Taller 9

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 13-nov-2020 11:59 PM

[Nicolás Garcés R]

[nicolas.garces@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/), ejercicios:

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

In [2]:

```
import nltk
nltk.download('gutenberg')
nltk.download('genesis')
nltk.download('inaugural')
nltk.download('nps_chat')
nltk.download('webtext')
nltk.download('treebank')

from nltk.book import *
```

```
[nltk_data] Downloading package gutenberg to
[nltk_data]     /Users/acidrain/nltk_data...
[nltk_data]   Package gutenberg is already up-to-date!
[nltk_data] Downloading package genesis to
[nltk_data]     /Users/acidrain/nltk_data...
[nltk_data]   Package genesis is already up-to-date!
[nltk_data] Downloading package inaugural to
[nltk_data]     /Users/acidrain/nltk_data...
[nltk_data]   Package inaugural is already up-to-date!
[nltk_data] Downloading package nps_chat to
[nltk_data]     /Users/acidrain/nltk_data...
[nltk_data]   Package nps_chat is already up-to-date!
[nltk_data] Downloading package webtext to
[nltk_data]     /Users/acidrain/nltk_data...
[nltk_data]   Package webtext is already up-to-date!
[nltk_data] Downloading package treebank to
[nltk_data]     /Users/acidrain/nltk_data...
[nltk_data]   Package treebank is already up-to-date!
```

```
[nltk_data]   Package treebank is already up-to-date!
```

```
*** 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
```

Capítulo 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 [54]:

```
four=[w for w in text5 if len(w)== 4]
cuatro=[w for w in four if w.isalpha()]
```

In [4]:

```
c =  FreqDist(cuatro)
c.most_common()
```

Out[4]:

```
[('JOIN', 1021),
 ('PART', 1016),
 ('that', 274),
 ('what', 183),
 ('here', 181),
 ('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),
 ('MODE', 41),
 ('will', 40),
 ('much', 40),
 ('then', 40),
 ('over', 39),
 ('work', 38),
 ('were', 38),
 ('take', 37),
 ('song', 36),
 ('even', 35),
 ('does', 35),
 ('seen', 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),
 ('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),
 ('hugs', 18),
 ('What', 18),
 ('find', 18),
 ('cant', 18),
 ('left', 17),
 ('shit', 17),
 ('nite', 17),
 ('busy', 17),
 ('hair', 17),
 ('lost', 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),
 ('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),
 ('doin', 11),
 ('kids', 11),
 ('perv', 11),
 ('wont', 11),
 ('face', 11),
 ('home', 11),
 ('year', 11),
 ('babe', 11),
 ('into', 11),
 ('yall', 11),
 ('hard', 10),
 ('show', 10),
 ('once', 10),
 ('Well', 10),
 ('help', 10),
 ('mind', 10),
 ('Yeah', 10),
 ('week', 10),
 ('Liam', 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),
 ('hmmm', 9),
 ('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),
 ('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),
 ('most', 6),
 ('thru', 6),
 ('list', 6),
 ('seem', 6),
 ('sing', 6),
 ('next', 6),
 ('done', 6),
 ('ride', 6),
 ('comp', 6),
 ('main', 6),
 ('goin', 6),
 ('pink', 6),
 ('poor', 6),
 ('gone', 6),
 ('oops', 6),
 ('knew', 6),
 ('ball', 6),
 ('send', 6),
 ('Song', 6),
 ('blah', 6),
 ('They', 6),
 ('part', 6),
 ('Last', 6),
 ('whos', 6),
 ('food', 6),
 ('sock', 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),
 ('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),
 ('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),
 ('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),
 ('lord', 4),
 ('mmmm', 4),
 ('ones', 4),
 ('huge', 4),
 ('woot', 4),
 ('shot', 4),
 ('team', 4),
 ('ways', 4),
 ('beat', 4),
 ('kent', 4),
 ('turn', 4),
 ('lame', 4),
 ('puff', 4),
 ('clap', 3),
 ('itch', 3),
 ('guyz', 3),
 ('gold', 3),
 ('ring', 3),
 ('isnt', 3),
 ('Only', 3),
 ('Your', 3),
 ('deal', 3),
 ('wash', 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),
 ('CHAT', 3),
 ('ahem', 3),
 ('hola', 3),
 ('butt', 3),
 ('imma', 3),
 ('town', 3),
 ('hawt', 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),
    ('hump', 3),
    ('elle', 3),
    ('yada', 3),
    ('tune', 3),
    ('hank', 3),
    ('slow', 3),
    ('rubs', 3),
    ('skin', 3),
    ('died', 3),
    ('swim', 3),
    ('army', 3),
    ('THAT', 3),
    ('wazz', 3),
    ('toes', 3),
    ('golf', 2),
    ('drew', 2),
    ('cast', 2),
    ('Days', 2),
    ('opps', 2),
    ('plan', 2),
    ('Just', 2),
    ('deaf', 2),
    ('deep', 2),
    ('phil', 2),
    ('hmph', 2),
    ('Poor', 2),
    ('Lies', 2),
    ('bite', 2),
    ('mins', 2),
    ('eats', 2),
    ('cell', 2),
    ('cmon', 2),
    ('wats', 2),
    ('kind', 2),
    ('mike', 2),
    ('whoa', 2),
    ('dumb', 2),
    ('park', 2),
    ('Sure', 2),
    ('Come', 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),
 ('argh', 2),
 ('limp', 2),
 ('rent', 2),
 ('cars', 2),
 ('Tell', 2),
 ('shop', 2),
 ('five', 2),
 ('sell', 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),
 ('Ohio', 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),
 ('John', 2),
 ('john', 2),
 ('sooo', 2),
 ('cost', 2),
 ('trip', 2),
 ('babi', 2),
 ('rich', 2),
 ('Ahhh', 2),
 ('moon', 2),
 ('STOP', 2),
 ('yeas', 2),
 ('wooo', 2),
 ('tick', 2),
 ('tock', 2),
 ('WITH', 2),
 ('FROM', 2),
 ('side', 2),
 ('Heyy', 2),
 ('howz', 2),
 ('Cool', 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),
 ('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),
 ('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),
 ('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),
 ('hugs', 1),
 ('peek', 1),
 ('MORE', 1),
 ('TIME', 1),
 ('loud', 1),
 ('Sexy', 1),
 ('Ctrl', 1),
 ('hots', 1),
 ('Need', 1),
 ('frst', 1),
 ('crop', 1),
 ('bomb', 1),
 ('Pour', 1),
 ('pour', 1),
 ('Swim', 1),
 ('Hard', 1),
 ('eeek', 1),
 ('tjhe', 1),
 ('heee', 1),
 ('peel', 1),
 ('fock', 1),
 ('Kold', 1),
 ('exit', 1),
 ('kold', 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),
 ('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),
 ('puts', 1),
 ('fish', 1),
 ('size', 1),
 ('syck', 1),
 ('tere', 1),
 ('sent', 1),
 ('Werd', 1),
 ('Rofl', 1),
 ('mode', 1),
 ('nawt', 1),
 ('sign', 1),
 ('woof', 1),
 ('wool', 1),
 ('ghet', 1),
 ('brad', 1),
 ('offa', 1),
 ('Dood', 1),
 ('LOUD', 1),
 ('sink', 1),
 ('FINE', 1),
 ('cums', 1),
 ('loss', 1),
 ('Life', 1),
 ('Damn', 1),
 ('wrap', 1),
 ('hide', 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),
 ('ally', 1),
 ('Judy', 1),
 ('cyas', 1),
 ('shup', 1),
 ('tooo', 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),
 ('Time', 1),
 ('pimp', 1),
 ('haaa', 1),
 ('Mono', 1),
 ('mono', 1),
 ('Bone', 1),
 ('Hero', 1),
 ('Came', 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)
 ('held', 1),
 ('Born', 1),
 ('Away', 1),
 ('Tide', 1),
 ('jush', 1),
 ('Cute', 1),
 ('GrlZ', 1),
 ('lung', 1),
 ('SOME', 1),
 ('Lion', 1),
 ('brat', 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),
 ('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),
 ('hgey', 1),
 ('caan', 1),
 ('Elle', 1),
 ('nude', 1),
 ('allo', 1),
 ('yesh', 1),
 ('wind', 1),
 ('Reub', 1),
 ('heat', 1),
 ('kmph', 1),
 ('pope', 1),
 ('yess', 1),
 ('duet', 1),
 ('wuts', 1),
 ('west', 1),
 ('quiz', 1),
 ('scar', 1),
 ('Girl', 1),
 ('pair', 1),
 ('Rang', 1),
 ('rang', 1),
 ('bell', 1),
 ('dawg', 1)
 ('dawg', 1),
 ('febe', 1),
 ('Prof', 1),
 ('Kewl', 1),
 ('jude', 1),
 ('Yoko', 1),
 ('seee', 1),
 ('whou', 1),
 ('idnt', 1),
 ('perk', 1),
 ('http', 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),
 ('seat', 1),
 ('cock', 1),
 ('SExy', 1),
 ('otay', 1),
 ('firs', 1),
 ('site', 1),
 ('dump', 1),
 ('toop', 1),
 ('four', 1),
 ('sets', 1),
 ('asss', 1),
 ('paid', 1),
 ('Iowa', 1),
 ('Teck', 1),
 ('jeff', 1),
 ('crib', 1),
 ('drug', 1),
 ('cook', 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),
 ('Haha', 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),
 ('jack', 1),
 ('Rule', 1),
 ('CAPS', 1),
 ('junk', 1),
 ('tips', 1),
 ('rush', 1),
 ('Nooo', 1),
 ('Nooo', 1),
 ('Troy', 1),
 ('tail', 1),
 ('Seee', 1),
 ('dyed', 1),
 ('beam', 1),
 ('daft', 1),
 ('twit', 1),
 ('scum', 1),
 ('Type', 1),
 ('WHOA', 1),
 ('toke', 1),
 ('ribs', 1),
 ('Eggs', 1),
 ('Wyte', 1),
 ('moms', 1),
 ('Over', 1),
 ('West', 1),
 ('Rock', 1),
 ('goof', 1),
 ('able', 1),
 ('vamp', 1),
 ('Nope', 1),
 ('Kent', 1),
 ('ther', 1),
 ('TEXT', 1),
 ('SIZE', 1),
 ('gear', 1),
 ('CALI', 1),
 ('Matt', 1),
 ('Rush', 1),
 ('AWAY', 1),
 ('NTMN', 1),
 ('Kiss', 1),
 ('grea', 1),
 ('Look', 1),
 ('guts', 1),
 ('wrek', 1),
 ('Fort', 1),
 ('AKST', 1),
 ('wire', 1),
 ('soda', 1),
 ('gray', 1),
 ('tlak', 1),
 ('ltnc', 1),
 ('sayn', 1),
 ('evah', 1),
 ('bike', 1),
 ('hill', 1),
 ('ohwa', 1),
 ('caca', 1),
 ('prep', 1),
 ('pull', 1),
 ('dirt', 1),
 ('vent', 1),
 ('safe', 1),
 ('dogs', 1),
 ('bull', 1),
 ('asks', 1),
 ('Road', 1),
 ('chit', 1),
 ('grin', 1),
 ('bred', 1),
 ('rats', 1),
 ('samn', 1),
 ('Phil', 1),
 ('nuff', 1),
 ('rose', 1),
 ('Ruth', 1),
 ('grew', 1),
 ('mena', 1),
 ('ROFL', 1),
 ('lapd', 1),
 ('surf', 1),
 ('City', 1),
 ('hazy', 1),
 ('thot', 1),
 ('acid', 1),
('acid', 1),
('wide', 1),
('keys', 1),
('salt', 1),
('mess', 1),
('base', 1),
('byes', 1),
('yout', 1),
('numb', 1),
('thah', 1),
('mahn', 1),
('King', 1),
('TALK', 1),
('GIRL', 1),
('WHEN', 1),
('HOTT', 1),
('HERE', 1),
('soup', 1),
('Mine', 1),
('vega', 1),
('pigs', 1),
('king', 1),
('poof', 1),
('Nova', 1),
('mofo', 1),
('Ohhh', 1),
('Holy', 1),
('sips', 1),
('clay', 1),
('None', 1),
('Male', 1),
('bacl', 1),
('body', 1),
('akon', 1),
('yoll', 1),
('boom', 1),
('News', 1),
('Maps', 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 [26]:

```
sum(len(w) for w in text1)
```

Out[26]:

```
999044
```

Este codigo suma la cantidad de characters en text 1

In [27]:

```
#average word length
sum(len(w) for w in text1)/len(text1)
```

Out[27]:

```
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 [7]:

```
def percent(word, text):
    f= FreqDist(text)
    per = f[word]
    per= per/len(text)
    per= per*100
    p= str(per)+'%'
    p= str(per)+'%'
    return p
```

In [8]:

```
percent("of",text1)
```

Out[8]:

```
'2.505952403774265%'
```

In [9]:

```
percent('kids', text5)
```

Out[9]:

```
'0.024439013552543878%'
```

Capitulo 2

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

In [11]:

```
import nltk
nltk.corpus.gutenberg.fileids()
['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']
austen = nltk.corpus.gutenberg.words('austen-persuasion.txt')
```

In [12]:

```
len(austen)
```

Out[12]:

```
98171
```

austen-persuasion tiene 98171 palabras.

In [14]:

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

Out[14]:

```
6132
```

austen-persuasion tiene 98171 tipos de palabras.

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 [18]:

```
import nltk
from nltk.corpus import state_union
word=state_union.words()
al=state_union.fileids()
mn=[]
for i in state_union.fileids():
    x= state_union.words(i)
    f=FreqDist(x)
    mn.append(f['men'])
    print(f['men'])
```

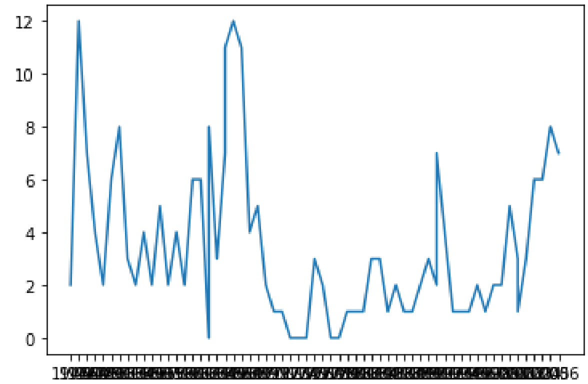
```
2
12
7
4
2
6
8
3
2
4
2
5
2
4
2
6
6
0
8
3
7
11
12
11
4
5
2
1
1
0
0
0
3
2
0
0
1
1
1
3
3
1
2
1
1
2
3
2
7
4
1
1
1
2
1
2
2
5
3
1
3
6
6
8
7
```

In [34]:

```
import matplotlib.pyplot as plt
years=[]
```

```
for i in a1:
    a=str(i)
    years.append(a[0:4])

plt.plot(years,mn);
```



El uso de la palabra men ha caido en los discursos a lo largo del tiempo.

In [35]:

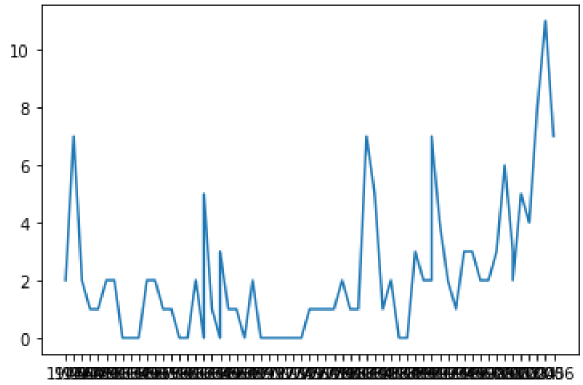
```
wm=[]
for i in state_union.fileids():
    x= state_union.words(i)
    f=FreqDist(x)
    wm.append(f['women'])
    print(f['women'])
```

```
2
7
2
1
1
2
2
0
0
0
2
2
1
1
0
0
2
0
5
1
0
3
1
1
0
2
0
0
0
0
0
0
1
1
1
1
2
1
1
7
5
1
```

```
2
0
0
3
2
2
7
4
2
1
3
3
2
2
3
6
3
2
5
4
8
11
7
```

In [37]:

```
plt.plot(years,wm);
```



El uso de la palabra women ha tenido una tendencia positiva en los ultimos discursos.

In [39]:

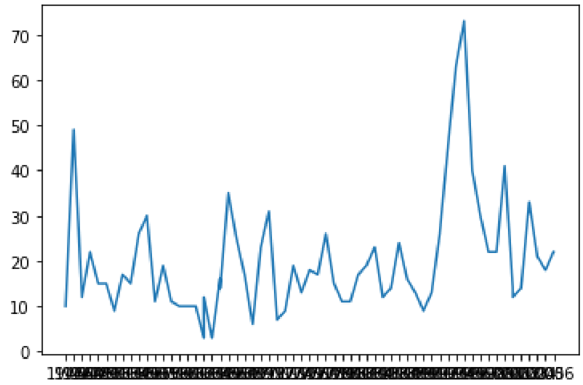
```
pl=[]
for i in state_union.fileids():
    x= state_union.words(i)
    f=FreqDist(x)
    pl.append(f['people'])
    print(f['people'])
```

```
10
49
12
22
15
15
9
17
15
26
30
11
19
11
10
10
10
3
```

```
12
3
16
14
35
25
17
6
23
31
7
9
19
13
18
17
26
15
11
11
17
19
23
12
14
24
16
13
9
13
13
26
45
63
73
40
30
22
22
41
14
12
14
33
21
18
22
```

In [40]:

```
plt.plot(years,pl);
```



El uso de la palabra people se ha mantenido estable en estos discursos

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 [53]:

```
import nltk
from nltk.corpus import brown
nltk.download('brown')
one= nltk.ConditionalFreqDist((genre,word)for genre in brown.categories() for word in brown.words(c
ategories=genre))
cat= ['news', 'religion', 'hobbies', 'science_fiction', 'romance', 'humor']
feel = ["love", "hate", "doubt", "fear", "pain", "joy",'wrath', 'happiness']
one.tabulate(conditions=cat, samples=feel)
```

```
[nltk_data] Downloading package brown to /Users/acidrain/nltk_data...
[nltk_data]   Package brown is already up-to-date!
```

```
                     love      hate     doubt      fear      pain       joy     wrath happiness
           news         3         1         6         4         1         0         0         0
       religion        13         3         6        24         3         3         3         0
        hobbies         6         0         3         2         0         1         0         0
science_fiction         3         0         2         1         7         0         0         5
        romance        32         9         8         4         5         4         0         5
          humor         4         0         4         5         1         1         0         0
```

religion tiende a tener más sentimientos negativos que el resto de categorias (fear,wrath). love aparece principalmente en romance y religion. sciencie fiction tiene el mayor numero de apariciones de pain y happiness. Hate tambien parece ser un sentimiento repetitivo en romance.El sentimiento mas comun en news es doubt. En general, romance parece ser la categoria que evoca más sentimientos.