Taller 9

Métodos Computacionales para Políticas Públicas - URosario

Entrega: viernes 30-abr-2021 11:59 PM

[Juan Diego Castro Rodríguez]

[juand.castro@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

Capítulo 1, ejercicio 22

In [1]:
```
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = [18.0, 8.0]
```

In [2]:
```
import nltk
```

In [4]:
```
nltk.download()
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

Out[4]:
```
True
```

In [5]:
```
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
```

In [7]:
```
fdist5 = FreqDist(text5)
fdist5
```

Out[7]:
```
FreqDist({'.': 1268, 'JOIN': 1021, 'PART': 1016, '?': 737, 'lol': 704, 'to': 658, 'i': 648, 'the': 646, 'you': 635, ',': 596, ...})
```

In [9]:
```
four_letter_words = [w for w in text5 if len(w) == 4]
```

In [10]:
```
fdist4lw= FreqDist(four_letter_words)
fdist4lw
```

Out[10]:
```
FreqDist({'JOIN': 1021, 'PART': 1016, 'that': 274, 'what': 183, 'here': 181, '....': 170, 'have': 164, 'like': 156, 'with': 152, 'chat': 142, ...})
```

Ejercicio 26

In [11]:
```
sum(len(w) for w in text1)
```

Out[11]:
```
999044
```

In [12]:
```
sum([len(w) for w in text1]) / len(text1)
```

Out[12]:
```
3.830411128023649
```

Ejercicio 28

In [20]:
```
def percent(word, text):
    count = 0
    for w in text:
        if word == w:
            count+=1
    return str(float(count) / float(len(text)) *100)+"%"
```

In [21]:
```
percent("monstrous", text1)
```

Out[21]:
```
'0.0038340765051625 84%'
```

Capítulo 2, Ejercicio 2

In [22]:
```
from nltk.corpus import gutenberg
```

In [23]:
```
austen=gutenberg.words('austen-persuasion.txt')
```

In [24]:
```
len(austen)
```

Out[24]:
```
98171
```

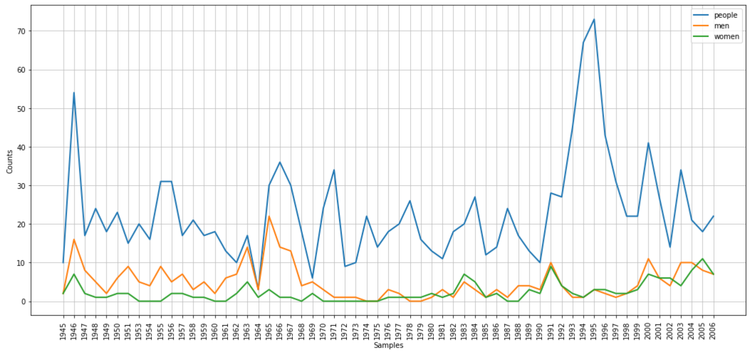
In [25]:
```
len(set(s.lower() for s in austen if s.isalpha()))
```

Out[25]:
```
5739
```

Ejercicio 4

In [26]:
```
from nltk.corpus import state_union
```

In [27]:
```
cfd = nltk.ConditionalFreqDist(
          (target, fileid[:4])
          for fileid in state_union.fileids()
          for w in state_union.words(fileid)
          for target in ['men', 'women', 'people']
          if w.lower().startswith(target))
cfd.plot()
```



Out[27]:
```
<AxesSubplot:xlabel='Samples', ylabel='Counts'>
```

Ejercicio 11

In [29]:
```
from nltk.corpus import brown
modals = ['can', 'could', 'may', 'might', 'must', 'will']
cfd = nltk.ConditionalFreqDist(
          (genre, word.lower())
    for genre in brown.categories()
          for word in brown.words(categories=genre))
genres = ['news', 'religion', 'hobbies', 'science_fiction', 'romance', 'humor']
cfd.tabulate(conditions=genres, samples=modals)
```

```
                can could   may might  must  will
          news   94    87    93    38    53   389
      religion   84    59    79    12    54    72
       hobbies  276    59   143    22    84   269
science_fiction   16    49     4    12     8    17
       romance   79   195    11    51    46    49
         humor   17    33     8     8     9    13
```

In [30]:
```
pronouns = ['I', 'you', 'he', 'she', 'it', 'we', 'they']
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']
cfd.tabulate(conditions=genres, samples=pronouns)
```

```
                 I  you   he  she   it   we they
          news  179   55  451   42  363   77  205
      religion  155  100  137   10  264  176  115
       hobbies  154  383  155   21  476  100  177
science_fiction   98   81  139   36  129   30   53
       romance  951  456  702  496  573   78  168
         humor  239  131  146   58  162   32   70
```

Sobre verbos modales las noticias especulan sobre posibles consecuencias, luego el "will" es más frecuente, en los Hobbies se dan situaciones donde se quiere mostrar lo que se puede hacer o pedir un favor, de ahí que "may" "will" y "can" se usen frecuentemente. Y sobre los pronombres, en las noticias "he" es el más frecuente, de pronto por algún enfoque de género. En Hobbies "it" es más frecuente porque se habla justamente de aquella acción. En romance "I" se utiliza para expresar emociones propias y "he" para hacer referencia a la persona que normalmente toma la iniciativa.