Taller 9

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 26-abr-2019 11:59 PM

Juan Sebastián Valbuena Silva

juans.valbuena@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/), Exercises:

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

In [1]:

```
import nltk
```

In [21]:

```
dler = nltk.downloader.Downloader()
dler._update_index()
dler._status_cache['panlex_lite'] = 'installed' # Trick the index to treat panlex_lite 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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package alpino to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package alpino is already up-to-date!
[nltk_data]    | Downloading package biocreative_ppi to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package biocreative_ppi is already up-to-date!
[nltk_data]    | Downloading package brown to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package brown is already up-to-date!
[nltk_data]    | Downloading package brown_tei to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package brown_tei is already up-to-date!
```

```
[nltk_data]    |   Package brown_tei is already up-to-date!
[nltk_data]    | Downloading package cess_cat to
[nltk_data]    |     C:\Users\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cess_esp is already up-to-date!
[nltk_data]    | Downloading package chat80 to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package chat80 is already up-to-date!
[nltk_data]    | Downloading package city_database to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package city_database is already up-to-date!
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package comparative_sentences to
[nltk_data]    |     C:\Users\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package comtrans is already up-to-date!
[nltk_data]    | Downloading package conll2000 to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2000 is already up-to-date!
[nltk_data]    | Downloading package conll2002 to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2002 is already up-to-date!
[nltk_data]    | Downloading package conll2007 to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package conll2007 is already up-to-date!
[nltk_data]    | Downloading package crubadan to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package crubadan is already up-to-date!
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package dependency_treebank is already up-to-date!
[nltk_data]    | Downloading package dolch to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package dolch is already up-to-date!
[nltk_data]    | Downloading package europarl_raw to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package europarl_raw is already up-to-date!
[nltk_data]    | Downloading package floresta to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package floresta is already up-to-date!
[nltk_data]    | Downloading package framenet_v15 to
[nltk_data]    |     C:\Users\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package framenet_v17 is already up-to-date!
[nltk_data]    | Downloading package gazetteers to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package gazetteers is already up-to-date!
[nltk_data]    | Downloading package genesis to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package genesis is already up-to-date!
[nltk_data]    | Downloading package gutenberg to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package gutenberg is already up-to-date!
[nltk_data]    | Downloading package ieer to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ieer is already up-to-date!
[nltk_data]    | Downloading package inaugural to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package inaugural is already up-to-date!
[nltk_data]    | Downloading package indian to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package indian is already up-to-date!
[nltk_data]    | Downloading package jeita to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package jeita is already up-to-date!
[nltk_data]    | Downloading package kimmo to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package kimmo is already up-to-date!
[nltk_data]    | Downloading package knbc to
```

```
[nltk_data]    | Downloading package knbc to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package knbc is already up-to-date!
[nltk_data]    | Downloading package lin_thesaurus to
[nltk_data]    |     C:\Users\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package mac_morpho is already up-to-date!
[nltk_data]    | Downloading package machado to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package machado is already up-to-date!
[nltk_data]    | Downloading package masc_tagged to
[nltk_data]    |     C:\Users\usuario\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\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package movie_reviews is already up-to-date!
[nltk_data]    | Downloading package names to
[nltk_data]    |     C:\Users\usuario\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\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package nps_chat is already up-to-date!
[nltk_data]    | Downloading package omw to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package omw is already up-to-date!
[nltk_data]    | Downloading package opinion_lexicon to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package opinion_lexicon is already up-to-date!
[nltk_data]    | Downloading package paradigms to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package paradigms is already up-to-date!
[nltk_data]    | Downloading package pil to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package pil is already up-to-date!
[nltk_data]    | Downloading package pl196x to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package pl196x is already up-to-date!
[nltk_data]    | Downloading package ppattach to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ppattach is already up-to-date!
[nltk_data]    | Downloading package problem_reports to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package problem_reports is already up-to-date!
[nltk_data]    | Downloading package propbank to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package propbank is already up-to-date!
[nltk_data]    | Downloading package ptb to
[nltk_data]    |     C:\Users\usuario\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\usuario\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\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package pros_cons is already up-to-date!
[nltk_data]    | Downloading package qc to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package qc is already up-to-date!
[nltk_data]    | Downloading package reuters to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package reuters is already up-to-date!
[nltk_data]    | Downloading package rte to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package rte is already up-to-date!
[nltk_data]    | Downloading package semcor to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package semcor is already up-to-date!
```

```
[nltk_data]    |   Package semcor is already up-to-date!
[nltk_data]    | Downloading package senseval to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package senseval is already up-to-date!
[nltk_data]    | Downloading package sentiwordnet to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sentiwordnet is already up-to-date!
[nltk_data]    | Downloading package sentence_polarity to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sentence_polarity is already up-to-date!
[nltk_data]    | Downloading package shakespeare to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package shakespeare is already up-to-date!
[nltk_data]    | Downloading package sinica_treebank to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package sinica_treebank is already up-to-date!
[nltk_data]    | Downloading package smultron to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package smultron is already up-to-date!
[nltk_data]    | Downloading package state_union to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package state_union is already up-to-date!
[nltk_data]    | Downloading package stopwords to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package stopwords is already up-to-date!
[nltk_data]    | Downloading package subjectivity to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package subjectivity is already up-to-date!
[nltk_data]    | Downloading package swadesh to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package swadesh is already up-to-date!
[nltk_data]    | Downloading package switchboard to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package switchboard is already up-to-date!
[nltk_data]    | Downloading package timit to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package timit is already up-to-date!
[nltk_data]    | Downloading package toolbox to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package toolbox is already up-to-date!
[nltk_data]    | Downloading package treebank to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package treebank is already up-to-date!
[nltk_data]    | Downloading package twitter_samples to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package twitter_samples is already up-to-date!
[nltk_data]    | Downloading package udhr to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package udhr is already up-to-date!
[nltk_data]    | Downloading package udhr2 to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package udhr2 is already up-to-date!
[nltk_data]    | Downloading package unicode_samples to
[nltk_data]    |     C:\Users\usuario\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\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package verbnet is already up-to-date!
[nltk_data]    | Downloading package verbnet3 to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package verbnet3 is already up-to-date!
[nltk_data]    | Downloading package webtext to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package webtext is already up-to-date!
[nltk_data]    | Downloading package wordnet to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet is already up-to-date!
[nltk_data]    | Downloading package wordnet_ic to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet_ic is already up-to-date!
[nltk_data]    | Downloading package words to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package words is already up-to-date!
```

```
[nltk_data]    | Downloading package ycoe to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package ycoe is already up-to-date!
[nltk_data]    | Downloading package rslp to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package rslp is already up-to-date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger to
[nltk_data]    |     C:\Users\usuario\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\usuario\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\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package punkt is already up-to-date!
[nltk_data]    | Downloading package book_grammars to
[nltk_data]    |     C:\Users\usuario\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\usuario\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\usuario\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\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package large_grammars is already up-to-date!
[nltk_data]    | Downloading package tagsets to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package tagsets is already up-to-date!
[nltk_data]    | Downloading package snowball_data to
[nltk_data]    |     C:\Users\usuario\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\usuario\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\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package panlex_swadesh is already up-to-date!
[nltk_data]    | Downloading package mte_teip5 to
[nltk_data]    |     C:\Users\usuario\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\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger_ru to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger_ru is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package perluniprops to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package perluniprops is already up-to-date!
[nltk_data]    | Downloading package nonbreaking_prefixes to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package nonbreaking_prefixes is already up-to-date!
[nltk_data]    | Downloading package vader_lexicon to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package vader_lexicon is already up-to-date!
[nltk_data]    | Downloading package porter_test to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package porter_test is already up-to-date!
[nltk_data]    | Downloading package wmt15_eval to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wmt15_eval is already up-to-date!
[nltk_data]    | Downloading package mwa_ppdb to
[nltk_data]    |     C:\Users\usuario\AppData\Roaming\nltk_data...
```

```
[nltk_data]    |   Package mwa_ppdb is already up-to-date!
[nltk_data]    |
[nltk_data]  Done downloading collection all
```

Out[21]:

```
True
```

In [32]:

```
from nltk.book import *
```

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 [33]:

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

Out[33]:

```
[('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),
 ('dont', 75),
 ('yeah', 75),
 ('want', 71),
 ('love', 60),
 ('guys', 58),
 ('some', 58),
 ('been', 57),
 ('talk', 56),
 ('nice', 52),
 ('time', 50),
 ('when', 48),
 ('haha', 44),
 ('make', 44),
 ('girl', 43),
 ('need', 43),
 ('U122', 42),
 ('MODE', 41),
 ('will', 40),
 ('much', 40),
 ('then', 40),
 ('over', 39),
 ('work', 38),
 ('were', 38),
 ('take', 37),
 ('U121', 36),
 ('U115', 36),
    ('song', 36),
    ('even', 35),
    ('does', 35),
    ('seen', 35),
    ('U156', 35),
    ('U105', 35),
    ('more', 34),
    ('damn', 34),
    ('only', 33),
    ('come', 33),
    ('hell', 29),
    ('long', 28),
    ('them', 28),
    ('name', 27),
    ('tell', 27),
    ('away', 26),
    ('sure', 26),
    ('look', 26),
    ('baby', 26),
    ('call', 26),
    ('play', 25),
    ('U110', 25),
    ('U114', 25),
    ('NICK', 24),
    ('down', 24),
    ('cool', 24),
    ('sexy', 23),
    ('many', 23),
    ('hate', 23),
    ('said', 23),
    ('last', 22),
    ('ever', 22),
    ('hear', 21),
    ('life', 21),
    ('live', 20),
    ('feel', 19),
    ('very', 19),
    ('mean', 19),
    ('give', 19),
    ('same', 19),
    ('must', 19),
    ('stop', 19),
    ('LMAO', 19),
    ('!!!!', 18),
    ('hugs', 18),
    ('What', 18),
    ('find', 18),
    ('cant', 18),
    ('left', 17),
    ('????', 17),
    ('shit', 17),
    ('nite', 17),
    ('busy', 17),
    ('hair', 17),
    ('lost', 17),
    ('U104', 17),
    ('fine', 16),
    ('real', 16),
    ('game', 16),
    ('fuck', 15),
    ('sits', 15),
    ('eyes', 15),
    ('lets', 15),
    ('heya', 15),
    ('kill', 15),
    ('read', 14),
    ('shut', 14),
    ('wait', 14),
    ('goes', 14),
    ('keep', 14),
    ('true', 14),
    ('pick', 13),
    ('free', 13),
    ('else', 13),
    ('near', 13),
    ('nope', 13),
    ('U168', 13),
 ('hope', 12),
 ('head', 12),
 ('male', 12),
 ('than', 12),
 ('gets', 12),
 ('cold', 12),
 ('hehe', 12),
 ('bout', 12),
 ('stay', 12),
 ('used', 12),
 ('awww', 12),
 ('told', 12),
 ('This', 12),
 ('U102', 12),
 ('doin', 11),
 ('kids', 11),
 ('perv', 11),
 ('wont', 11),
 ('face', 11),
 ('home', 11),
 ('year', 11),
 ('babe', 11),
 ('into', 11),
 ('yall', 11),
 ('.. .', 11),
 ('U119', 11),
 ('U107', 11),
 ('hard', 10),
 ('show', 10),
 ('U101', 10),
 ('once', 10),
 ('Well', 10),
 ('help', 10),
 ('mind', 10),
 ('Yeah', 10),
 ('week', 10),
 ('Liam', 10),
 ('U132', 10),
 ('pics', 9),
 ('such', 9),
 ('type', 9),
 ('best', 9),
 ('neck', 9),
 ('dang', 9),
 ('dead', 9),
 ('runs', 9),
 ('aint', 9),
 ('rock', 9),
 ('days', 9),
 ('mine', 9),
 ('book', 9),
 ('crap', 9),
 ('soon', 9),
 ('care', 9),
 ('full', 9),
 ('kiss', 9),
 ('hour', 9),
 ('nick', 9),
 ('sick', 9),
 ('; ..', 9),
 ('hmmm', 9),
 ('U139', 8),
 ('word', 8),
 ('heyy', 8),
 ('case', 8),
 ('wana', 8),
 ('hows', 8),
 ('went', 8),
 ('lady', 8),
 ('blue', 8),
 ('says', 8),
 ('suck', 8),
 ('made', 8),
 ('wife', 8),
 ('sang', 8),
 ('U144', 8),
 ('fast', 7),
 ('rule', 7),
 ('dude', 7),
 ('okay', 7),
 ('alot', 7),
 ('hand', 7),
 ('took', 7),
 ('wear', 7),
 ('Hiya', 7),
 ('kick', 7),
 ('ahhh', 7),
 ('dear', 7),
 ('That', 7),
 ('U108', 7),
 ('U169', 7),
 ('U129', 6),
 ('U116', 6),
 ('most', 6),
 ('thru', 6),
 ('U165', 6),
 ('list', 6),
 ('seem', 6),
 ('sing', 6),
 ('next', 6),
 ('done', 6),
 ('ride', 6),
 ('comp', 6),
 ('main', 6),
 ('))))', 6),
 ('goin', 6),
 ('U520', 6),
 ('pink', 6),
 ('poor', 6),
 ('gone', 6),
 ('oops', 6),
 ('knew', 6),
 ('<---', 6),
 ('ball', 6),
 ('send', 6),
 ('Song', 6),
 ('blah', 6),
 ('They', 6),
 ('part', 6),
 ('U103', 6),
 ('U120', 6),
 ('Last', 6),
 ('whos', 6),
 ('food', 6),
 ('U142', 6),
 ('sock', 6),
 ('U197', 6),
 ('legs', 5),
 ('fire', 5),
 ('warm', 5),
 ('late', 5),
 ('hang', 5),
 ('miss', 5),
 ('boys', 5),
 ('land', 5),
 ('nose', 5),
 ('lick', 5),
 ('caps', 5),
 ('wish', 5),
 ('U128', 5),
 ('came', 5),
 ('cali', 5),
 ('roll', 5),
 ('easy', 5),
 ('lose', 5),
 ('When', 5),
 ('soul', 5),
 ('luck', 5),
 ('also', 5),
 ('kool', 5),
 ('fall', 5),
 ('boss', 5),
 ('beer', 5),
 ('ohhh', 5),
 ('####', 5),
 ('wall', 5),
 ('Have', 5),
 ('meet', 5),
 ('till', 5),
 ('feet', 5),
 ('xbox', 5),
 ('idea', 5),
 ('heck', 5),
 ('joke', 5),
 ('fool', 5),
 ('felt', 5),
 ('yoko', 5),
 ('meds', 5),
 ('both', 5),
 ('Lime', 5),
 ('glad', 4),
 ('U133', 4),
 ('U126', 4),
 ('jerk', 4),
 ('ugly', 4),
 ('date', 4),
 ('ummm', 4),
 ('quit', 4),
 ('rest', 4),
 ('door', 4),
 ('none', 4),
 ('self', 4),
 ('pass', 4),
 ('line', 4),
 ('cute', 4),
 ('holy', 4),
 ('hook', 4),
 ('Like', 4),
 ('each', 4),
 ('open', 4),
 ('high', 4),
 ('ouch', 4),
 ('evil', 4),
 ('fart', 4),
 ('grrr', 4),
 ('pain', 4),
 ('pfft', 4),
 ('sigh', 4),
 ('shes', 4),
 ('ROOM', 4),
 (',,,,', 4),
 ('lord', 4),
 ('mmmm', 4),
 ('ones', 4),
 ('huge', 4),
 ('woot', 4),
 ('shot', 4),
 ('team', 4),
 ('ways', 4),
 ('beat', 4),
 ('kent', 4),
 ('U130', 4),
 ('U196', 4),
 ('U219', 4),
 ('turn', 4),
 ('lame', 4),
 ('U123', 4),
 ('U154', 4),
 ('U988', 4),
 ('puff', 4),
 ('U146', 4),
 ('U989', 4),
 ('U117', 4),
 ('U819', 4),
 ('U820', 4),
 ('clap', 3),
 ('itch', 3),
 ('guyz', 3),
 ('U136', 3),
 ('gold', 3),
 ('ring', 3),
 ('isnt', 3),
 ('U141', 3),
 ('Only', 3),
 ('U148', 3),
 ('Your', 3),
 ('deal', 3),
 ('wash', 3),
 ('U109', 3),
 ('piff', 3),
 ('jump', 3),
 ('band', 3),
 ('orgy', 3),
 ('slap', 3),
 ('soft', 3),
 ('bend', 3),
 ('toss', 3),
 ('amen', 3),
 ('rain', 3),
 ('deop', 3),
 ('roof', 3),
 ('((((', 3),
 ('CHAT', 3),
 ('ahem', 3),
 ('hola', 3),
 ('butt', 3),
 ('imma', 3),
 ('town', 3),
 ('hawt', 3),
 ('2006', 3),
 ('Elev', 3),
 ('Wind', 3),
 ('AKDT', 3),
 ('lead', 3),
 ('DING', 3),
 ('note', 3),
 ('gawd', 3),
 ('half', 3),
 ('mary', 3),
 ('ello', 3),
 ('hick', 3),
 ('wine', 3),
 ('hiii', 3),
 ('bare', 3),
 ('vote', 3),
 ('Same', 3),
 ('wack', 3),
 ('snow', 3),
 ('hurt', 3),
 ('move', 3),
 ('road', 3),
 ('walk', 3),
 ('yawn', 3),
 ('hail', 3),
 ('nana', 3),
 ('U106', 3),
 ('hump', 3),
 ('elle', 3),
 ('yada', 3),
 ('tune', 3),
 ('hank', 3),
 ('slow', 3),
 ('rubs', 3),
 ('skin', 3),
 ('died', 3),
 ('U145', 3),
 ('swim', 3),
 ('U163', 3),
 ('army', 3),
 ('THAT', 3),
 ('wazz', 3),
 ('toes', 3),
 ('U153', 3),
 ('golf', 2),
 ('drew', 2),
 ('cast', 2),
 ('Days', 2),
 ('opps', 2),
 ('U138', 2),
 ('plan', 2),
 ('Just', 2),
 ('deaf', 2),
 ('deep', 2),
 ('phil', 2),
 ('hmph', 2),
 ('U155', 2),
 ('Poor', 2),
 ('Lies', 2),
 ('bite', 2),
 ('mins', 2),
 ('eats', 2),
 ('>:->', 2),
 ('cell', 2),
 ('cmon', 2),
 ('wats', 2),
 ('kind', 2),
 ('mike', 2),
 ('whoa', 2),
 ('dumb', 2),
 ('park', 2),
 ('Sure', 2),
 ('Come', 2),
 ('O.k.', 2),
 ('mama', 2),
 ('Nice', 2),
 ('hold', 2),
 ('ohio', 2),
 ('whip', 2),
 ('twin', 2),
 ('burp', 2),
 ('blew', 2),
 ('temp', 2),
 ('corn', 2),
 ('pool', 2),
 ('cash', 2),
 ('ears', 2),
 ('From', 2),
 ('porn', 2),
 ('heal', 2),
 ('Dang', 2),
 ('ciao', 2),
 ('DOES', 2),
 ('typo', 2),
 ('Stop', 2),
 ('eric', 2),
 ('Drew', 2),
 ('sore', 2),
 ('Live', 2),
 ('High', 2),
 ('hits', 2),
 ('KoOL', 2),
 ('past', 2),
 ('Love', 2),
 ('meat', 2),
 ('!!!.', 2),
 ('argh', 2),
 ('limp', 2),
 ('rent', 2),
 ('cars', 2),
 ('Tell', 2),
 ('shop', 2),
 ('U172', 2),
 ('five', 2),
 ('sell', 2),
 ('<<<<', 2),
 ('city', 2),
 ('yard', 2),
 ('grrl', 2),
 ('chip', 2),
 ('bear', 2),
 ('foot', 2),
 ('uses', 2),
 ('DONT', 2),
 ('sort', 2),
 ('lies', 2),
 ('whud', 2),
 ('hott', 2),
 ('Down', 2),
 ('Lets', 2),
 ('club', 2),
 ('adds', 2),
 ('Here', 2),
 ('born', 2),
 ('wOOt', 2),
 ('area', 2),
 ('?!?!', 2),
 ('Ohio', 2),
 ('U112', 2),
 ('humm', 2),
 ('newp', 2),
 ('gays', 2),
 ('zone', 2),
 ('hint', 2),
 ('spin', 2),
 ('ewww', 2),
 ('pies', 2),
 ('doll', 2),
 ('drop', 2),
 ('gimp', 2),
 ('spot', 2),
 ('ages', 2),
 ('clue', 2),
 ('mass', 2),
 ('Ummm', 2),
 ('Gosh', 2),
 ('flow', 2),
 ('kewl', 2),
 ('hall', 2),
 ('haze', 2),
 ('1996', 2),
 ('John', 2),
 ('john', 2),
 ('sooo', 2),
 ('cost', 2),
 ('trip', 2),
 ('babi', 2),
 ('rich', 2),
 ('U100', 2),
 ('n9ne', 2),
 ('Ahhh', 2),
 ('??!!', 2),
 ('U111', 2),
 ('moon', 2),
 ('STOP', 2),
 ('any1', 2),
 ('yeas', 2),
 ('wooo', 2),
 ('<333', 2),
 ('tick', 2),
 ('tock', 2),
 ('WITH', 2),
 ('FROM', 2),
 ('side', 2),
 ('Heyy', 2),
 ('howz', 2),
 ("ex's", 2),
 ('Cool', 2),
 ('U170', 2),
 ('U175', 2),
 ('root', 2),
 ('tyvm', 2),
 ('luvs', 2),
 ('fits', 2),
 ('rofl', 2),
 ('sand', 2),
 ('ltns', 2),
 ('flaw', 2),
 ('aunt', 2),
 ('lawl', 2),
 ('Okay', 2),
 ('HAVE', 2),
 ('NONE', 2),
    ('YOUR', 2),
    ('Lmao', 2),
    ('Tisk', 2),
    ('U190', 2),
    ('tisk', 2),
    ('draw', 1),
    ('docs', 1),
    ('Slip', 1),
    ('Fade', 1),
    ('bowl', 1),
    ('bong', 1),
    ('ogan', 1),
    ('cams', 1),
    ('gooo', 1),
    ('yeee', 1),
    ('ahah', 1),
    ('jeep', 1),
    ('Deep', 1),
    ('Show', 1),
    ('Turn', 1),
    ('Hand', 1),
    ('VBox', 1),
    ('ELSE', 1),
    ('serg', 1),
    ('bein', 1),
    ('whys', 1),
    ('tape', 1),
    ('sexs', 1),
    ('form', 1),
    ('HUGE', 1),
    ('nads', 1),
    ('owww', 1),
    ('gags', 1),
    ('Meep', 1),
    ('LAst', 1),
    ("pm's", 1),
    ('1.99', 1),
    ('lool', 1),
    ('kina', 1),
    ('sext', 1),
    ('lazy', 1),
    ('calm', 1),
    ('arms', 1),
    ('smax', 1),
    ('VVil', 1),
    ('este', 1),
    ('chik', 1),
    ('Boyz', 1),
    ('coat', 1),
    ('Eyes', 1),
    ('Dawn', 1),
    ('LIVE', 1),
    ('mauh', 1),
    ('ques', 1),
    ('4.20', 1),
    ('gosh', 1),
    ('ruff', 1),
    ('mame', 1),
    ('nada', 1),
    ('push', 1),
    ('prob', 1),
    ('wild', 1),
    ('whew', 1),
    ('dark', 1),
    ('waht', 1),
    ('test', 1),
    ('boot', 1),
    ('hiom', 1),
    ('HAHA', 1),
    ('dman', 1),
    ('jail', 1),
    ('cops', 1),
    ('hogs', 1),
    ('peek', 1),
    ('MORE', 1),
    ('TIME', 1),
    ('loud', 1),
 ('o.k.', 1),
 ('Sexy', 1),
 ('Ctrl', 1),
 ('hots', 1),
 ('Need', 1),
 ('frst', 1),
 ('1200', 1),
 ('crop', 1),
 ('bomb', 1),
 ('Pour', 1),
 ('pour', 1),
 ('Swim', 1),
 ('Hard', 1),
 ('eeek', 1),
 ('tjhe', 1),
 ('10th', 1),
 ('heee', 1),
 ('peel', 1),
 ('fock', 1),
 ('Kold', 1),
 ('exit', 1),
 ('kold', 1),
 ('3:45', 1),
 ('MRIs', 1),
 ('buff', 1),
 ('plus', 1),
 ('tory', 1),
 ('knee', 1),
 ('OOPS', 1),
 ('oooh', 1),
 ('lala', 1),
 ('fake', 1),
 ('ssid', 1),
 ('poot', 1),
 ('poop', 1),
 ('bird', 1),
 ('plow', 1),
 ('thnx', 1),
 ('card', 1),
 ('Hugs', 1),
 ('Lord', 1),
 ('uyes', 1),
 ('benz', 1),
 ('<~~~', 1),
 ('disc', 1),
 ('LONG', 1),
 ('Been', 1),
 ('Will', 1),
 ('bloe', 1),
 ('blow', 1),
 ('hooo', 1),
 ('thje', 1),
 ('Jess', 1),
 ('term', 1),
 ('Tina', 1),
 ('ooer', 1),
 ('HALO', 1),
 ('Awww', 1),
 ('anal', 1),
 ('Drop', 1),
 ('dojn', 1),
 ('wubs', 1),
 ('mkay', 1),
 ('spat', 1),
 ('gees', 1),
 ('hawT', 1),
 ('yes.', 1),
 ('puts', 1),
 ('fish', 1),
 ('size', 1),
 ('39.3', 1),
 ('1980', 1),
 ('64.8', 1),
 ('syck', 1),
 ('tere', 1),
 ('U542', 1),
 ('sent', 1),
```

```
('some', 1),
('45.5', 1),
('98.5', 1),
('1299', 1),
('1900', 1),
('1930', 1),
('Werd', 1),
('Rofl', 1),
('mode', 1),
('nawt', 1),
('sign', 1),
('woof', 1),
('sum1', 1),
('ghet', 1),
('brad', 1),
('offa', 1),
('Dood', 1),
('out.', 1),
('LOUD', 1),
('sink', 1),
('FINE', 1),
('cums', 1),
('loss', 1),
('Life', 1),
('Damn', 1),
('wrap', 1),
('hide', 1),
("PM's", 1),
('Talk', 1),
('okey', 1),
('worl', 1),
('Hold', 1),
('cepn', 1),
('lots', 1),
('Mary', 1),
('nawp', 1),
('addy', 1),
('lake', 1),
('slip', 1),
('mite', 1),
('wood', 1),
('orta', 1),
('wins', 1),
('ebay', 1),
('coem', 1),
('giva', 1),
('1.98', 1),
('ally', 1),
('Judy', 1),
('cyas', 1),
('shup', 1),
('tooo', 1),
("pm'n", 1),
('choc', 1),
('wher', 1),
('whoo', 1),
('dint', 1),
('tend', 1),
('menu', 1),
('lust', 1),
('nods', 1),
('NAME', 1),
('kept', 1),
('scuk', 1),
('raed', 1),
('Then', 1),
('bugs', 1),
('nerd', 1),
('Hill', 1),
('Evil', 1),
('saME', 1),
('2Pac', 1),
('Time', 1),
('pimp', 1),
('haaa', 1),
('98.6', 1),
("it's", 1),
('Mono', 1),
('none', 1),
('mono', 1),
('Bone', 1),
('Hero', 1),
('Came', 1),
('.op.', 1),
('Hott', 1),
('Joey', 1),
('Jane', 1),
('span', 1),
('wore', 1),
('QUIT', 1),
('pasa', 1),
('barn', 1),
('Kick', 1),
('feat', 1),
('Back', 1),
('dork', 1),
('laid', 1),
('Home', 1),
('herd', 1),
('Born', 1),
('Away', 1),
('Tide', 1),
('jush', 1),
('Cute', 1),
('GrlZ', 1),
('lung', 1),
('SOME', 1),
('Lion', 1),
('brat', 1),
(':o *', 1),
('MUAH', 1),
('fawk', 1),
('dust', 1),
('Help', 1),
('seth', 1),
('Heya', 1),
('bone', 1),
('abou', 1),
('tthe', 1),
('Even', 1),
('herE', 1),
('Hail', 1),
('halo', 1),
('pork', 1),
('1cos', 1),
("yw's", 1),
('mark', 1),
('dotn', 1),
('PMSL', 1),
('pmsl', 1),
('gift', 1),
('outs', 1),
('Paul', 1),
('outa', 1),
('York', 1),
('Care', 1),
('Chat', 1),
('fear', 1),
('dies', 1),
('givs', 1),
('bust', 1),
('xmas', 1),
('enuf', 1),
('LoVe', 1),
('eeww', 1),
('dick', 1),
('fair', 1),
('lyin', 1),
('lois', 1),
('cuss', 1),
('LATE', 1),
('THEY', 1),
('GOOD', 1),
('rape', 1),
('geez', 1),
('tart', 1),
('tart', 1),
('hgey', 1),
('caan', 1),
('lol.', 1),
('Elle', 1),
('nude', 1),
('allo', 1),
('yesh', 1),
('wind', 1),
('Reub', 1),
('!???', 1),
('heat', 1),
('kmph', 1),
('pope', 1),
('yess', 1),
('!...', 1),
('duet', 1),
('wuts', 1),
('west', 1),
('quiz', 1),
('scar', 1),
('Girl', 1),
('pair', 1),
('Rang', 1),
('rang', 1),
('bell', 1),
('dawg', 1),
('febe', 1),
('Prof', 1),
('Kewl', 1),
('jude', 1),
('Yoko', 1),
('seee', 1),
('whou', 1),
('idnt', 1),
('perk', 1),
('http', 1),
('2DAY', 1),
('yell', 1),
('mang', 1),
('SSRI', 1),
('cure', 1),
('wean', 1),
('post', 1),
('anti', 1),
('noth', 1),
('tall', 1),
('pray', 1),
('weed', 1),
('icky', 1),
('Rick', 1),
('spit', 1),
('lube', 1),
('mami', 1),
('east', 1),
('18ST', 1),
('seat', 1),
('cock', 1),
('SExy', 1),
('otay', 1),
('firs', 1),
('site', 1),
('U113', 1),
('dump', 1),
('toop', 1),
('four', 1),
('U118', 1),
('sets', 1),
('asss', 1),
('paid', 1),
('Iowa', 1),
('Teck', 1),
('"...', 1),
('jeff', 1),
('crib', 1),
('drug', 1),
('cook', 1),
('9:10', 1),
 ('9.10', 1),
 ('ladz', 1),
 ('aime', 1),
 ('hong', 1),
 ('kong', 1),
 ('Oops', 1),
 ('tits', 1),
 ('gret', 1),
 ('guns', 1),
 ('inch', 1),
 ('sean', 1),
 ('howl', 1),
 ('Take', 1),
 ('z-ro', 1),
 ('U137', 1),
 ('Haha', 1),
 ('1985', 1),
 ('slam', 1),
 ('pine', 1),
 ('puke', 1),
 ('waaa', 1),
 ('urls', 1),
 ('star', 1),
 ('Save', 1),
 ('teck', 1),
 ('Room', 1),
 ('sori', 1),
 ('Long', 1),
 ('poem', 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 [24]:

```
sum([len(w) for w in text1]) / len(text1)
```

Out[24]:

```
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 [16]:

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

```
0.003834076505162584%
```

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 [17]:

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

Out[17]:

```
['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-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 [18]:

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

Out[18]:

```
98171
```

In [19]:

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

Out[19]:

```
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 [27]:

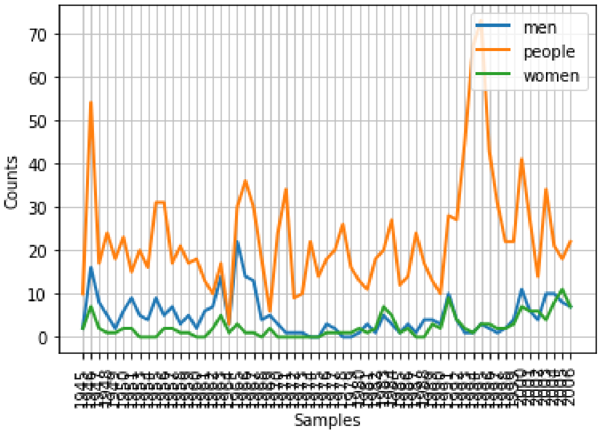
```
nltk.corpus.state_union.fileids()
```

Out[27]:

```
['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 [35]:

```
cfd = nltk.ConditionalFreqDist(
    (target, fileid[:4])
    for fileid in nltk.corpus.state_union.fileids()
    for w in nltk.corpus.state_union.words(fileid)
    for target in ['men', 'women','people']
    if w.lower().startswith(target))
cfd.plot()
```



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 [38]:

```
from nltk.corpus import brown
brown.categories()
```

Out[38]:

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

In [39]:

```
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','learned','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
```

Will es una palabra que tiene gran preponderancia en textos news, hobbies, goverment y lore: cumple la lógica de que esta palabra se relaciona directamente para estructurar promesas, intenciones y acciones espontaneas, lo que se confirma con la espontaneidad aspectos como hobbies y las promesas e intenciones en temas de goverment, lore y news (que en muchos casos esta asociado a goverment). En segunda instancia la palabra mas usada es can (casi al mismo nivel de could) en donde se toma para can como referencia el uso en learned y belle_lettres, asociandose directamente con el caracter estricto de learned y el caracter subjetivo de belle:lettres.

In [40]:

```
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','learned','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   10   10
           lore   15   19   19
        mystery   21    7    8
        reviews    3    7    0
```

Las palabras dead, love y pain tienen una relacion directa entre dead en el ambito fiction, debido a que en ella se concentra gran parte del género. En contextos ficticios podemos asumir que la palabra muerte siempre sera el eje principalal igual que Belles Lettres. Por sentido compun, la palabra amor y romance están totalmente vinculados.

Tambien podríamos decir que al existir una relacion de la palabra love en el ambito de belles lettres, parte de un desarrollo del romaticismo e idealismo. Finalmente podemos observar que en el ambito adventure, la palabra dead tiene una relación por el extremismo y el peligro de estos contextos.