesis to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package genesis is already up-to-date!
[nltk_data]     | Downloading package gutenberg to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package gutenberg is already up-to-date!
[nltk_data]     | Downloading package ieer to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package ieer is already up-to-date!
[nltk_data]     | Downloading package inaugural to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package inaugural is already up-to-date!
[nltk_data]     | Downloading package indian to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package indian is already up-to-date!
[nltk_data]     | Downloading package jeita to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package jeita is already up-to-date!
[nltk_data]     | Downloading package kimmo to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package kimmo is already up-to-date!
[nltk_data]     | Downloading package knbc to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package knbc is already up-to-date!
[nltk_data]     | Downloading package lin_thesaurus to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package lin_thesaurus is already up-to-date!
[nltk_data]     | Downloading package mac_morpho to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package mac_morpho is already up-to-date!
[nltk_data]     | Downloading package machado to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package machado is already up-to-date!
[nltk_data]     | Downloading package masc_tagged to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package masc_tagged is already up-to-date!
[nltk_data]     | Downloading package moses_sample to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package moses_sample is already up-to-date!
[nltk_data]     | Downloading package movie_reviews to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package movie_reviews is already up-to-date!
[nltk_data]     | Downloading package names to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package names is already up-to-date!
[nltk_data]     | Downloading package nombank.1.0 to
```

```
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package nombank.1.0 is already up-to-date!
[nltk_data]     | Downloading package nps_chat to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package nps_chat is already up-to-date!
[nltk_data]     | Downloading package omw to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package omw is already up-to-date!
[nltk_data]     | Downloading package opinion_lexicon to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package opinion_lexicon is already up-to-date!
[nltk_data]     | Downloading package paradigms to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package paradigms is already up-to-date!
[nltk_data]     | Downloading package pil to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package pil is already up-to-date!
[nltk_data]     | Downloading package pl196x to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package pl196x is already up-to-date!
[nltk_data]     | Downloading package ppattach to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package ppattach is already up-to-date!
[nltk_data]     | Downloading package problem_reports to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package problem_reports is already up-to-date!
[nltk_data]     | Downloading package propbank to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package propbank is already up-to-date!
[nltk_data]     | Downloading package ptb to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package ptb is already up-to-date!
[nltk_data]     | Downloading package product_reviews_1 to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package product_reviews_1 is already up-to-date!
[nltk_data]     | Downloading package product_reviews_2 to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package product_reviews_2 is already up-to-date!
[nltk_data]     | Downloading package pros_cons to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package pros_cons is already up-to-date!
[nltk_data]     | Downloading package qc to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package qc is already up-to-date!
[nltk_data]     | Downloading package reuters to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package reuters is already up-to-date!
[nltk_data]     | Downloading package rte to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package rte is already up-to-date!
[nltk_data]     | Downloading package semcor to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package semcor is already up-to-date!
[nltk_data]     | Downloading package senseval to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package senseval is already up-to-date!
[nltk_data]     | Downloading package sentiwordnet to
[nltk_data]     |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]     |   Package sentiwordnet is already up-to-date!
```

```
[nltk_data]    | Downloading package sentence_polarity to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sentence_polarity is already up-to-date!
[nltk_data]    | Downloading package shakespeare to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package shakespeare is already up-to-date!
[nltk_data]    | Downloading package sinica_treebank to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sinica_treebank is already up-to-date!
[nltk_data]    | Downloading package smultron to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package smultron is already up-to-date!
[nltk_data]    | Downloading package state_union to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package state_union is already up-to-date!
[nltk_data]    | Downloading package stopwords to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package stopwords is already up-to-date!
[nltk_data]    | Downloading package subjectivity to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package subjectivity is already up-to-date!
[nltk_data]    | Downloading package swadesh to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package swadesh is already up-to-date!
[nltk_data]    | Downloading package switchboard to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package switchboard is already up-to-date!
[nltk_data]    | Downloading package timit to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package timit is already up-to-date!
[nltk_data]    | Downloading package toolbox to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package toolbox is already up-to-date!
[nltk_data]    | Downloading package treebank to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package treebank is already up-to-date!
[nltk_data]    | Downloading package twitter_samples to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package twitter_samples is already up-to-date!
[nltk_data]    | Downloading package udhr to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package udhr is already up-to-date!
[nltk_data]    | Downloading package udhr2 to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package udhr2 is already up-to-date!
[nltk_data]    | Downloading package unicode_samples to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package unicode_samples is already up-to-date!
[nltk_data]    | Downloading package universal_treebanks_v20 to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package universal_treebanks_v20 is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package verbnet to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package verbnet is already up-to-date!
[nltk_data]    | Downloading package webtext to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package webtext is already up-to-date!
[nltk_data]    | Downloading package wordnet to
```

```
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet is already up-to-date!
[nltk_data]    | Downloading package wordnet_ic to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet_ic is already up-to-date!
[nltk_data]    | Downloading package words to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package words is already up-to-date!
[nltk_data]    | Downloading package ycoe to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ycoe is already up-to-date!
[nltk_data]    | Downloading package rslp to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package rslp is already up-to-date!
[nltk_data]    | Downloading package hmm_treebank_pos_tagger to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package hmm_treebank_pos_tagger is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package maxent_treebank_pos_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package universal_tagset to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package universal_tagset is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package maxent_ne_chunker is already up-to-date!
[nltk_data]    | Downloading package punkt to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package punkt is already up-to-date!
[nltk_data]    | Downloading package book_grammars to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package book_grammars is already up-to-date!
[nltk_data]    | Downloading package sample_grammars to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sample_grammars is already up-to-date!
[nltk_data]    | Downloading package spanish_grammars to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package spanish_grammars is already up-to-date!
[nltk_data]    | Downloading package basque_grammars to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package basque_grammars is already up-to-date!
[nltk_data]    | Downloading package large_grammars to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package large_grammars is already up-to-date!
[nltk_data]    | Downloading package tagsets to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package tagsets is already up-to-date!
[nltk_data]    | Downloading package snowball_data to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package snowball_data is already up-to-date!
[nltk_data]    | Downloading package bllip_wsj_no_aux to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package bllip_wsj_no_aux is already up-to-date!
[nltk_data]    | Downloading package word2vec_sample to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package word2vec_sample is already up-to-date!
[nltk_data]    | Downloading package panlex_swadesh to
```

```
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    | Downloading package mte_teip5 to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package mte_teip5 is already up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package panlex_lite to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Package panlex_lite is already up-to-date!
[nltk_data]    | Downloading package perluniprops to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping misc\perluniprops.zip.
[nltk_data]    | Downloading package nonbreaking_prefixes to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping corpora\nonbreaking_prefixes.zip.
[nltk_data]    | Downloading package vader_lexicon to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    | Downloading package porter_test to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping stemmers\porter_test.zip.
[nltk_data]    | Downloading package wmt15_eval to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping models\wmt15_eval.zip.
[nltk_data]    | Downloading package mwa_ppdb to
[nltk_data]    |     C:\Users\PC\AppData\Roaming\nltk_data...
[nltk_data]    |   Unzipping misc\mwa_ppdb.zip.
[nltk_data]    |
[nltk_data]  Done downloading collection all

Out[2]:

True


In [3]:
```

```
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
```

Chapter 1

22.Find all the four-letter words in the Chat Corpus (text5). With the help of a frequency distribution (FreqDist), show these words in decreasing order of frequency.

In [53]:

```
words_4 = [w for w in (text5) if len(w) == 4]
fdist=FreqDist(words_4)
fdist.most_common()
```

Out[53]:

```
[('JOIN', 1021),
 ('PART', 1016),
 ('that', 274),
 ('what', 183),
 ('here', 181),
 ('....', 170),
 ('have', 164),
 ('like', 156),
 ('with', 152),
 ('chat', 142),
 ('your', 137),
 ('good', 130),
 ('just', 125),
 ('lmao', 107),
 ('know', 103),
 ('room', 98),
 ('from', 92),
 ('this', 86),
 ('well', 81),
 ('back', 78),
 ('hiya', 78),
 ('they', 77),
 ('yeah', 75),
 ('dont', 75),
 ('want', 71),
 ('love', 60),
 ('some', 58),
 ('guys', 58),
 ('been', 57),
 ('talk', 56),
 ('nice', 52),
 ('time', 50),
 ('when', 48),
 ('make', 44),
 ('haha', 44),
 ('need', 43),
 ('girl', 43),
 ('U122', 42),
 ('MODE', 41),
 ('much', 40),
 ('will', 40),
 ('then', 40),
 ('over', 39),
 ('work', 38),
 ('were', 38),
 ('take', 37),
 ('U121', 36),
 ('U115', 36),
 ('song', 36),
 ('U156', 35),
 ('even', 35),
 ('U105', 35),
 ('seen', 35),
 ('does', 35),
 ('more', 34),
 ('damn', 34),
 ('come', 33),
```

```
('only', 33),

('hell', 29),
('long', 28),
('them', 28),
('tell', 27),
('name', 27),
('away', 26),
('call', 26),
('look', 26),
('baby', 26),
('sure', 26),
('play', 25),
('U110', 25),
('U114', 25),
('cool', 24),
('down', 24),
('NICK', 24),
('many', 23),
('said', 23),
('sexy', 23),
('hate', 23),
('last', 22),
('ever', 22),
('life', 21),
('hear', 21),
('live', 20),
('very', 19),
('must', 19),
('LMAO', 19),
('give', 19),
('mean', 19),
('feel', 19),
('stop', 19),
('same', 19),
('cant', 18),
('What', 18),
('find', 18),
('hugs', 18),
('!!!!', 18),
('shit', 17),
('U104', 17),
('????', 17),
('nite', 17),
('left', 17),
('lost', 17),
('hair', 17),
('busy', 17),
('real', 16),
('game', 16),
('fine', 16),
('sits', 15),
('eyes', 15),
('heya', 15),
('lets', 15),
('kill', 15),
('fuck', 15),
('read', 14),
('wait', 14),
('shut', 14)
```

('shut', 14),
('true', 14),
('keep', 14),
('goes', 14),
('free', 13),
('near', 13),
('U168', 13),
('nope', 13),
('else', 13),
('pick', 13),
('male', 12),
('hehe', 12),
('told', 12),
('cold', 12),
('than', 12),
('used', 12),
('U102', 12),
('hope', 12),
('head', 12),
('gets', 12),
('This', 12),
('awww', 12),
('stay', 12),
('bout', 12),
('yall', 11),
('perv', 11),
('face', 11),
('babe', 11),
('doin', 11),
('home', 11),
('wont', 11),
('year', 11),
('U107', 11),
('.. .', 11),
('into', 11),
('kids', 11),
('U119', 11),
('mind', 10),
('once', 10),
('Liam', 10),
('Yeah', 10),
('U132', 10),
('week', 10),
('U101', 10),
('help', 10),
('Well', 10),
('hard', 10),
('show', 10),
('hmmm', 9),
('book', 9),
('nick', 9),
('type', 9),
('dang', 9),
('crap', 9),
('runs', 9),
('rock', 9),
('dead', 9),
('soon', 9),
('sick', 9),
('best', 9)

('best', 9),
('care', 9),
('days', 9),
('aint', 9),
('kiss', 9),
('mine', 9),
('pics', 9),
('neck', 9),
('; ..', 9),
('such', 9),
('full', 9),
('hour', 9),
('heyy', 8),
('suck', 8),
('U139', 8),
('U144', 8),
('hows', 8),
('sang', 8),
('blue', 8),
('lady', 8),
('word', 8),
('made', 8),
('wana', 8),
('says', 8),
('went', 8),
('case', 8),
('wife', 8),
('hand', 7),
('dude', 7),
('dear', 7),
('Hiya', 7),
('wear', 7),
('U108', 7),
('fast', 7),
('That', 7),
('alot', 7),
('okay', 7),
('U169', 7),
('took', 7),
('ahhh', 7),
('kick', 7),
('rule', 7),
('done', 6),
('U165', 6),
('whos', 6),
('comp', 6),
('sock', 6),
('sing', 6),
('U103', 6),
('Song', 6),
('))))', 6),
('poor', 6),
('part', 6),
('send', 6),
('pink', 6),
('blah', 6),
('U116', 6),
('ball', 6),
('goin', 6),
('last', 6)

('last', 6),
('oops', 6),
('main', 6),
('gone', 6),
('thru', 6),
('U129', 6),
('They', 6),
('U197', 6),
('next', 6),
('U120', 6),
('knew', 6),
('list', 6),
('U142', 6),
('food', 6),
('ride', 6),
('most', 6),
('seem', 6),
('U520', 6),
('<---', 6),
('miss', 5),
('beer', 5),
('lose', 5),
('boys', 5),
('boss', 5),
('fall', 5),
('luck', 5),
('Have', 5),
('heck', 5),
('ohhh', 5),
('idea', 5),
('feet', 5),
('also', 5),
('Lime', 5),
('meds', 5),
('till', 5),
('legs', 5),
('late', 5),
('When', 5),
('wish', 5),
('cali', 5),
('warm', 5),
('xbox', 5),
('came', 5),
('fool', 5),
('joke', 5),
('yoko', 5),
('hang', 5),
('nose', 5),
('wall', 5),
('U128', 5),
('####', 5),
('lick', 5),
('kool', 5),
('both', 5),
('soul', 5),
('land', 5),
('meet', 5),
('caps', 5),
('fire', 5),
('roll', 5)

('foil', 5),
('easy', 5),
('felt', 5),
('pass', 4),
('evil', 4),
('ouch', 4),
('fart', 4),
('mmmm', 4),
('door', 4),
('high', 4),
('quit', 4),
('shot', 4),
('turn', 4),
('each', 4),
('U126', 4),
('kent', 4),
('U988', 4),
('team', 4),
('ways', 4),
('U146', 4),
('U133', 4),
('hook', 4),
('U219', 4),
('lord', 4),
('glad', 4),
('beat', 4),
('U130', 4),
('U154', 4),
('sigh', 4),
('lame', 4),
('cute', 4),
('self', 4),
('ones', 4),
('Like', 4),
('ROOM', 4),
('jerk', 4),
('none', 4),
('line', 4),
('U117', 4),
('shes', 4),
('huge', 4),
('U819', 4),
('U123', 4),
('pain', 4),
(',,,,', 4),
('puff', 4),
('pfft', 4),
('U989', 4),
('U820', 4),
('open', 4),
('holy', 4),
('U196', 4),
('rest', 4),
('ummm', 4),
('grrr', 4),
('date', 4),
('ugly', 4),
('woot', 4),
('deop', 3),
('gold', 3)

('gold', 3),
('toss', 3),
('vote', 3),
('mary', 3),
('U141', 3),
('hank', 3),
('deal', 3),
('amen', 3),
('half', 3),
('((((', 3),
('elle', 3),
('nana', 3),
('wash', 3),
('Only', 3),
('orgy', 3),
('piff', 3),
('rubs', 3),
('bend', 3),
('DING', 3),
('wazz', 3),
('U109', 3),
('yawn', 3),
('band', 3),
('skin', 3),
('slow', 3),
('CHAT', 3),
('note', 3),
('jump', 3),
('ahem', 3),
('butt', 3),
('move', 3),
('hiii', 3),
('itch', 3),
('died', 3),
('Elev', 3),
('tune', 3),
('snow', 3),
('U106', 3),
('hola', 3),
('guyz', 3),
('imma', 3),
('U145', 3),
('Wind', 3),
('hick', 3),
('rain', 3),
('bare', 3),
('army', 3),
('clap', 3),
('walk', 3),
('AKDT', 3),
('soft', 3),
('toes', 3),
('U136', 3),
('isnt', 3),
('ring', 3),
('Same', 3),
('road', 3),
('swim', 3),
('ello', 3),
('U163', 3)

('U103', 3),
('hump', 3),
('gawd', 3),
('lead', 3),
('wack', 3),
('hawt', 3),
('U148', 3),
('Your', 3),
('2006', 3),
('roof', 3),
('THAT', 3),
('wine', 3),
('slap', 3),
('hail', 3),
('U153', 3),
('yada', 3),
('hurt', 3),
('town', 3),
('flaw', 2),
('pies', 2),
('gays', 2),
('five', 2),
('temp', 2),
('wOOt', 2),
('john', 2),
('Ohio', 2),
('KoOL', 2),
('plan', 2),
('sell', 2),
('Just', 2),
('U100', 2),
('shop', 2),
('born', 2),
('Lmao', 2),
('adds', 2),
('typo', 2),
('eats', 2),
('DONT', 2),
('Gosh', 2),
('NONE', 2),
('newp', 2),
('mama', 2),
('Heyy', 2),
('opps', 2),
('ohio', 2),
('Days', 2),
('cmon', 2),
('tyvm', 2),
('root', 2),
('ciao', 2),
('whoa', 2),
('tisk', 2),
('Tell', 2),
('Love', 2),
('O.k.', 2),
('U112', 2),
('deaf', 2),
('blew', 2),
('any1', 2),
('rofl', 2)

('?!?!', 2),
('STOP', 2),
('Okay', 2),
('rent', 2),
('Ummm', 2),
('drew', 2),
('Here', 2),
('1996', 2),
('Dang', 2),
('ltns', 2),
('eric', 2),
('haze', 2),
('hmph', 2),
('HAVE', 2),
('Nice', 2),
('rich', 2),
('Come', 2),
('past', 2),
('>:->', 2),
('cost', 2),
('lies', 2),
('Stop', 2),
('park', 2),
('Lets', 2),
('FROM', 2),
('Tisk', 2),
('cars', 2),
('sooo', 2),
('club', 2),
('howz', 2),
('sort', 2),
('chip', 2),
('hint', 2),
('argh', 2),
('kewl', 2),
('bear', 2),
('doll', 2),
('wats', 2),
('cast', 2),
('sand', 2),
('?!?!', 2),
('yeas', 2),
('hold', 2),
('U172', 2),
('mike', 2),
('limp', 2),
('gimp', 2),
('Down', 2),
('uses', 2),
('??!!', 2),
('hott', 2),
('U190', 2),
('aunt', 2),
('tock', 2),
('phil', 2),
('<333', 2),
('babi', 2),
('tick', 2),
('WITH', 2),
('pool', 2)

('pool', 2),
('ewww', 2),
('U155', 2),
('foot', 2),
('<<<<', 2),
('cash', 2),
('meat', 2),
('luvs', 2),
('ages', 2),
('twin', 2),
('Ahhh', 2),
('mins', 2),
('whip', 2),
('hall', 2),
('DOES', 2),
('U138', 2),
('zone', 2),
('spot', 2),
('golf', 2),
('kind', 2),
('U175', 2),
('Live', 2),
('U111', 2),
('hits', 2),
('High', 2),
('bite', 2),
('From', 2),
('Drew', 2),
('area', 2),
('moon', 2),
('city', 2),
('side', 2),
('Lies', 2),
('lawl', 2),
('corn', 2),
('flow', 2),
('sore', 2),
('trip', 2),
('Cool', 2),
('burp', 2),
('porn', 2),
('Poor', 2),
('grrl', 2),
('John', 2),
('whud', 2),
('heal', 2),
('drop', 2),
('wooo', 2),
('spin', 2),
("ex's", 2),
('!!!.', 2),
('yard', 2),
('n9ne', 2),
('fits', 2),
('dumb', 2),
('ears', 2),
('clue', 2),
('deep', 2),
('mass', 2),
('U170', 2)

('U170', 2),
('humm', 2),
('YOUR', 2),
('Sure', 2),
('cell', 2),
('drug', 1),
('wood', 1),
('bomb', 1),
('jeff', 1),
('Care', 1),
('docs', 1),
('U134', 1),
('HOTT', 1),
('hots', 1),
('Hand', 1),
('mofo', 1),
('eeww', 1),
('thnx', 1),
('asss', 1),
('owww', 1),
('Hard', 1),
('nods', 1),
('givs', 1),
('lapd', 1),
('Troy', 1),
('sayn', 1),
('dyed', 1),
('Jess', 1),
('tiff', 1),
('Deep', 1),
('bust', 1),
('febe', 1),
('choc', 1),
('clay', 1),
(':o *', 1),
('urls', 1),
('gift', 1),
('SExy', 1),
('scar', 1),
('halo', 1),
('ebay', 1),
('pray', 1),
('MRIs', 1),
('heat', 1),
('lol.', 1),
('cure', 1),
('poem', 1),
('bike', 1),
('evah', 1),
('SIZE', 1),
('serg', 1),
('yes.', 1),
('safe', 1),
('keys', 1),
('boed', 1),
('weed', 1),
('cums', 1),
('CAPS', 1),
('tere', 1),
('1 98', 1)

('1.98', 1),
('bell', 1),
('prep', 1),
('nads', 1),
('bacl', 1),
('smax', 1),
('Male', 1),
('yell', 1),
('2DAY', 1),
('mauh', 1),
('benz', 1),
('seat', 1),
('tenn', 1),
('Teck', 1),
('Mine', 1),
('dump', 1),
('JUST', 1),
('18ST', 1),
('chit', 1),
('nawt', 1),
('beam', 1),
('poll', 1),
('jush', 1),
('sexs', 1),
('howl', 1),
('base', 1),
('plow', 1),
('HERE', 1),
('waaa', 1),
('firs', 1),
('lois', 1),
('lisa', 1),
('ther', 1),
('pope', 1),
('COME', 1),
('fock', 1),
('Dawn', 1),
('out.', 1),
('wire', 1),
('soup', 1),
('outa', 1),
('3:45', 1),
('cuss', 1),
('hill', 1),
('ussy', 1),
('pair', 1),
('wide', 1),
('wore', 1),
('teck', 1),
('nuff', 1),
('rose', 1),
('pure', 1),
('ques', 1),
('raed', 1),
('woof', 1),
('ltnc', 1),
('Bone', 1),
('Very', 1),
('gooo', 1),
('prob', 1)

('prob', 1),
('caca', 1),
('west', 1),
('pigs', 1),
('Dude', 1),
('bong', 1),
('98.5', 1),
('yess', 1),
('Lord', 1),
('span', 1),
('Hill', 1),
('hooo', 1),
('scuk', 1),
('Iowa', 1),
('Kold', 1),
('brwn', 1),
('nawp', 1),
('SOME', 1),
('1200', 1),
('cook', 1),
('Lion', 1),
('Heya', 1),
('2Pac', 1),
('Awww', 1),
('THEY', 1),
('U181', 1),
('Damn', 1),
('moms', 1),
('YALL', 1),
('outs', 1),
('100%', 1),
('Hugs', 1),
('Matt', 1),
('WILL', 1),
('9:10', 1),
('GOOD', 1),
('blow', 1),
('brat', 1),
('tory', 1),
('brad', 1),
('mode', 1),
("ok'd", 1),
('tooo', 1),
('lyin', 1),
('daft', 1),
('bull', 1),
('gals', 1),
('Hott', 1),
('Need', 1),
('Eyes', 1),
('HUGE', 1),
('KNOW', 1),
('okey', 1),
("pm'n", 1),
('!...', 1),
('!???', 1),
('duet', 1),
('ohwa', 1),
('puke', 1),
('Elle', 1)

('Elle', 1),
('U147', 1),
('Phil', 1),
('tall', 1),
('aime', 1),
('ogan', 1),
('kold', 1),
('calm', 1),
('ally', 1),
('Heys', 1),
('64.8', 1),
('slam', 1),
('herd', 1),
('Yoko', 1),
('mkay', 1),
('Kids', 1),
('Good', 1),
('bred', 1),
('10th', 1),
('bone', 1),
('Take', 1),
('mess', 1),
('Boyz', 1),
('puts', 1),
('otay', 1),
('Kick', 1),
('AWAY', 1),
('loud', 1),
('push', 1),
('goof', 1),
('wild', 1),
('Long', 1),
('dawg', 1),
('LONG', 1),
('1299', 1),
('tjhe', 1),
('abou', 1),
('boom', 1),
('able', 1),
('ghet', 1),
('tits', 1),
('guns', 1),
('kept', 1),
('peek', 1),
('U118', 1),
('VBox', 1),
('toop', 1),
('soda', 1),
('LIVE', 1),
('Time', 1),
('poot', 1),
('gees', 1),
('cepn', 1),
('thje', 1),
('hide', 1),
('cams', 1),
('surf', 1),
('acid', 1),
('Evil', 1),
('bloe', 1)

('bide', 1),
('Away', 1),
('Ctrl', 1),
('Then', 1),
("PM's", 1),
('Girl', 1),
('TALK', 1),
('mame', 1),
('idnt', 1),
('Uhhh', 1),
('gosh', 1),
('mite', 1),
('LATE', 1),
('98.6', 1),
('HALO', 1),
('http', 1),
('wind', 1),
('noth', 1),
('rang', 1),
('Chop', 1),
('tlak', 1),
('arms', 1),
('spat', 1),
('Judy', 1),
('News', 1),
('jack', 1),
('seth', 1),
('xmas', 1),
('jail', 1),
('term', 1),
('vent', 1),
('nada', 1),
('PMSL', 1),
('coat', 1),
('fair', 1),
('coem', 1),
('four', 1),
('sexi', 1),
('U113', 1),
('hong', 1),
('gear', 1),
('gags', 1),
('grea', 1),
('scum', 1),
('enuf', 1),
('worl', 1),
('pork', 1),
('barn', 1),
('Rang', 1),
('Show', 1),
('Came', 1),
('Food', 1),
('Been', 1),
('bied', 1),
('Reub', 1),
('gret', 1),
('6:51', 1),
('ahah', 1),
('bois', 1),
('kmph', 1)

('Kmph', 1),
('York', 1),
('U149', 1),
('King', 1),
('asks', 1),
('samn', 1),
('waht', 1),
('Holy', 1),
('Dood', 1),
('rats', 1),
('mang', 1),
('LoVe', 1),
('LOUD', 1),
('yeee', 1),
('wrap', 1),
('Type', 1),
('VVil', 1),
('Pour', 1),
('allo', 1),
("pm's", 1),
('NAME', 1),
('Fade', 1),
('bird', 1),
('ELSE', 1),
('Drop', 1),
('geez', 1),
('Swim', 1),
('loss', 1),
('ruff', 1),
('exit', 1),
('buff', 1),
('Tiff', 1),
('syck', 1),
('TIME', 1),
('Ohhh', 1),
('Even', 1),
('Turn', 1),
('hogs', 1),
('Werd', 1),
('GrlZ', 1),
('wuts', 1),
('sum1', 1),
('woah', 1),
('Nope', 1),
('thot', 1),
('tips', 1),
('Talk', 1),
('orta', 1),
('ssid', 1),
('rush', 1),
('Tina', 1),
('6:38', 1),
('sori', 1),
('LAst', 1),
('Rule', 1),
('Hail', 1),
('fawk', 1),
('Haha', 1),
('Rock', 1),
('MORE', 1)

('MORE', 1),
('dick', 1),
('inch', 1),
('akon', 1),
('SSRI', 1),
('dint', 1),
('paid', 1),
('Look', 1),
('Born', 1),
('page', 1),
('4:03', 1),
('wins', 1),
('NTMN', 1),
('wher', 1),
('dotn', 1),
('WHOA', 1),
('sext', 1),
('kong', 1),
('SEEN', 1),
('haaa', 1),
('vamp', 1),
('junk', 1),
('whew', 1),
('1cos', 1),
('gray', 1),
('cops', 1),
('lots', 1),
('Sat.', 1),
('U158', 1),
('Rofl', 1),
('Rush', 1),
('Over', 1),
('herE', 1),
('Slip', 1),
('1980', 1),
('jude', 1),
('sent', 1),
('este', 1),
('pull', 1),
('Kiss', 1),
('WHEN', 1),
('7:45', 1),
('U164', 1),
('DAMN', 1),
('cock', 1),
('Mono', 1),
('ribs', 1),
('AKST', 1),
('star', 1),
("yw's", 1),
('twit', 1),
('Kent', 1),
('Save', 1),
('laid', 1),
('2:55', 1),
('yesh', 1),
('rape', 1),
('kina', 1),
('Oops', 1),
]

```
...」
```

26.What does the following Python code do? sum(len(w) for w in text1) Can you use it to work out the average word length of a text?

In [62]:

```
#Suma las longitudes de cada palabra en el texto 1.
sum([len(w) for w in text1]) / len(text1)
```

Out[62]:

```
3.830411128023649
```

28.Define a function percent(word, text) that calculates how often a given word occurs in a text, and expresses the result as a percentage.

In [70]:

```
def percent(word, text):
    return 100 * text.count(word) / len(text)
print (str(percent('monstrous', text1)) + '%')
```

```
0.00383407650516258 4%
```

Chapter 2

2.Use the corpus module to explore austen-persuasion.txt. How many word tokens does this book have? How many word types?

In [93]:

```
nltk.corpus.gutenberg.fileids()
```

Out[93]:

```
['austen-emma.txt',
 'austen-persuasion.txt',
 'austen-sense.txt',
 'bible-kjv.txt',
 'blake-poems.txt',
 'bryant-stories.txt',
 'burgess-busterbrown.txt',
 'carroll-alice.txt',
 'chesterton-ball.txt',
 'chesterton-brown.txt',
 'chesterton-thursday.txt',
 'edgeworth-parents.txt',
 'melville-moby_dick.txt',
 'milton-paradise.txt',
 'shakespeare-caesar.txt',
 'shakespeare-hamlet.txt',
 'shakespeare-macbeth.txt',
 'whitman-leaves.txt']
```

In [98]:

```
austen = gutenberg.words('austen-persuasion.txt')
len(austen)
```

Out[98]:

98171

In [102]:

```
len(set(austen))
```

Out[102]:

6132

4.Read in the texts of the State of the Union addresses, using the state_union corpus reader. Count occurrences of men, women, and people in each document. What has happened to the usage of these words over time?

In [103]:

```
nltk.corpus.state_union.fileids()
```

Out[103]:

```
['1945-Truman.txt',
 '1946-Truman.txt',
 '1947-Truman.txt',
 '1948-Truman.txt',
 '1949-Truman.txt',
 '1950-Truman.txt',
 '1951-Truman.txt',
 '1953-Eisenhower.txt',
 '1954-Eisenhower.txt',
 '1955-Eisenhower.txt',
 '1956-Eisenhower.txt',
 '1957-Eisenhower.txt',
 '1958-Eisenhower.txt',
 '1959-Eisenhower.txt',
 '1960-Eisenhower.txt',
 '1961-Kennedy.txt',
 '1962-Kennedy.txt',
 '1963-Johnson.txt',
 '1963-Kennedy.txt',
 '1964-Johnson.txt',
 '1965-Johnson-1.txt',
 '1965-Johnson-2.txt',
 '1966-Johnson.txt',
 '1967-Johnson.txt',
 '1968-Johnson.txt',
 '1969-Johnson.txt',
 '1970-Nixon.txt',
 '1971-Nixon.txt',
 '1972-Nixon.txt',
 '1973-Nixon.txt',
 '1974-Nixon.txt',
 '1975-Ford.txt',
 '1976-Ford.txt',
 '1977-Ford.txt',
 '1978-Carter.txt',
 '1979-Carter.txt',
 '1980-Carter.txt',
 '1981-Reagan.txt',
 '1982-Reagan.txt',
 '1983-Reagan.txt',
 '1984-Reagan.txt',
 '1985-Reagan.txt',
 '1986-Reagan.txt',
 '1987-Reagan.txt',
 '1988-Reagan.txt',
 '1989-Bush.txt',
 '1990-Bush.txt',
 '1991-Bush-1.txt',
 '1991-Bush-2.txt',
 '1992-Bush.txt',
 '1993-Clinton.txt',
 '1994-Clinton.txt',
 '1995-Clinton.txt',
 '1996-Clinton.txt',
 '1997-Clinton.txt',
 '1998-Clinton.txt',
 '1999-Clinton.txt',
```

```
'2000-Clinton.txt',

'2001-GWBush-1.txt',
'2001-GWBush-2.txt',
'2002-GWBush.txt',
'2003-GWBush.txt',
'2004-GWBush.txt',
'2005-GWBush.txt',
'2006-GWBush.txt']
```

In [124]:

```
for fileid in state_union.fileids():
    print(fileid,"men=", FreqDist(fileid)['men'])
```

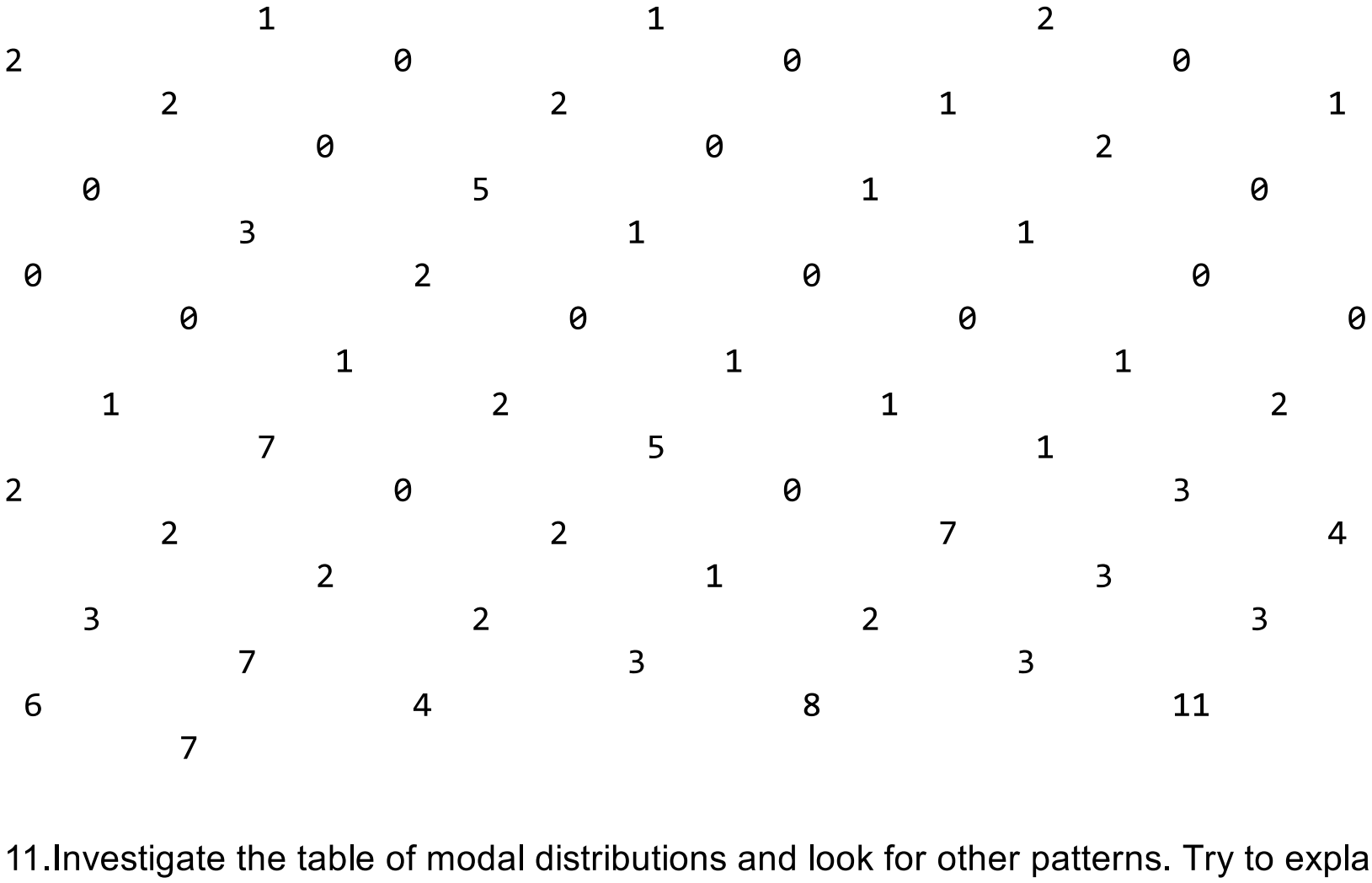
```
1945-Truman.txt men= 0
1946-Truman.txt men= 0
1947-Truman.txt men= 0
1948-Truman.txt men= 0
1949-Truman.txt men= 0
1950-Truman.txt men= 0
1951-Truman.txt men= 0
1953-Eisenhower.txt men= 0
1954-Eisenhower.txt men= 0
1955-Eisenhower.txt men= 0
1956-Eisenhower.txt men= 0
1957-Eisenhower.txt men= 0
1958-Eisenhower.txt men= 0
1959-Eisenhower.txt men= 0
1960-Eisenhower.txt men= 0
1961-Kennedy.txt men= 0
1962-Kennedy.txt men= 0
1963-Johnson.txt men= 0
1963-Kennedy.txt men= 0
1964-Johnson.txt men= 0
1965-Johnson-1.txt men= 0
1965-Johnson-2.txt men= 0
1966-Johnson.txt men= 0
1967-Johnson.txt men= 0
1968-Johnson.txt men= 0
1969-Johnson.txt men= 0
1970-Nixon.txt men= 0
1971-Nixon.txt men= 0
1972-Nixon.txt men= 0
1973-Nixon.txt men= 0
1974-Nixon.txt men= 0
1975-Ford.txt men= 0
1976-Ford.txt men= 0
1977-Ford.txt men= 0
1978-Carter.txt men= 0
1979-Carter.txt men= 0
1980-Carter.txt men= 0
1981-Reagan.txt men= 0
1982-Reagan.txt men= 0
1983-Reagan.txt men= 0
1984-Reagan.txt men= 0
1985-Reagan.txt men= 0
1986-Reagan.txt men= 0
1987-Reagan.txt men= 0
1988-Reagan.txt men= 0
1989-Bush.txt men= 0
1990-Bush.txt men= 0
1991-Bush-1.txt men= 0
1991-Bush-2.txt men= 0
1992-Bush.txt men= 0
1993-Clinton.txt men= 0
1994-Clinton.txt men= 0
1995-Clinton.txt men= 0
1996-Clinton.txt men= 0
1997-Clinton.txt men= 0
1998-Clinton.txt men= 0
1999-Clinton.txt men= 0
2000-Clinton.txt men= 0
```

```
2001-GWBush-1.txt men= 0
2001-GWBush-2.txt men= 0
2002-GWBush.txt men= 0
2003-GWBush.txt men= 0
2004-GWBush.txt men= 0
2005-GWBush.txt men= 0
2006-GWBush.txt men= 0
```

In [115]:

```
cfd = nltk.ConditionalFreqDist(
    (target, fileid[:])
    for fileid in state_union.fileids()
    for w in state_union.words(fileid)
    for target in ['men', 'women', 'people']
    if w.lower() == target)
cfd.tabulate()
```

1945-Truman.txt 1946-Truman.txt 1947-Truman.txt 1
948-Truman.txt 1949-Truman.txt 1950-Truman.txt 1951-Truman.t
xt 1953-Eisenhower.txt 1954-Eisenhower.txt 1955-Eisenhower.txt 1956-Eise
nhower.txt 1957-Eisenhower.txt 1958-Eisenhower.txt 1959-Eisenhower.txt 1
960-Eisenhower.txt 1961-Kennedy.txt 1962-Kennedy.txt 1963-Johns
on.txt 1963-Kennedy.txt 1964-Johnson.txt 1965-Johnson-1.txt 1965
-Johnson-2.txt 1966-Johnson.txt 1967-Johnson.txt 1968-Johnson.t
xt 1969-Johnson.txt 1970-Nixon.txt 1971-Nixon.txt 1972
-Nixon.txt 1973-Nixon.txt 1974-Nixon.txt 1975-Ford.txt
1976-Ford.txt 1977-Ford.txt 1978-Carter.txt 1979-Cart
er.txt 1980-Carter.txt 1981-Reagan.txt 1982-Reagan.txt 1
983-Reagan.txt 1984-Reagan.txt 1985-Reagan.txt 1986-Reagan.t
xt 1987-Reagan.txt 1988-Reagan.txt 1989-Bush.txt 199
0-Bush.txt 1991-Bush-1.txt 1991-Bush-2.txt 1992-Bush.txt
1993-Clinton.txt 1994-Clinton.txt 1995-Clinton.txt 1996-Clint
on.txt 1997-Clinton.txt 1998-Clinton.txt 1999-Clinton.txt 20
00-Clinton.txt 2001-GWBush-1.txt 2001-GWBush-2.txt 2002-GWBush.t
xt 2003-GWBush.txt 2004-GWBush.txt 2005-GWBush.txt 2006-
GWBush.txt
men 2 12 7
5 2 6
8 3 2 4
2 5 2 4
2 6 6
0 8 3 7
12 12 11
4 5 2 1
1 1 0 0
3 2 0
0 1 1 1
3 3 1
2 1 1 2
3 2 7 4
1 1 1
2 1 2 2
5 3 1
3 6 7 8
7
people 10 49 12
22 15 15
10 17 15 26
30 11 19 11
10 10 10
3 12 3 16
14 35 25
17 6 23 32
7 9 20 14
18 19 26
15 12 11 17
19 27 12
14 24 17 13
9 14 13 27
45 66 73
43 31 22 22
41 15 12
14 33 21 18
22
women 2 7 2



11.Investigate the table of modal distributions and look for other patterns. Try to explain them in terms of your own impressionistic understanding of the different genres. Can you find other closed classes of words that exhibit significant differences across different genres?

In [127]:

```
brown.categories()
```

Out[127]:

```
['adventure',
 'belles_lettres',
 'editorial',
 'fiction',
 'government',
 'hobbies',
 'humor',
 'learned',
 'lore',
 'mystery',
 'news',
 'religion',
 'reviews',
 'romance',
 'science_fiction']
```

In [131]:

```
cfd = nltk.ConditionalFreqDist(
    (genre, word)
    for genre in brown.categories()
    for word in brown.words(categories=genre))
genres = ['news', 'religion', 'hobbies', 'science_fiction', 'romance',
          'humor','adventure','belles_lettres','editorial', 'fiction','government','l
earned','lore','mystery','reviews']
modals = ['can', 'could', 'may', 'might', 'must', 'will']
cfd.tabulate(conditions=genres, samples=modals)
```

```
                  can could   may might  must  will
           news    93    86    66    38    50   389
       religion    82    59    78    12    54    71
        hobbies   268    58   131    22    83   264
science_fiction    16    49     4    12     8    16
        romance    74   193    11    51    45    43
          humor    16    30     8     8     9    13
      adventure    46   151     5    58    27    50
 belles_lettres   246   213   207   113   170   236
      editorial   121    56    74    39    53   233
        fiction    37   166     8    44    55    52
     government   117    38   153    13   102   244
        learned   365   159   324   128   202   340
           lore   170   141   165    49    96   175
        mystery    42   141    13    57    30    20
        reviews    45    40    45    26    19    58
```

-Usando todos los generos, podemos ver por ejemplo que el govierno usa más will, lo sería coherente con la idea de que los gobiernos justifican sus acciones por lo que ellos consideran que estas van a lograr. -En el texto del genero de misterio la palabra que más se usa es could (podría), lo que prodia indicar que lo releante es la incertidumbre sobre cualquier evento y las posible consecuencias.

In [137]:

```
cfd = nltk.ConditionalFreqDist(
    (genre, word)
    for genre in brown.categories()
    for word in brown.words(categories=genre))
genres = ['news', 'religion', 'hobbies', 'science_fiction', 'romance',
          'humor','adventure','belles_lettres','editorial', 'fiction','government','l
earned','lore','mystery','reviews']
modals = ['dead','love', 'pain']
cfd.tabulate(conditions=genres, samples=modals)
```

```
                dead love pain
          news     8    3    1
      religion     9   13    3
       hobbies     3    6    0
science_fiction    7    3    7
       romance    15   32    5
         humor     3    4    1
     adventure    25    9    9
 belles_lettres   20   68    4
     editorial     5   13    2
       fiction    19   16   10
    government     1    1    0
       learned    12   13   18
          lore    15   19   19
       mystery    21    7    8
       reviews     3    7    0
```

-Usando las palabras claves love, pain y dead, vemos que la palabra love es principalmente usada en belles_lettres y romance la palabra pues obviamente en estas se idealiza las relaciones con el amor. Dead es usada principalmente en mystery y aventura pues en estos textos la muerte de un individuo es muy llamativo y provoca mayor interes en el lector.