Taller 9

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 26-oct-2018 11:59 PM

[Juan Camilo Perdomo]

[juan.perdomor@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 [3]:

```
%matplotlib inline
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = [18.0, 8.0]
import nltk
nltk.download()
from nltk.book import *
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
*** 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 exercise 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 [10]:

```
palabras = set([w for w in text5 if len(w) == 4])
palabras
```

Out[10]:

```
{'east',
 'each',
 'comp',
 'Ruth',
 'ahah',
 'quit',
 'read',
 'Yeah',
 'hail',
 'last',
 'safe',
 'Road',
 'away',
 'brwn',
 'yeah',
 'note',
 'What',
 'show',
 '?!?!',
 'evil',
 'used',
 '1900',
 'boot',
 'foot',
 'fool',
 '6:38',
 'teck',
 'open',
 'surf',
 'Dude',
 'U149',
 'z-ro',
 'Joey',
 't he',
 'half',
 'amen',
 'otay',
 'sexs',
 'left',
 'blah',
 'play',
 'peek',
 'girl',
 'baby',
 'feat',
 'sean',
 'knee',
 'U154',
 'wack',
 'West',
 'slow',
 'dear',
 'hows',
 'none',
 'Take',
 'need',
 'Well',
 '1299',
 'bust',
 '((((',
 'whys',
 'busy',
 'kewl',
 'must',
 'herE',
 'dust',
 'gals',
 'kids',
 'hook',
 'hawT',
 'time',
 'move',
 'hazy',
 'yes.',
```

```
'push',
'out.',
'imma',
'wild',
'Away',
'Deep',
'shes',
'yard',
'sore',
'tjhe',
'pasa',
'FROM',
'come',
'Love',
'live',
'Over',
'U150',
'ewww',
'Tina',
'city',
'VVil',
'urls',
'Chop',
'raed',
'page',
"ex's",
'evah',
'feet',
'pure',
'Care',
'yoko',
'jush',
'ciao',
'crap',
'Hill',
'QUIT',
'john',
'moms',
'boed',
'U122',
'City',
'U164',
',,,,',
'U109',
'calm',
'ROOM',
'sets',
'Haha',
'cold',
'Days',
'body',
'card',
'mike',
'>:->',
'star',
'Look',
'Fade',
'pimp',
'enuf',
'Food',
'golf',
'HERE',
'mind',
'thah',
'walk',
'gret',
'this',
'size',
'itch',
'down',
'stay',
'Wind',
'York',
'givs',
'arms',
'real',
'Pour',
```

```
'lord',
'Liam',
'U102',
'Reub',
'Ahhh',
'wooo',
'idea',
'guns',
'ages',
'doll',
'Heys',
'eats',
'will',
'cams',
'puke',
'U196',
'well',
'talk',
'U112',
'Does',
'Cute',
'rofl',
'dont',
'chik',
'hiii',
'know',
'seat',
'1985',
'rose',
'boom',
'cops',
'roll',
'Ctrl',
'WHEN',
'U542',
'rang',
'door',
'hehe',
'pain',
'part',
'blue',
'pink',
'OOPS',
'THAT',
'Ummm',
'tthe',
'male',
'felt',
'hots',
'sexy',
'HALO',
'just',
'nuff',
'swim',
'game',
'1.99',
'halo',
'wife',
'scar',
'grea',
'wire',
'aunt',
'Rock',
'tyvm',
'Were',
'FACE',
'DAMN',
'lapd',
'cock',
'U138',
'Damn',
'work',
'runs',
'outa',
'Elev',
'U219',
'roof',
```

```
'nude',
'Heyy',
'buff',
'idnt',
'deaf',
'send',
'cums',
'choc',
'Bone',
'When',
'rain',
'seem',
'ally',
'sayn',
'lala',
"yw's",
'disc',
'wean',
'rest',
'Hugs',
'Lord',
'hits',
'offa',
'cost',
'Dawn',
'whoo',
'WHOA',
'gosh',
'ride',
'sort',
'exit',
'Lion',
'U153',
'wins',
'twin',
'base',
'guys',
'mass',
'That',
'Nice',
'Here',
'Kewl',
'Live',
'hiya',
'haze',
'asks',
'flow',
'toop',
'post',
'park',
'EVEN',
'plan',
'blow',
'NONE',
'eeww',
'allo',
'1980',
'jude',
'Sat.',
'they',
'oops',
'AKST',
'jump',
'hate',
'oooh',
'past',
'waht',
'ghet',
'chip',
'moon',
'nerd',
'AKDT',
'wana',
'fuck',
'ummm',
'Kold',
'U132',
```

```
'tory',
'mark',
'Need',
'syck',
'1996',
'fire',
'Hero',
'18ST',
'limp',
'U115',
'<~~~',
'MUAH',
'lois',
'PART',
'Have',
'lube',
'zone',
'anti',
'prep',
'poor',
'heat',
'ugly',
'Holy',
'sink',
'ebay',
'spit',
'gawd',
'.. .',
'like',
'ELSE',
'flaw',
'U146',
'gees',
'word',
'Girl',
'duet',
'sigh',
'glad',
'gays',
'U142',
'KNOW',
'U169',
'!???',
'kill',
'wats',
'vent',
'draw',
'frst',
'heee',
'bacl',
'Dang',
'Lime',
'dint',
'kind',
'blew',
'slap',
'pigs',
'cant',
'most',
'thru',
'Came',
'thnx',
'Judy',
'tape',
"pm's",
'tooo',
'HAHA',
'kong',
'them',
'dick',
'help',
'much',
'lazy',
'!!!.',
'sing',
'west',
'any1',
```

'clue',
'slam',
'45.5',
'salt',
'days',
'whud',
'nods',
'Then',
'lame',
'ltnc',
'YALL',
'tell',
'dang',
'look',
'sang',
'corn',
'back',
'nads',
'whip',
'been',
'Wyte',
'ruff',
'fock',
'able',
'U170',
'News',
'tlak',
'luvs',
'brad',
'opps',
'tere',
'ello',
'wazz',
'hgey',
'bong',
'ogan',
'take',
'awww',
'also',
'Swim',
'sent',
'bare',
'Same',
'Tell',
'nice',
'Tide',
'John',
'woot',
'icky',
'64.8',
'okey',
'junk',
'slip',
'ways',
'Gosh',
'jack',
'2:55',
'Iowa',
'Matt',
'U175',
'Rang',
'98.5',
'Lets',
'ribs',
'over',
'lost',
'into',
'ssid',
'Will',
'Drew',
'98.6',
'U197',
'MODE',
'Rofl',
'puff',
'U128',
'daft',

'mins',
'Hiya',
'U113',
'noth',
'stop',
'1930',
'food',
'hugs',
'alot',
'WITH',
'bout',
'Seee',
'your',
'cyas',
'CHAT',
'pray',
'lead',
'crib',
'poop',
'Elle',
'SExy',
'deal',
'legs',
'yawn',
'wine',
'quiz',
'have',
'bugs',
'U123',
'ouch',
'hall',
'grew',
'plus',
'perv',
'heya',
'U144',
'U172',
'ring',
'wear',
'ther',
'brat',
'bull',
'wide',
'Okay',
'holy',
'They',
'road',
'area',
'.op.',
'Save',
'heyy',
'This',
'jail',
'From',
'U108',
'adds',
'scum',
'keep',
'lawl',
'went',
'tail',
'DOES',
'late',
'from',
'pool',
'lust',
'beat',
'U520',
'Werd',
'LATE',
'poem',
'orgy',
'nawt',
'pics',
'even',
'hard',
'fake',

'only',
'good',
'wish',
'thje',
'DING',
'loss',
'gone',
'mama',
'sure',
'face',
'Type',
'soup',
'U165',
'KoOL',
'U103',
'bear',
'clap',
'spat',
'lots',
'100%',
'lmao',
'Last',
'lool',
'wrap',
'typo',
'knew',
'GOOD',
'Awww',
'Teck',
'waaa',
'Cool',
'burp',
'Rush',
'Lmao',
'land',
'beam',
'mean',
'o.k.',
'isnt',
'cute',
'yall',
'U118',
'room',
'xmas',
'self',
'Oops',
'10th',
'soul',
'club',
'<<<<',
'Tisk',
'Nova',
'U136',
'cure',
'U190',
'Meep',
'rock',
'perk',
'cepn',
'pmsl',
'U988',
'benz',
'phil',
'<333',
'tips',
'Talk',
'2Pac',
'Turn',
'tart',
'poot',
'true',
'gooo',
'ahem',
'9.53',
'meds',
'that',
'wher',

'ever',
'song',
'Rule',
'Time',
'four',
'died',
'town',
'jeff',
'THEY',
'U137',
'U156',
'aint',
'turn',
'aime',
'poll',
'same',
'hong',
'Evil',
'soda',
'mono',
'CALI',
'Very',
'pork',
'smax',
'pine',
'MORE',
'Life',
'9:10',
'numb',
'than',
'U148',
'akon',
'ones',
'dojn',
'whou',
'kmph',
'toss',
'lake',
'U119',
'SEEN',
'YOUR',
'woah',
'fish',
'mite',
'easy',
'mkay',
'NAME',
'Nooo',
'NICK',
'4.20',
'uyes',
'high',
'else',
'newp',
'cali',
'hooo',
'hint',
'woof',
'argh',
'deep',
'kept',
'Show',
'mmmm',
'bike',
'temp',
'eyes',
'sell',
'Poor',
'care',
'shot',
'cuss',
'seee',
'hell',
'head',
'<---',
'dyed',
'coat',

```
'coat',
'inch',
'Come',
'VBox',
'TIME',
'wOOt',
'AWAY',
'howl',
'rape',
'mena',
'huge',
'hiom',
'hawt',
'army',
'pass',
'sips',
'U105',
'says',
'nawp',
'sand',
'span',
'hear',
'wrek',
'drop',
'clay',
'U141',
'mami',
'fair',
'Hand',
'form',
'U104',
'"...',
'once',
'U147',
'TALK',
'hide',
'very',
'cook',
'worl',
'Stop',
'Kick',
'yell',
'HUGE',
'MRIs',
'howz',
'when',
'U181',
'nope',
'piff',
'till',
'asss',
'bred',
'damn',
'bois',
'gets',
'3:45',
'ball',
'7:45',
'Even',
'caan',
'book',
'U120',
'nick',
'suck',
'????',
'goof',
'!...',
'hand',
'cool',
'team',
'band',
'U100',
'CAPS',
'Only',
'fall',
'guyz',
'wind',
'keys',
```

'keys',
'O.k.',
'year',
'ROFL',
'Eggs',
'pwns',
'ladz',
'HAVE',
'paid',
'LAst',
'2006',
'menu',
'rich',
'eric',
'seth',
'JUST',
'near',
'Jane',
'Room',
'rush',
'crop',
'FINE',
'tock',
'wuts',
'made',
'hmmm',
'U101',
'Hott',
'None',
'shup',
'grrr',
'SOME',
'soon',
'find',
'term',
'lisa',
'hour',
'gift',
'tits',
'boss',
'sext',
'heal',
'dump',
'sock',
'docs',
'U117',
'free',
'doin',
'kent',
'lets',
'ques',
'haaa',
'sori',
'mofo',
'shop',
'PMSL',
'bird',
'heck',
'Sexy',
'dawg',
"RN's",
'U820',
'Rick',
'hogs',
'chit',
'tisk',
'STOP',
'hola',
'herd',
'took',
'High',
'Ohhh',
'dude',
'1cos',
'yeee',
'Paul',
'mahn',
'boys',

'boys',
'life',
'febe',
'joke',
'Good',
'hmph',
'U163',
'Like',
'pfft',
'home',
'fine',
'lies',
'NTMN',
'Ohio',
'byes',
'yada',
'luck',
'rubs',
'!!!!',
'spot',
'U133',
'Sure',
'ears',
'give',
'bloe',
'said',
'kool',
'weed',
'mauh',
'test',
'Maps',
'Prof',
'U989',
'6:51',
'some',
'site',
'born',
'vamp',
'gold',
'caca',
'dogs',
'such',
'done',
'ohio',
'TYPR',
'lol.',
'n9ne',
'LoVe',
'toes',
'ooer',
'soft',
'6:53',
'scuk',
'bowl',
'39.3',
'coem',
'snow',
'Home',
'U121',
'U111',
'fart',
'dark',
'dead',
'Yoko',
'nada',
'SSRI',
'fast',
'poof',
'week',
'tick',
'cmon',
'owww',
'hang',
'1200',
'eeek',
'####',
'Help',
'mode',

'mode',
'Boyz',
'plow',
'rats',
'anal',
'bone',
'pies',
'name',
'hill',
'tiff',
'mine',
'fits',
'wait',
'hair',
'Male',
'serg',
'pour',
'addy',
'bein',
'best',
'este',
'elle',
'U107',
'geez',
'side',
'bied',
'U110',
'GrlZ',
'Troy',
'sits',
'gray',
'spin',
'LIVE',
'??!!',
'U114',
'seen',
'giva',
'cars',
'tend',
'neck',
'lick',
'fear',
'Fort',
'tenn',
'4:03',
'okay',
'U129',
'told',
'saME',
'does',
'U106',
"pm'n",
'U155',
'deop',
'; ..',
'peel',
'loud',
'drew',
'bomb',
'shit',
'mess',
'U126',
'call',
'U130',
'U134',
'goin',
'Uhhh',
'Your',
'pope',
'laid',
'full',
'Song',
'....',
'Lies',
'Hold',
'line',
'dies',
'fcuk',

```
 'rawk',
 'babe',
 'five',
 ...}
```

In [11]:

```
len(palabras)
```

Out[11]:

```
1181
```

Chapter 1 exercise 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(palabra) for palabra in text1)
```

Out[26]:

```
999044
```

La función retorna la suma de las longitudes de todas las palabras del texto 1.

Chapter 1 exercise 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 [45]:

```
def porcentaje(palabras, texto):
    palabras = len(text)
    texto = text.count(palabras)
    print(ocurencias / floac(total)) * 100
```

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

In [14]:

```
words = nltk.corpus.gutenberg.words('austen-emma.txt')
len(words)
```

Out[14]:

```
192427
```

In [15]:

```
len(set(words))
```

Out[15]:

```
7811
```

El libro tiene 192.427 'word tokens' y 7811 'types'.

Chapter 2 exercise 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 [24]:

```
from nltk.corpus import state_union
print(state_union.fileids())
```

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

```
['1945-Truman.txt', '1946-Truman.txt', '1947-Truman.txt', '1948-Truman.txt', '1949-Truman.txt', '1
950-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-For
d.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-B
ush-2.txt', '1992-Bush.txt', '1993-Clinton.txt', '1994-Clinton.txt', '1995-Clinton.txt', '1996-Cli
nton.txt', '1997-Clinton.txt', '1998-Clinton.txt', '1999-Clinton.txt', '2000-Clinton.txt', '2001-G
WBush-1.txt', '2001-GWBush-2.txt', '2002-GWBush.txt', '2003-GWBush.txt', '2004-GWBush.txt', '2005-
GWBush.txt', '2006-GWBush.txt']
```

Chapter 2 exercise 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 [25]:

```
from nltk.corpus import brown
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']
general_words = ["who", "what", "when", "where", "why", "how"]
cfd.tabulate(conditions=genres, samples=general_words)
```

```
                  who  what  when where   why   how
           news   268    76   128    58     9    37
       religion   100    64    53    20    14    23
        hobbies   103    78   119    72    10    40
science_fiction    13    27    21    10     4    12
        romance    89   121   126    54    34    60
          humor    48    36    52    15     9    18
```