**Almacenamiento y Procesamiento Masivo de Datos**

**Tarea03**

**Alumno:** Benjamín Corvalán

El primer problema que se presentó, fue que los datos son muchos (en el dataset original), por ende guardar los datos o tratar de leerlos todos de en una corrida, da un Memory Error. La solución a esto fue guardar los datos que se necesitaban como archivos separados e ir leyendo línea a línea, y no tratar de leer todas las líneas de un tirón.

Por ende el primer paso fue la separación en dos archivos, para los reviews con dos o menos estrellas y otro archivo para los con 3 o más estrellas, para que luego el trabajo sea más fácil ya que no se deberá separar entre ellos.

El procedimiento para los cuatro problemas, será trabajar en lo que se pide y probar con varios sub-sets con menor cantidad de datos. Los cuáles serán 10.000, 100.000 y en lo posible con todos los datos. En vez de los 2.685.066 de datos del dataset completo, para probar los códigos. Luego se verá si el resultado se obtiene en un tiempo razonable y estimar cuanto se demoraría con el dataset completo, si no fuera posible correrlo con este.

La estimación es a mano, o sea que se dejó la diferencia más grande entre los tiempos que le demoró a cada pequeño dataset realizar el proceso.

Claramente la estimación no va a estar correcta en relación al tiempo que va tomar en la realidad, puede ser que el tiempo real sea mucho menos o más que la estimación.

Lo ideal sería poder correr con el dataset completo, pero la realidad es que en algunos casos esto requeriría mucho tiempo. Por lo que se colocaran los tiempos demorados para el caso del dataset pequeño, la estimación y si fue posible el tiempo en el dataset completo.

Todos los tiempos están medidos con tiempo de procesador, y fueron corridos de forma lineal, o sea no de forma distribuida.

**1.**

Con 10.000 datos como dataset, se obtuvo las siguientes 20 palabras más ocupadas, para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

. 7431

the 6987

and 6821

a 6365

, 6147

i 6064

to 5847

is 5085

of 5061

it 4947

for 4888

in 4855

was 4095

but 3974

this 3808

my 3802

with 3720

that 3680

! 3655

on 3631

Top 20 words for 2- Stars ratings.

. 2248

the 2109

and 2021

to 1923

a 1919

i 1902

, 1856

was 1647

it 1636

of 1597

for 1520

in 1485

is 1379

this 1361

that 1344

not 1326

n't 1303

but 1299

my 1236

on 1195

Con 100.000 datos como dataset, se obtuvo las siguientes 20 palabras más ocupadas, para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

. 74280

the 69537

and 69101

a 61686

i 59975

, 59806

to 57169

is 49577

of 48632

for 48105

it 47452

in 45789

was 39859

my 38221

this 38081

! 38032

but 37309

with 36572

that 34926

have 34732

Top 20 words for 2- Stars ratings.

. 21504

the 20297

and 19519

to 18483

i 18301

a 18192

, 17268

was 15723

it 15155

of 14907

for 14748

in 14302

this 13154

not 13097

is 12944

that 12847

my 12532

n't 12349

but 11999

they 11728

Con todo el dataset, se obtuvo las siguientes 20 palabras más ocupadas, para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

. 2008608

the 1888234

and 1875192

a 1643547

i 1601473

, 1583491

to 1537206

is 1305858

of 1283949

it 1276403

for 1265706

in 1202669

was 1187749

! 1140592

this 1050349

my 1023627

with 994316

but 981175

that 908630

on 895858

Top 20 words for 2- Stars ratings.

. 536564

the 510018

and 489009

to 464309

i 454198

a 453061

, 424797

was 405422

it 382200

for 374019

of 369605

in 351232

not 333368

this 329697

is 323852

that 318525

n't 312062

my 306066

but 305211

they 300403

Tiempo Dataset 10.000: 9.94314530248

Tiempo Dataset 100.000: 93.3192574668

Tiempo Dataset completo: 2718.27574554

Como comentarios de este ejercicio, se puede ver claramente la influencia de los “stopwords” en las palabras más usadas, ya que son los conectores que en casi todo texto con sentido debieran estar para que este tenga coherencia. Lo que hay que destacar es que solo se contó el “stopword” que apareciese solo 1 vez por review, por lo que esto demuestra que todas las veces que fueron ocupadas, implica que fueron ocupadas en la misma cantidad de reviews, lo cual es algo importante de entender.

**2.**

Con 10.000 datos como dataset, se obtuvo las siguientes 20 bigrams (par de palabras) más ocupadas, para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

(u'.', u'i') 4071

(u'.', u'the') 3603

(u',', u'and') 2620

(u',', u'but') 2334

(u'of', u'the') 2203

(u'in', u'the') 2193

(u'and', u'the') 2029

(u'.', u'it') 1902

(u'it', u"'s") 1886

(u'it', u'was') 1722

(u',', u'i') 1665

(u'and', u'i') 1601

(u'for', u'a') 1454

(u'.', u'they') 1445

(u'this', u'place') 1427

(u'if', u'you') 1413

(u'i', u'was') 1400

(u'on', u'the') 1373

(u'is', u'a') 1341

(u'to', u'the') 1293

Top 20 words for 2- Stars ratings.

(u'.', u'i') 1340

(u'.', u'the') 1164

(u',', u'but') 789

(u',', u'and') 716

(u'it', u'was') 711

(u'of', u'the') 699

(u'in', u'the') 692

(u',', u'i') 633

(u'.', u'it') 600

(u'i', u'was') 582

(u'and', u'the') 579

(u'do', u"n't") 553

(u'this', u'place') 515

(u'to', u'the') 512

(u'and', u'i') 498

(u'on', u'the') 494

(u'to', u'be') 489

(u'it', u"'s") 484

(u'for', u'a') 470

(u',', u'the') 461

Con 100.000 datos como dataset, se obtuvo las siguientes 20 bigrams (par de palabras) más ocupadas, para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

(u'.', u'i') 38928

(u'.', u'the') 33775

(u',', u'and') 23345

(u',', u'but') 21138

(u'of', u'the') 19594

(u'and', u'the') 19517

(u'in', u'the') 18901

(u'.', u'it') 16695

(u'it', u"'s") 16339

(u'it', u'was') 15903

(u',', u'i') 15438

(u'and', u'i') 15225

(u'this', u'place') 14717

(u'.', u'they') 13791

(u'on', u'the') 12852

(u'i', u'have') 12782

(u'for', u'a') 12736

(u'i', u'was') 12651

(u'if', u'you') 12520

(u'is', u'a') 11775

Top 20 words for 2- Stars ratings.

(u'.', u'i') 12865

(u'.', u'the') 10177

(u'it', u'was') 6468

(u'in', u'the') 6440

(u',', u'but') 6378

(u'of', u'the') 6239

(u',', u'and') 6140

(u',', u'i') 5926

(u'i', u'was') 5804

(u'and', u'the') 5552

(u'do', u"n't") 5068

(u'and', u'i') 4896

(u'to', u'the') 4837

(u'.', u'it') 4781

(u'on', u'the') 4637

(u'this', u'place') 4605

(u'to', u'be') 4572

(u'for', u'the') 4371

(u'for', u'a') 4327

(u'did', u"n't") 4281

Con todo el dataset, se obtuvo las siguientes 20 bigrams (par de palabras) más ocupadas, para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

(u'.', u'i') 1013443

(u'.', u'the') 912966

(u',', u'and') 578063

(u'and', u'the') 541064

(u'of', u'the') 531518

(u',', u'but') 509996

(u'it', u'was') 481354

(u'in', u'the') 457076

(u'.', u'it') 437430

(u'this', u'place') 428720

(u'it', u"'s") 421315

(u'and', u'i') 413904

(u',', u'i') 388190

(u'on', u'the') 371163

(u'.', u'we') 347277

(u'.', u'they') 344078

(u'i', u'was') 339060

(u'for', u'a') 324845

(u'if', u'you') 324229

(u'the', u'food') 319681

Top 20 words for 2- Stars ratings.

(u'.', u'i') 310025

(u'.', u'the') 252652

(u'it', u'was') 168721

(u'of', u'the') 154668

(u'in', u'the') 152282

(u',', u'but') 147197

(u'i', u'was') 140295

(u'and', u'the') 139682

(u',', u'i') 138940

(u',', u'and') 138024

(u'do', u"n't") 127650

(u'this', u'place') 124085

(u'to', u'the') 122043

(u'and', u'i') 121878

(u'on', u'the') 120159

(u'.', u'it') 118476

(u'for', u'the') 112902

(u'did', u"n't") 112708

(u'.', u'we') 112637

(u'to', u'be') 110176

Tiempo Dataset 10.000: 11.7296530444

Tiempo Dataset 100.000: 115.858378627

Tiempo Dataset completo: 3066.1622093

Como comentarios de este ejercicio, se puede ver que patrones sigue cada review, o las personas en general a la hora de escribir, lo cual es interesante. También sirve para ver patrones de que se ocupa en general antes o después de una cierta palabra. Lo malo de este ejercicio es que tampoco limitamos a los “stopwords”, en nuestro bigrams, ni las puntuaciones, ni “shortcuts” del lenguaje propio. Lo que hay que destacar es que solo se contó cada para de palabras (1 bigram) que apareciese, solo 1 vez por review, por lo que esto demuestra que todas las veces que fueron ocupadas, implica que fueron ocupadas en la misma cantidad de reviews, lo cual es algo importante de entender. También es importante mencionar que el tiempo que toma el cálculo de las mejores 20 palabras mediante bigrams aumenta bastante.

**3.**

Con 10.000 datos como dataset, se obtuvo las siguientes 20 palabras más ocupadas sin stopwords (en inglés), para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

. 7431

, 6147

! 3655

's 3551

n't 3208

good 3013

place 2871

great 2740

food 2452

) 2369

( 2254

like 2038

get 2004

one 1921

go 1797

time 1771

would 1713

service 1705

also 1656

really 1596

Top 20 words for 2- Stars ratings.

. 2248

, 1856

n't 1303

's 1031

food 891

place 842

would 772

like 760

get 735

one 732

! 727

service 696

( 670

back 667

good 664

) 662

time 639

go 608

even 584

'' 542

Con 100.000 datos como dataset, se obtuvo las siguientes 20 palabras más ocupadas sin stopwords (en inglés), para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

Top 20 words for 3+ Stars ratings.

. 74280

, 59806

! 38032

's 30781

n't 29526

great 28938

good 28665

place 28494

food 26138

) 21851

( 20158

service 19908

like 19759

get 18518

time 18087

one 17694

go 17304

back 15747

would 15578

really 15155

Top 20 words for 2- Stars ratings.

. 21504

, 17268

n't 12349

's 8729

would 7627

! 7597

place 7477

food 7456

like 6958

one 6922

get 6767

service 6641

back 6590

time 6432

( 6221

) 6160

good 6091

go 5874

'' 5244

... 5218

Con todo el dataset, se obtuvo las siguientes 20 palabras más ocupadas sin stopwords (en inglés), para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos):

No se pudo.

Tiempo Dataset 10.000: 299.175021587

Tiempo Dataset 100.000: 2959.59097763

Estimación Tiempo Dataset completo: 22,1925 horas

Tiempo Dataset completo: No se pudo.

La estimación del dataset completo nos muestra que, ya que luego de tokenizar las palabras en un arreglo (sin repetidas) y luego buscar las palabras dentro de ese arreglo que sean stopwords es O(n) [para todas las palabras en mi arreglo] \* O(n) [buscar dentro del arreglo de stopwords] o sea esta parte es O(n^2) y considerando que esto es para cada línea entonces el algoritmo finalmente es de O(n^3), lo cual explica porque se demora tanto.

Y debido a que se demora tanto, no podre calcularlo antes de la entrega propia de la tarea.

Como comentarios de este ejercicio, se puede ver claramente como la influencia de los “stopwords” se elimina comparando con el ejercicio 1, permitiendo ver las palabras más importantes/relevantes, pero este ejercicio no está exento de problemas, ya que si bien sacamos los “stopwords”, aún queda por sacar todo lo que tiene que ver con puntuación en el lenguaje, dígase exclamaciones, preguntas, punto, coma. Y el tema de los “shortcuts” de palabras en el idioma inglés. Lo que hay que destacar es que solo se contó cada palabra que apareciese, solo 1 vez por review, por lo que esto demuestra que todas las veces que fueron ocupadas, implica que fueron ocupadas en la misma cantidad de reviews, lo cual es algo importante de entender. También es importante mencionar que el tiempo que toma el cálculo de las mejores 20 palabras sin stopwords aumenta considerablemente.

**4.**

Con 10.000 datos como dataset, se obtuvo las siguientes 20 palabras más ocupadas sin stopwords (en inglés), para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos), con sus estrellas respectivas (la interpretación es para los 3+ estrellas, el primer número del arreglo es la cantidad de veces que se dice en reviews con 3 estrellas, el siguiente es lo mismo pero para 4 estrellas y el ultimo lo mismo pero con 5 estrellas):

Top 20 words for 3+ Stars ratings.

. 7431 [1504, 2611, 3316]

, 6147 [1292, 2208, 2647]

! 3655 [463, 1204, 1988]

's 3551 [760, 1310, 1481]

n't 3208 [814, 1118, 1276]

good 3013 [786, 1267, 960]

place 2871 [572, 1049, 1250]

great 2740 [413, 973, 1354]

food 2452 [614, 955, 883]

) 2369 [510, 870, 989]

( 2254 [506, 831, 917]

like 2038 [520, 740, 778]

get 2004 [448, 713, 843]

one 1921 [442, 696, 783]

go 1797 [386, 631, 780]

time 1771 [403, 594, 774]

would 1713 [492, 609, 612]

service 1705 [392, 610, 703]

also 1656 [324, 605, 727]

really 1596 [386, 603, 607]

Top 20 words for 2- Stars ratings.

. 2248 [1231, 1017]

, 1856 [976, 880]

n't 1303 [691, 612]

's 1031 [504, 527]

food 891 [441, 450]

place 842 [451, 391]

would 772 [437, 335]

like 760 [385, 375]

get 735 [365, 370]

one 732 [400, 332]

! 727 [438, 289]

service 696 [371, 325]

( 670 [349, 321]

back 667 [394, 273]

good 664 [276, 388]

) 662 [342, 320]

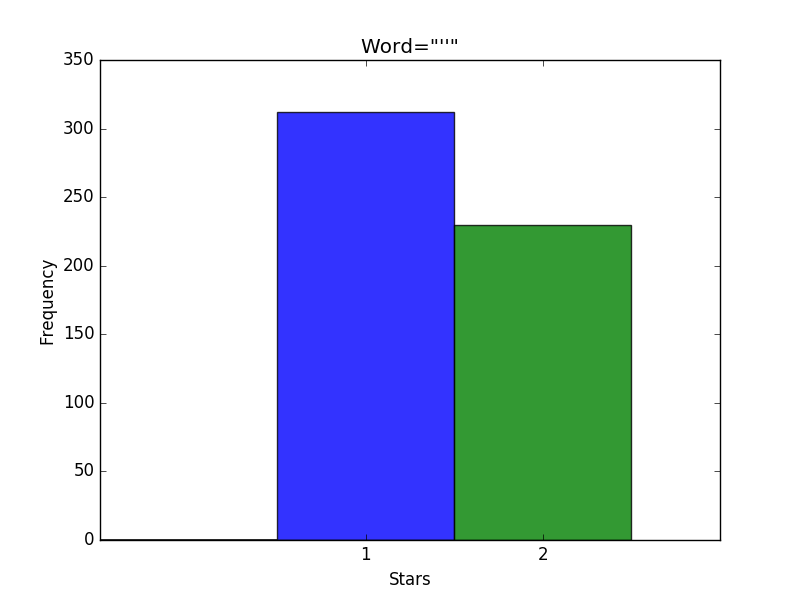
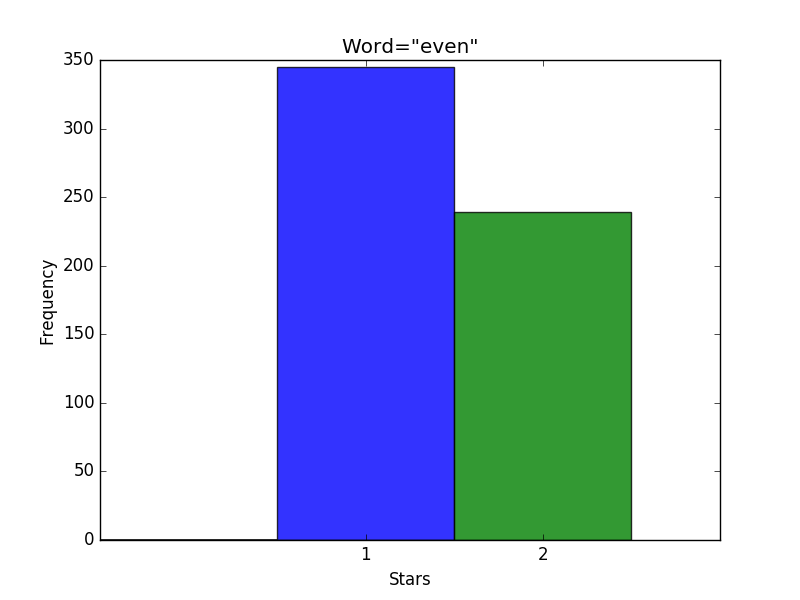
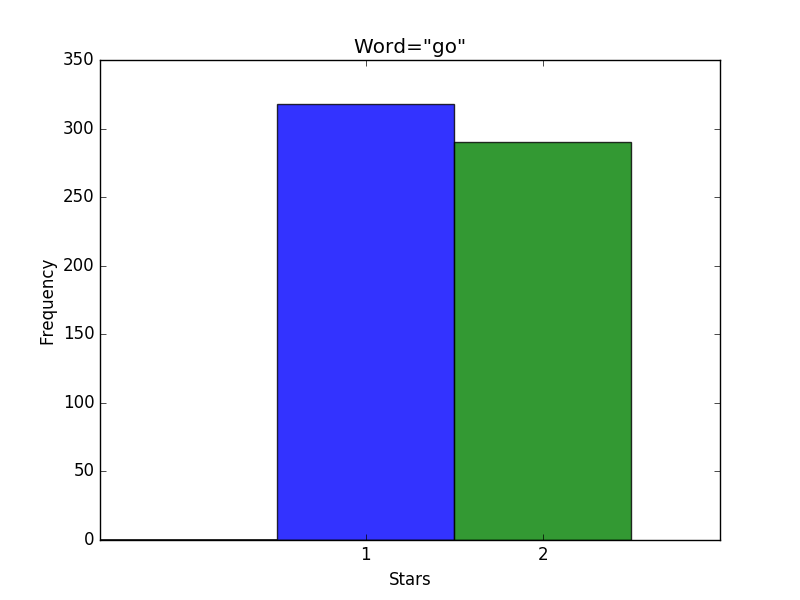
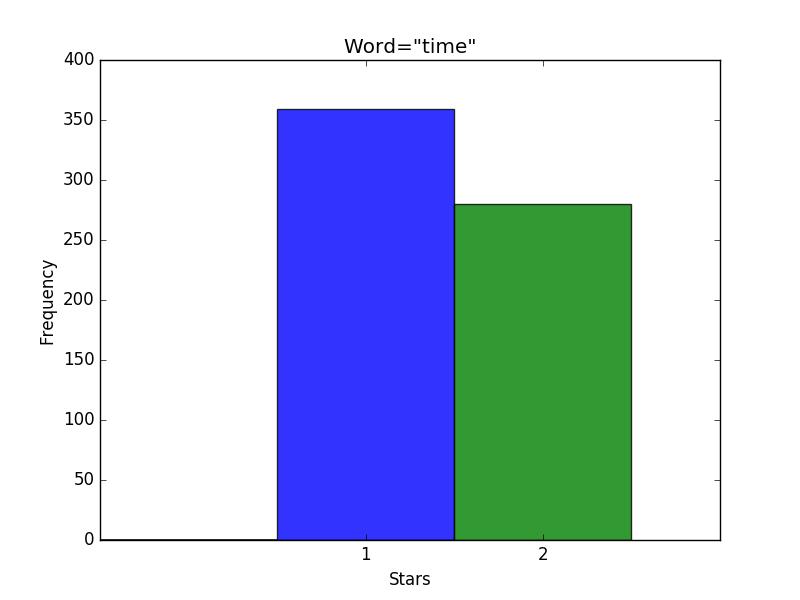
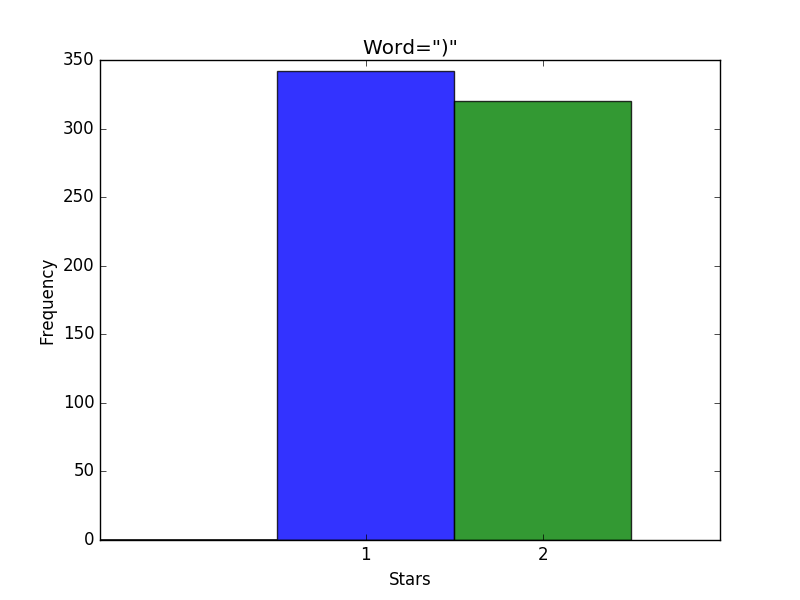
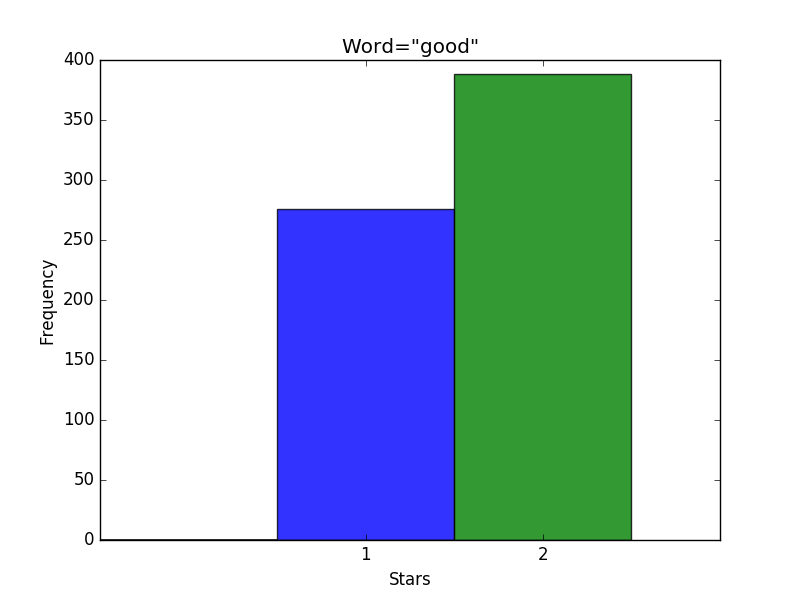
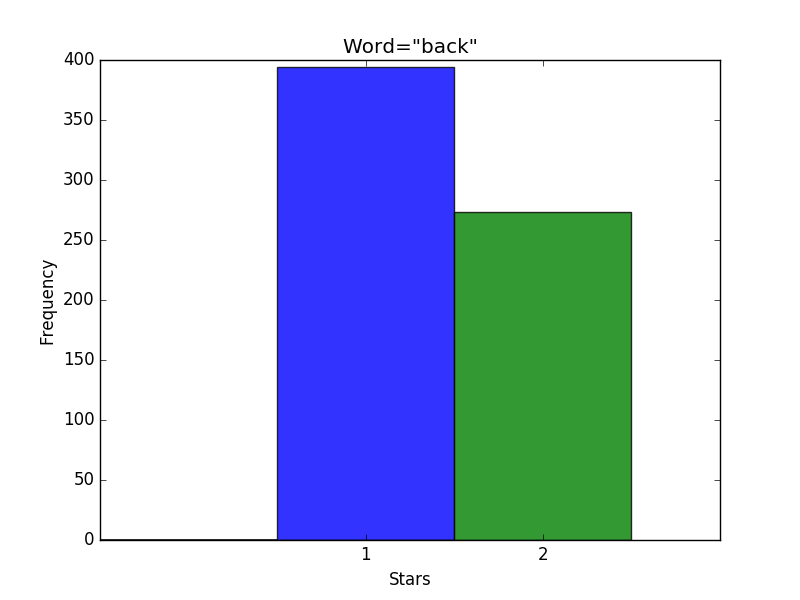
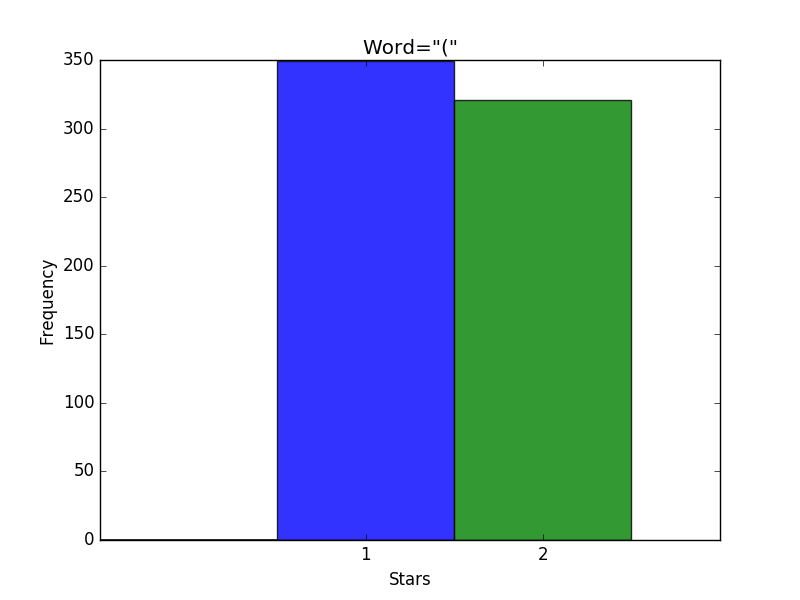
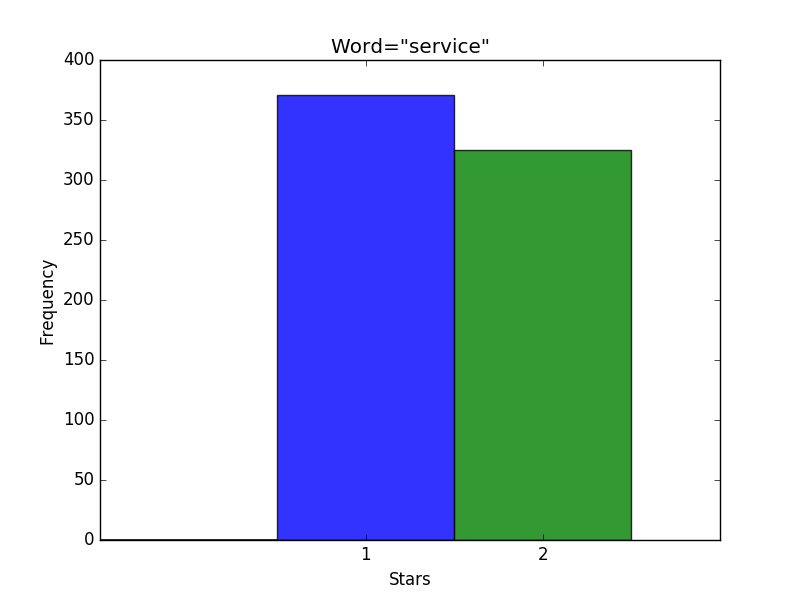
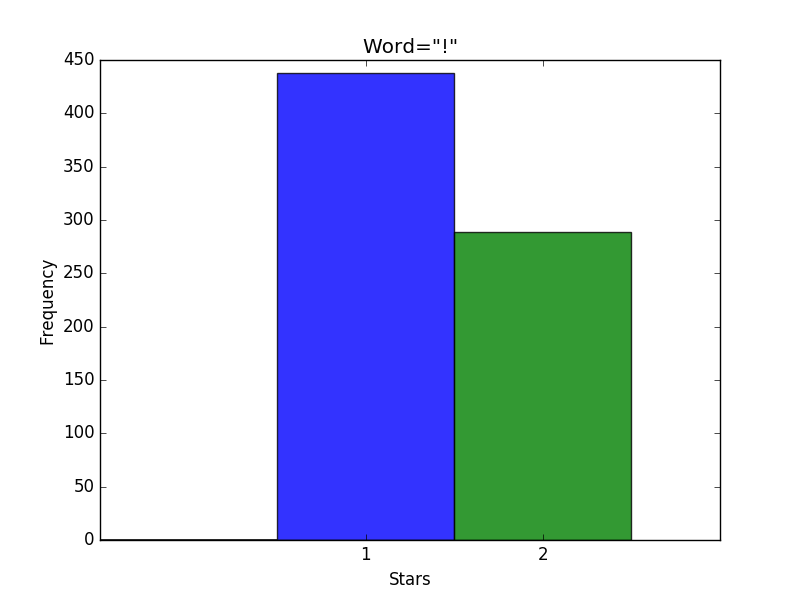
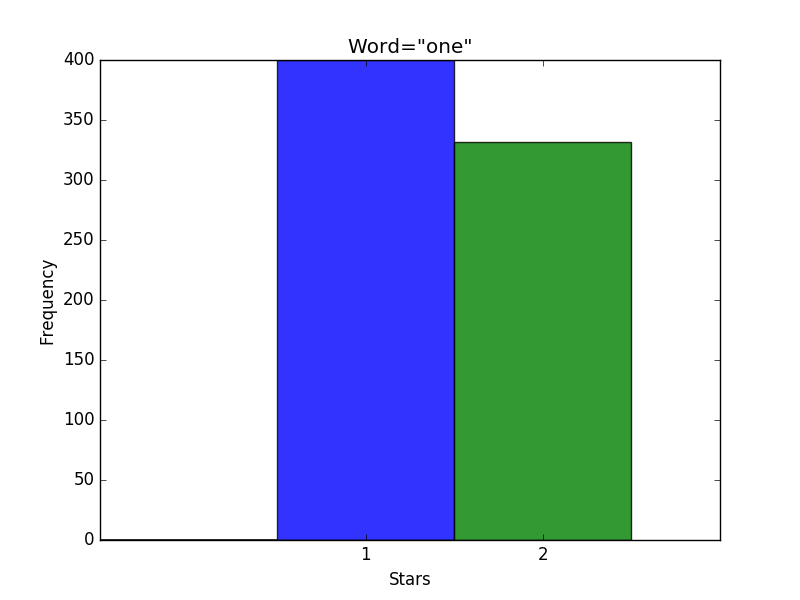
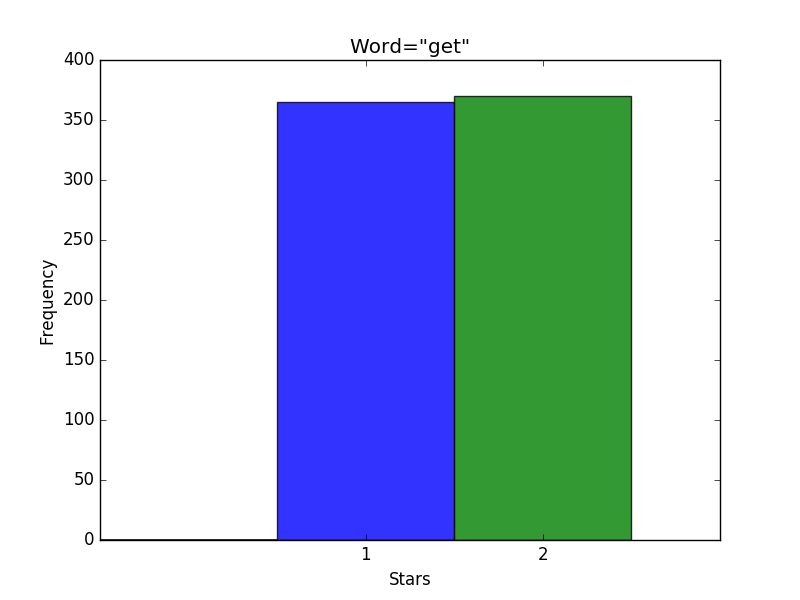
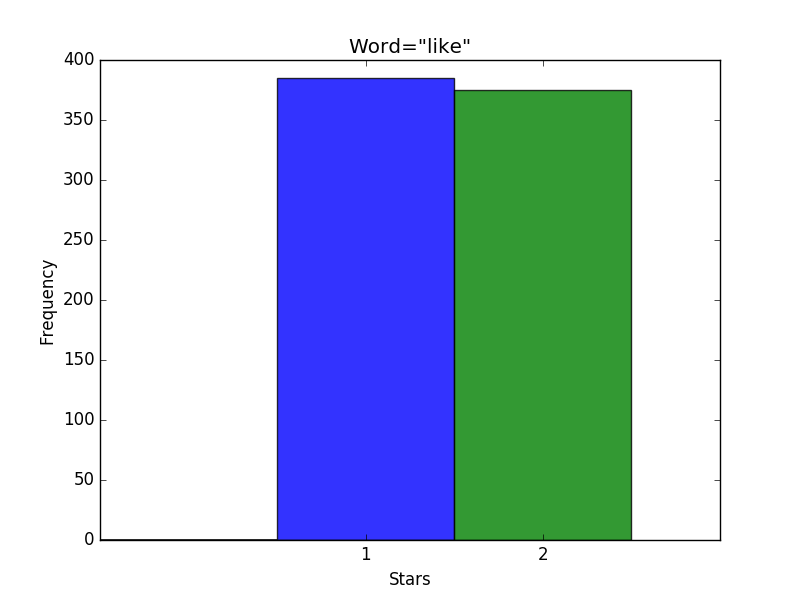
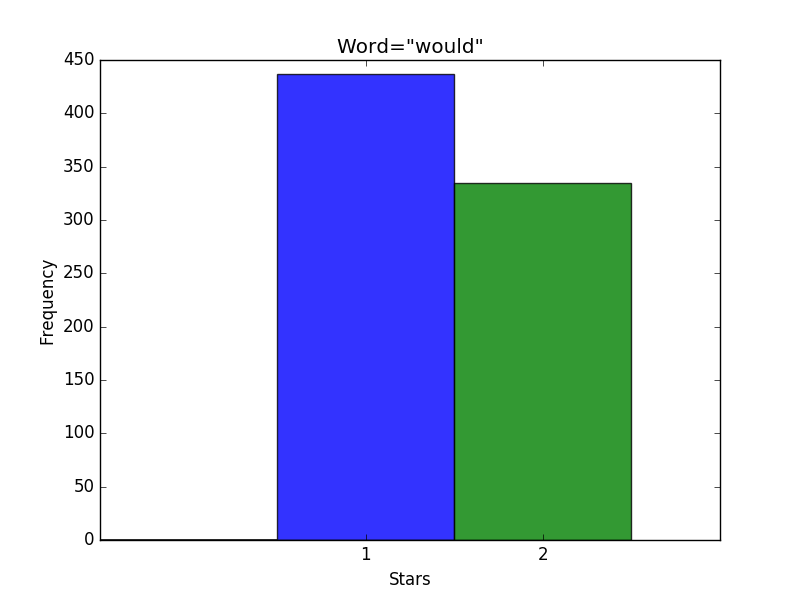
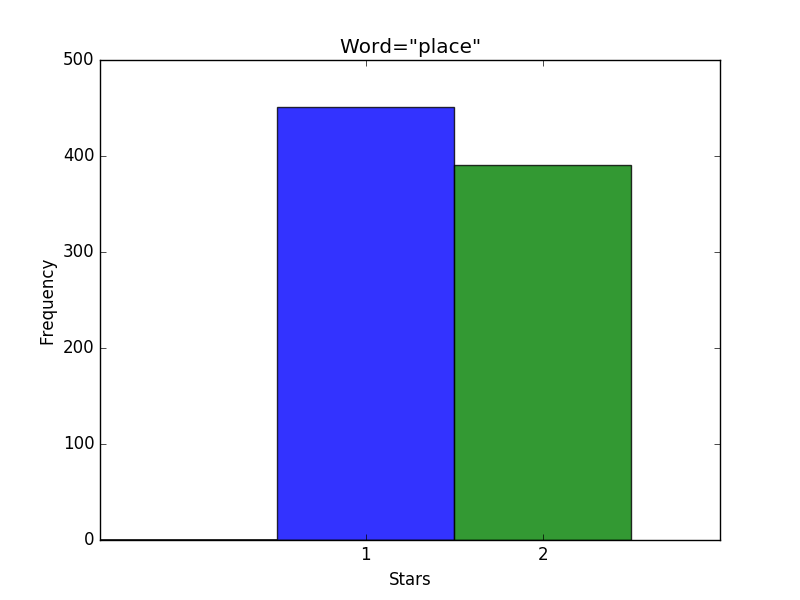
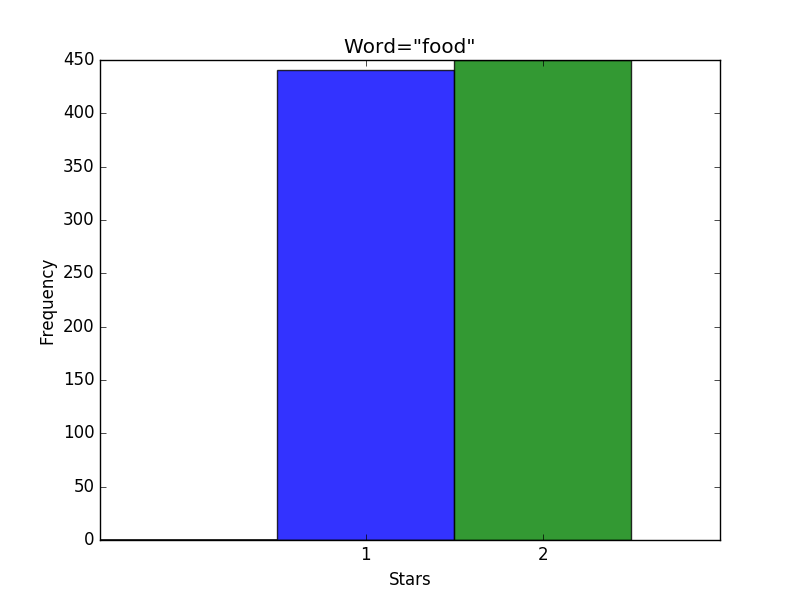
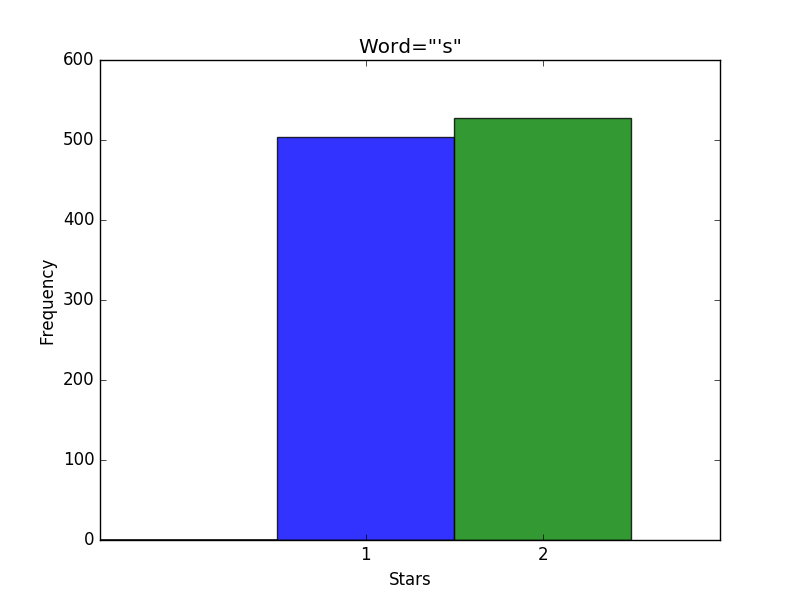
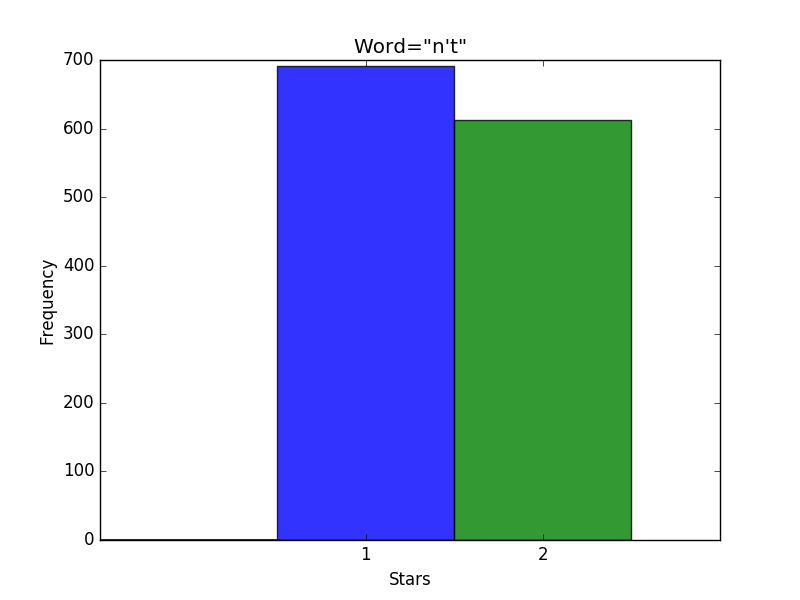
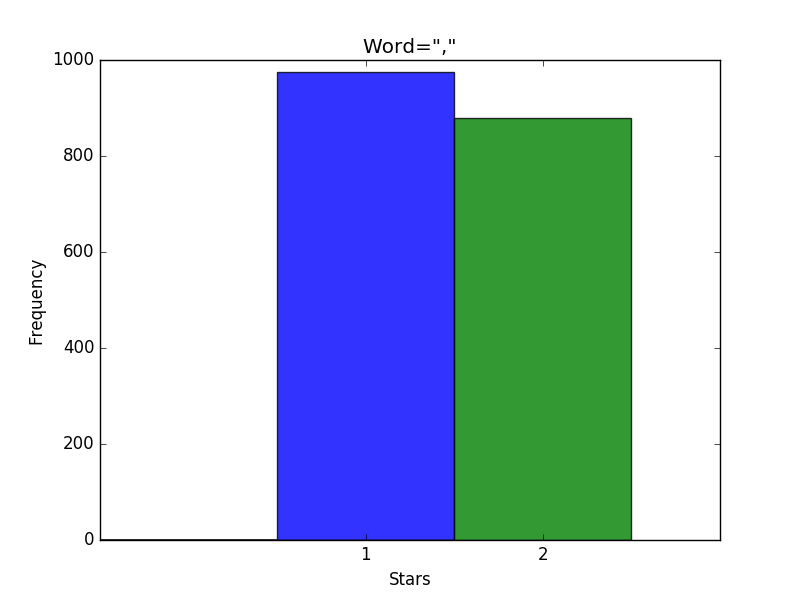
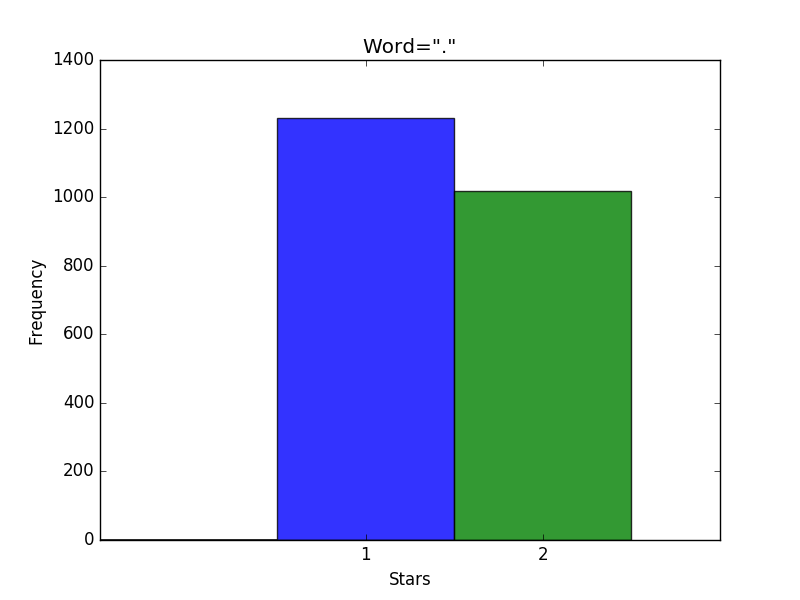
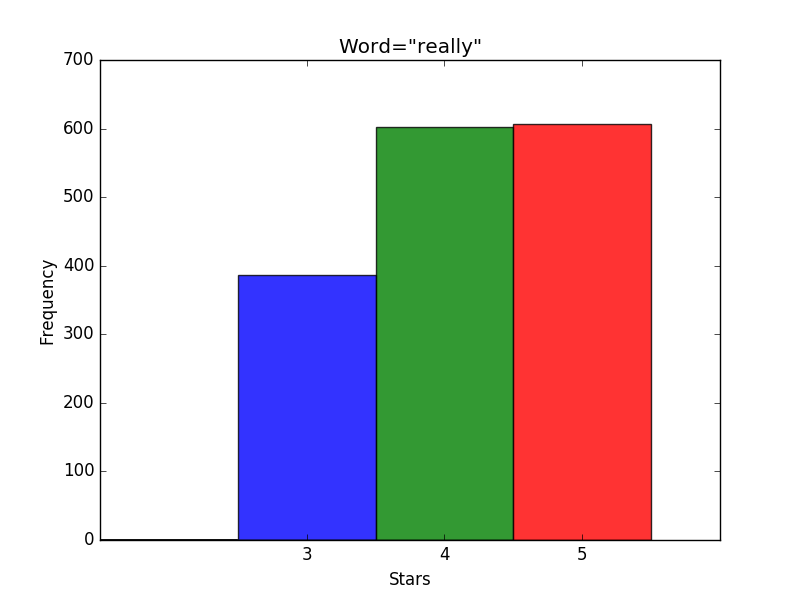
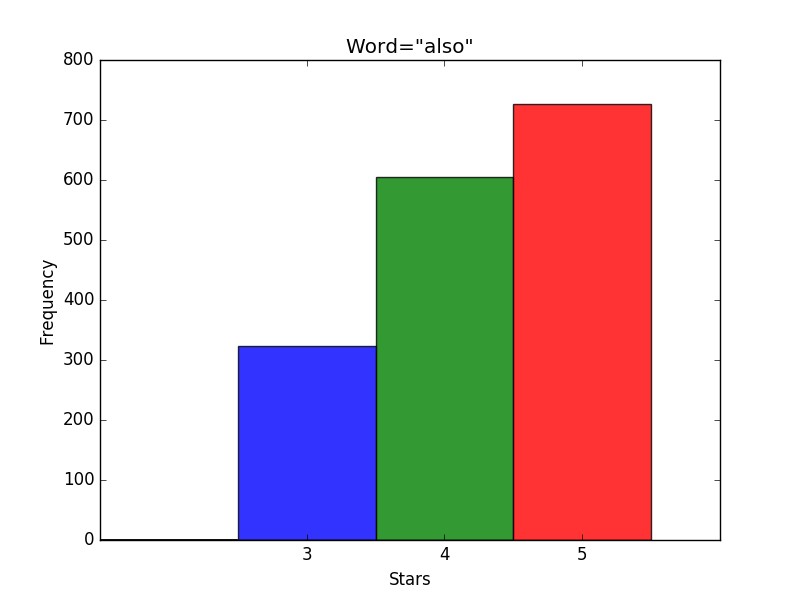
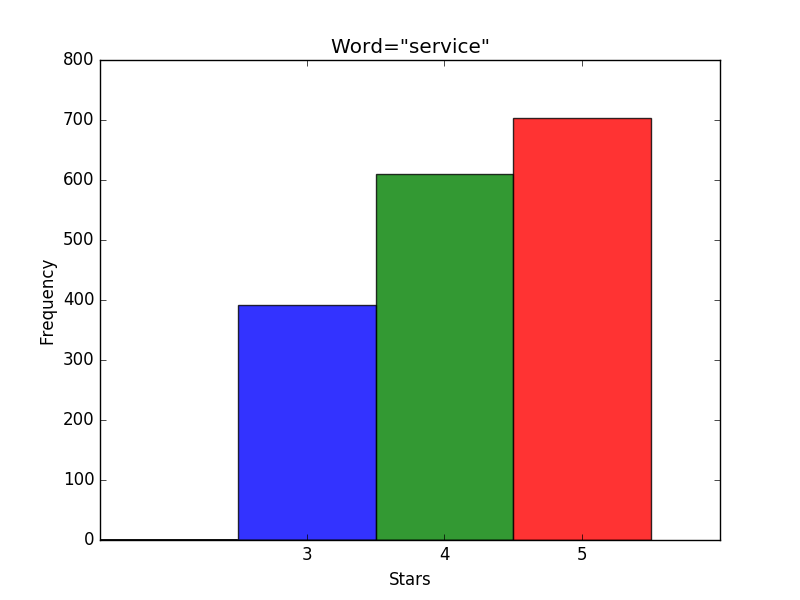
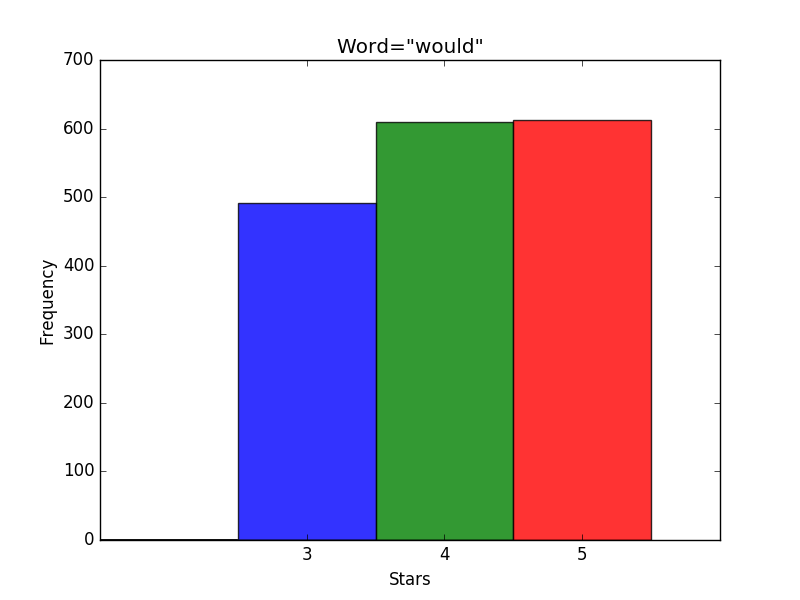
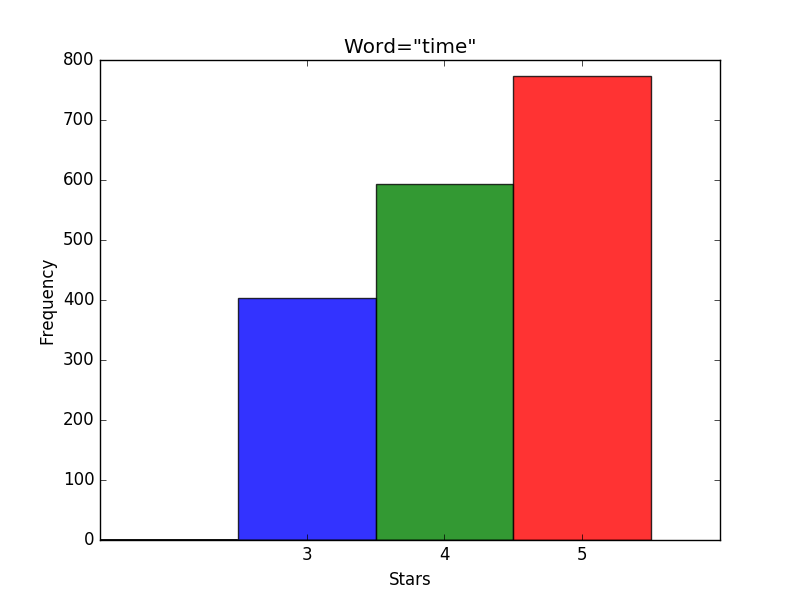
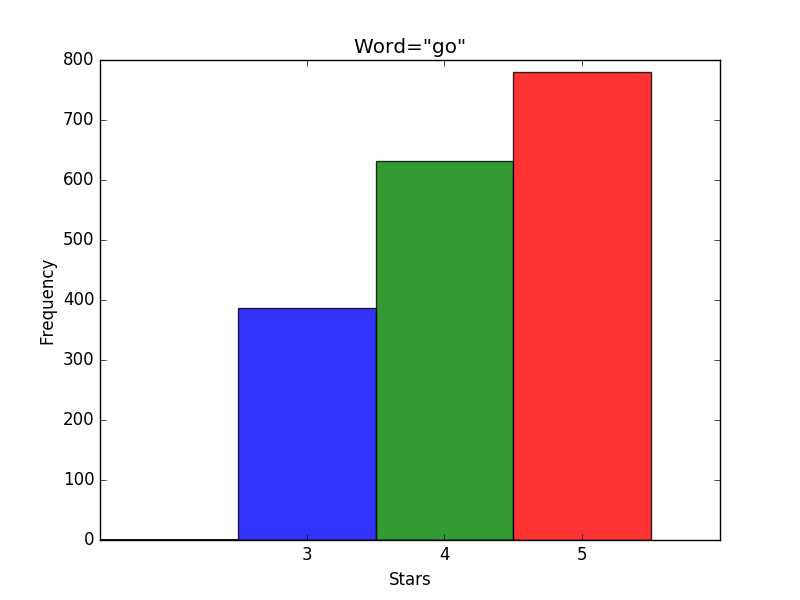
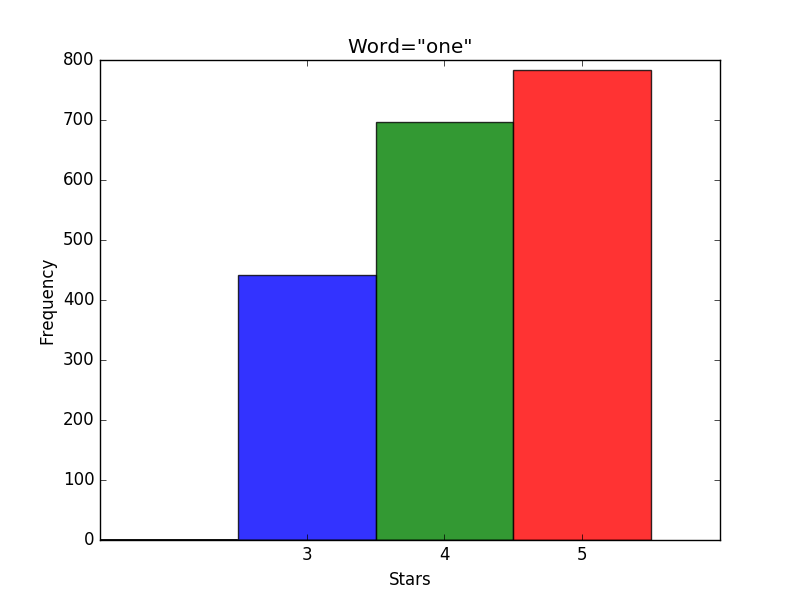
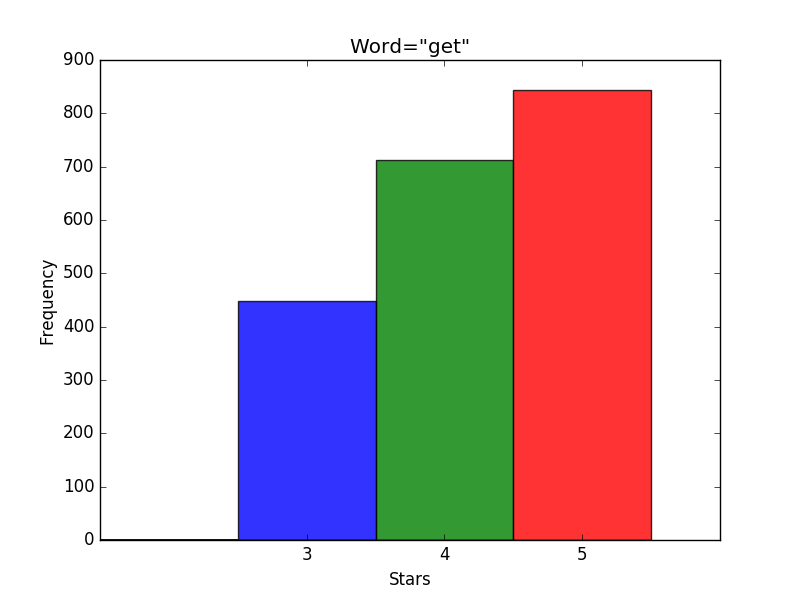
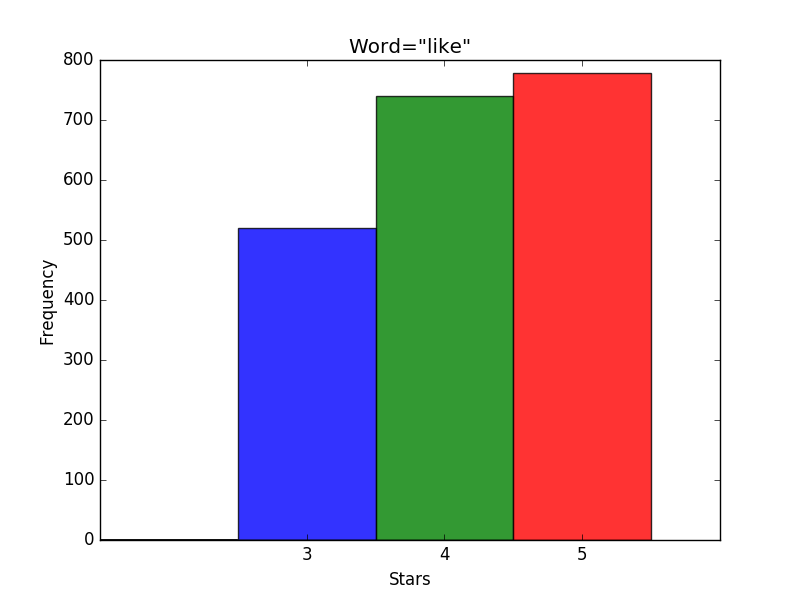
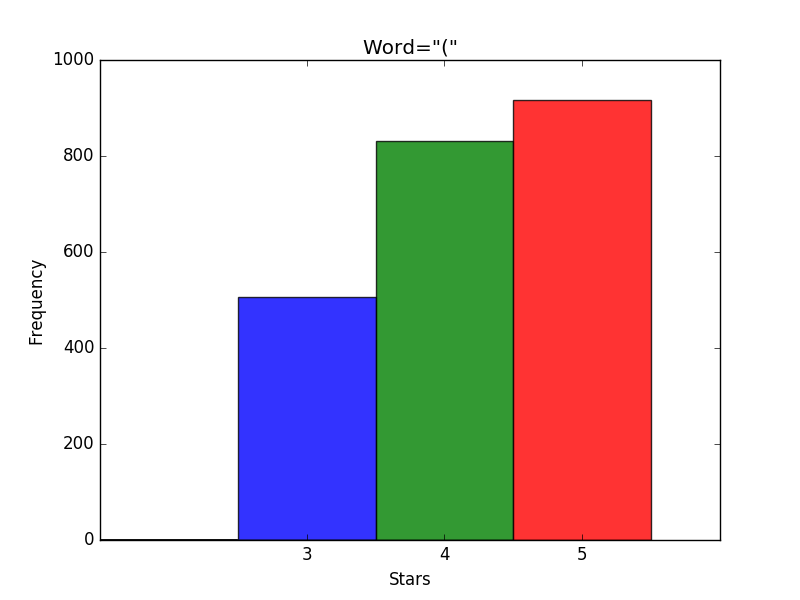
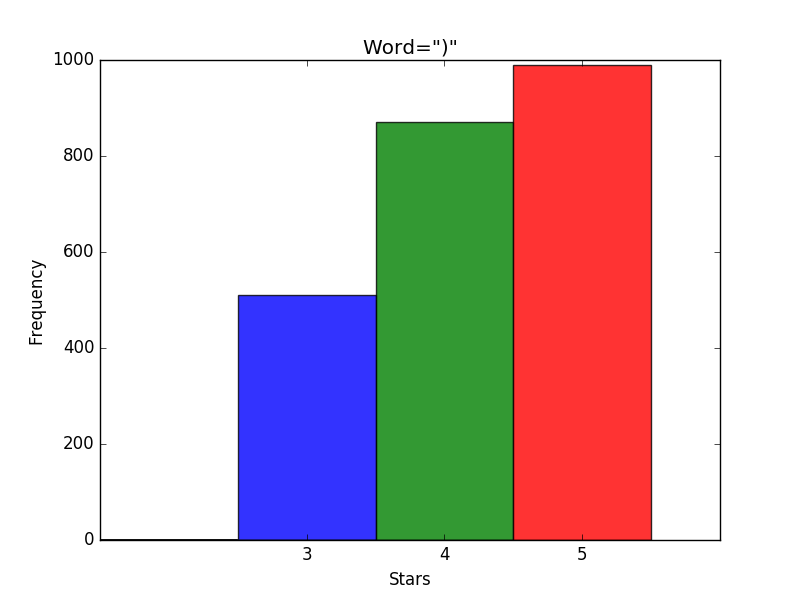
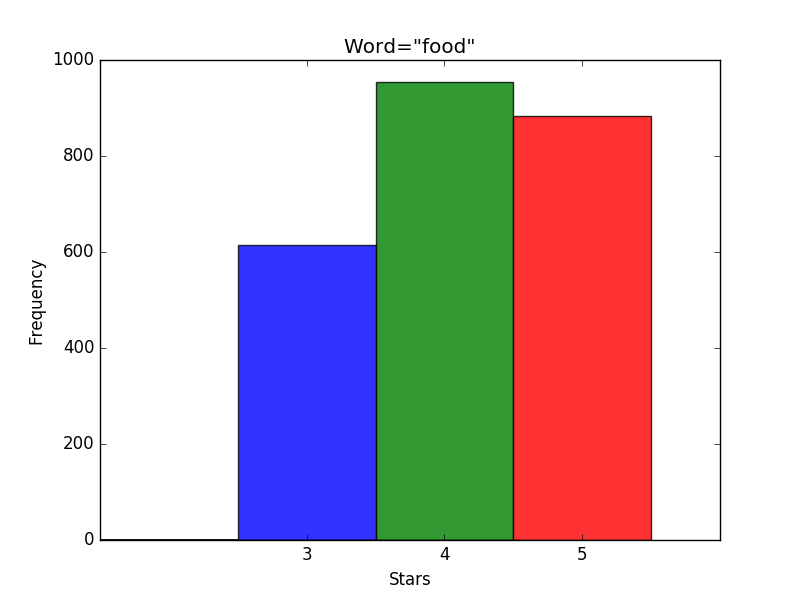
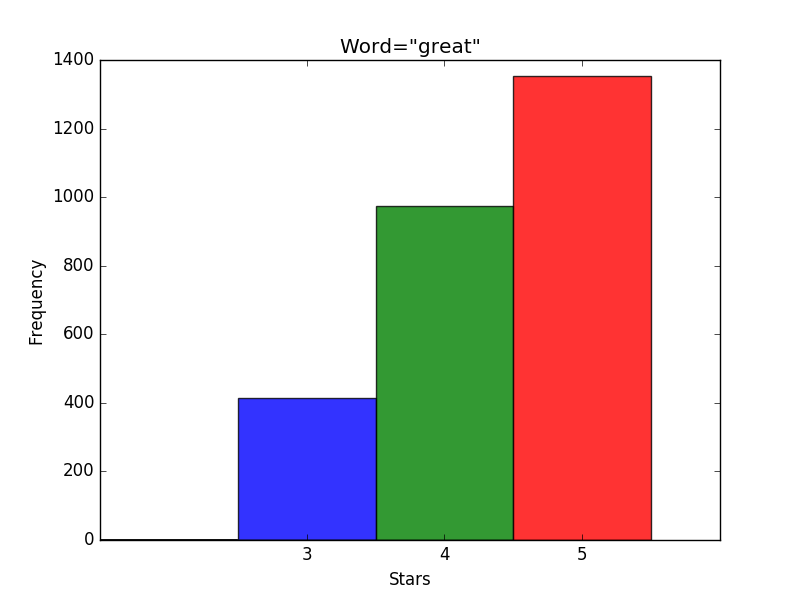
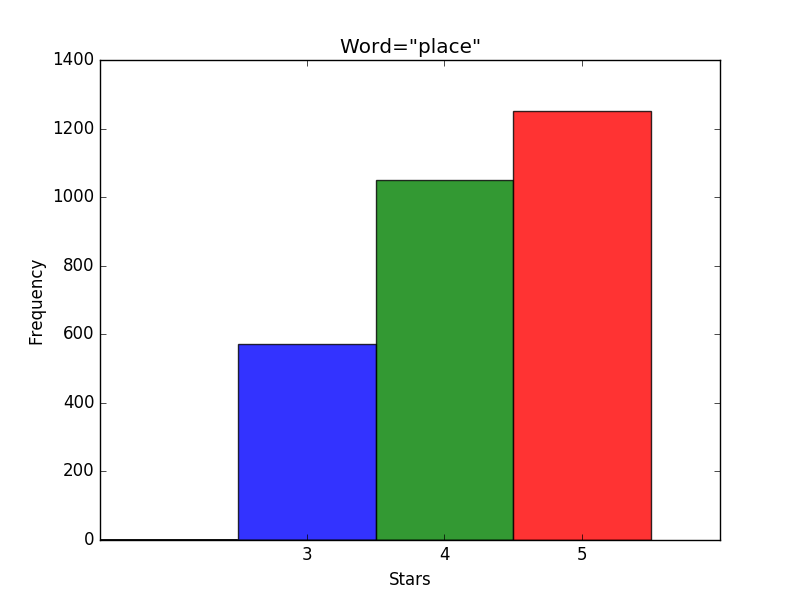
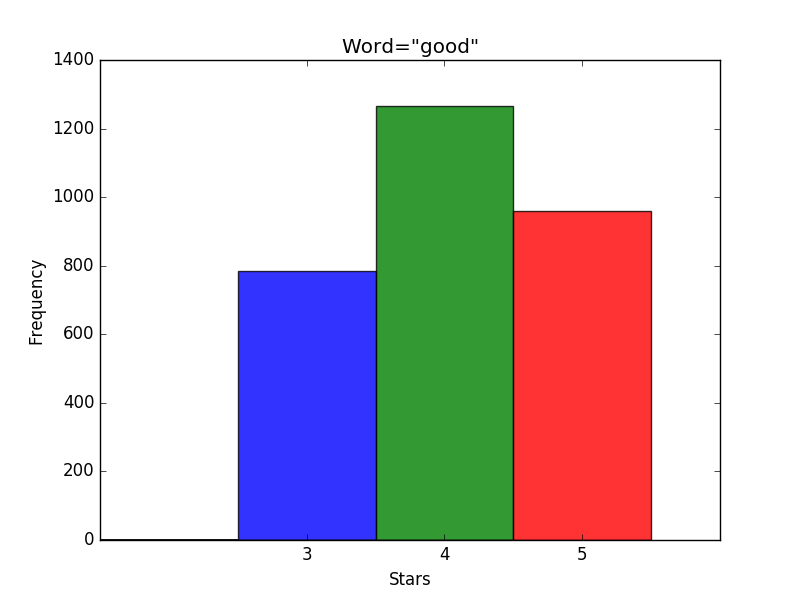
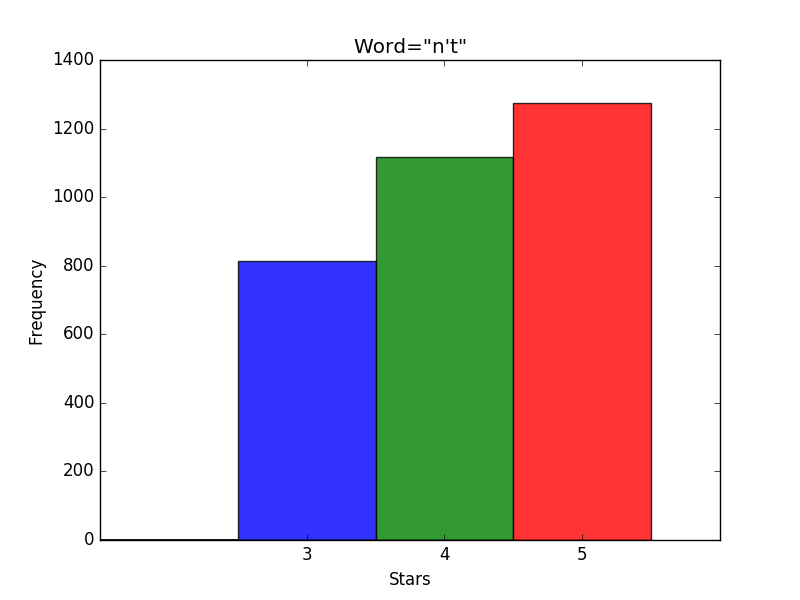
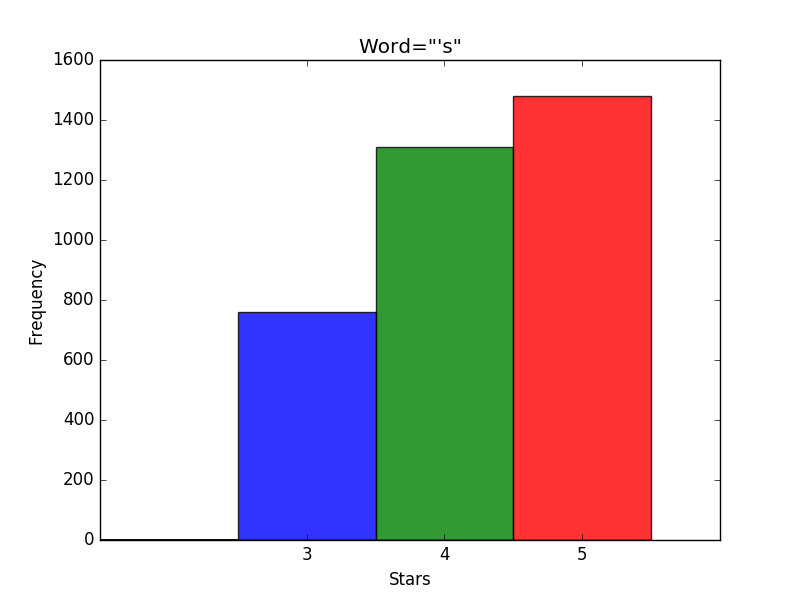
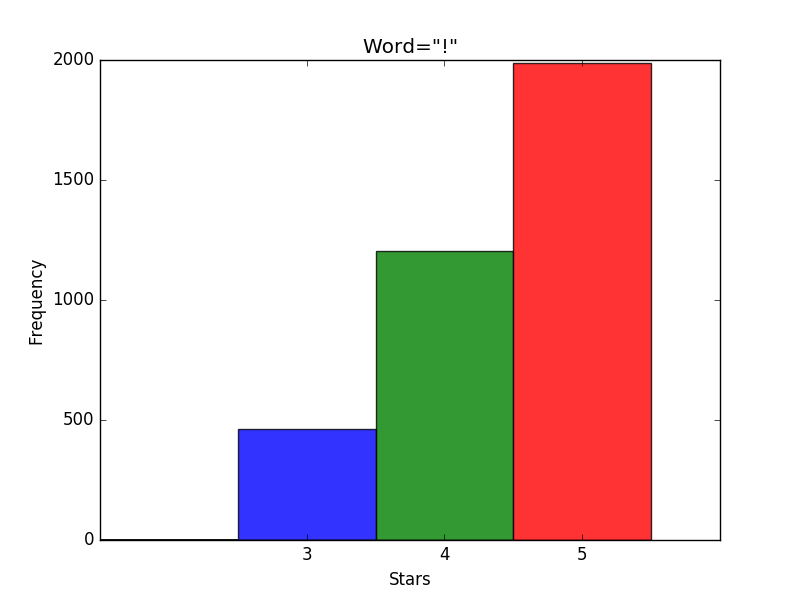
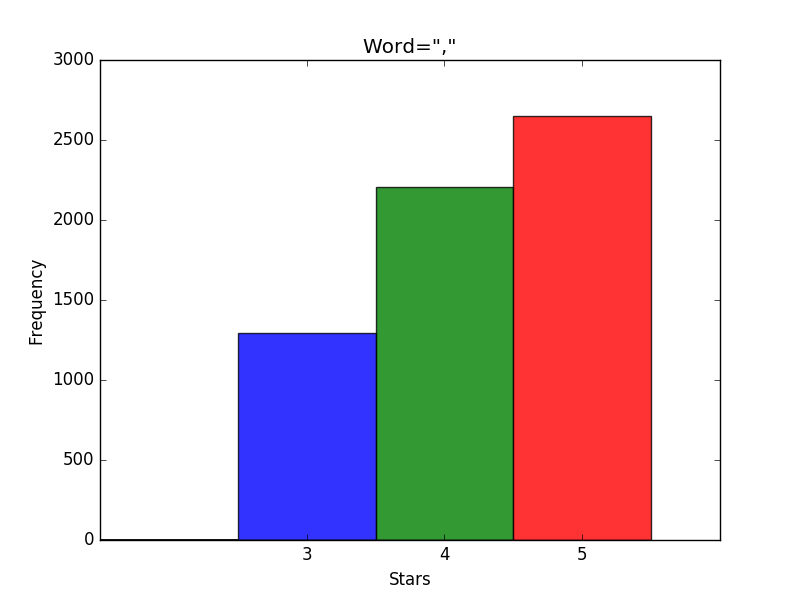
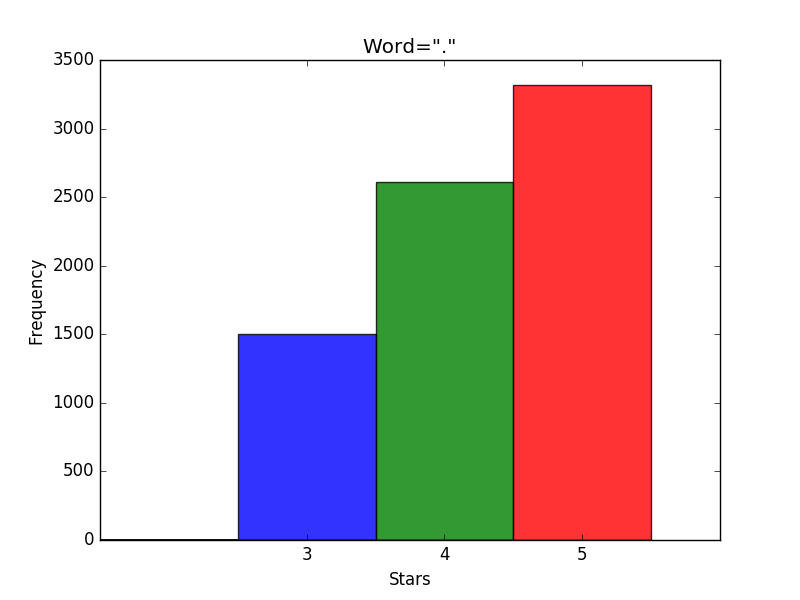
time 639 [359, 280]

go 608 [318, 290]

even 584 [345, 239]

'' 542 [312, 230]

Los gráficos obtenidos:



Con 100.000 datos como dataset, se obtuvo las siguientes 20 palabras más ocupadas sin stopwords (en inglés), para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos), con sus estrellas respectivas (la interpretación es para los 3+ estrellas, el primer número del arreglo es la cantidad de veces que se dice en reviews con 3 estrellas, el siguiente es lo mismo pero para 4 estrellas y el ultimo lo mismo pero con 5 estrellas):

Top 20 words for 3+ Stars ratings.

. 74280 [12648, 26242, 35390]

, 59806 [10783, 21549, 27474]

! 38032 [3799, 12392, 21841]

's 30781 [6100, 11536, 13145]

n't 29526 [6581, 10823, 12122]

great 28938 [3451, 10546, 14941]

good 28665 [6468, 12219, 9978]

place 28494 [4901, 10555, 13038]

food 26138 [5386, 10136, 10616]

) 21851 [4072, 8391, 9388]

( 20158 [3975, 7852, 8331]

service 19908 [3606, 6794, 9508]

like 19759 [4361, 7681, 7717]

get 18518 [3594, 6839, 8085]

time 18087 [3215, 6281, 8591]

one 17694 [3367, 6648, 7679]

go 17304 [3331, 6234, 7739]

back 15747 [2939, 5676, 7132]

would 15578 [3754, 5692, 6132]

really 15155 [3210, 6053, 5892]

Top 20 words for 2- Stars ratings.

. 21504 [12608, 8896]

, 17268 [9834, 7434]

n't 12349 [7147, 5202]

's 8729 [4714, 4015]

would 7627 [4663, 2964]

! 7597 [5044, 2553]

place 7477 [4209, 3268]

food 7456 [3655, 3801]

like 6958 [3748, 3210]

one 6922 [4099, 2823]

get 6767 [4019, 2748]

service 6641 [3892, 2749]

back 6590 [4105, 2485]

time 6432 [3895, 2537]

( 6221 [3407, 2814]

) 6160 [3359, 2801]

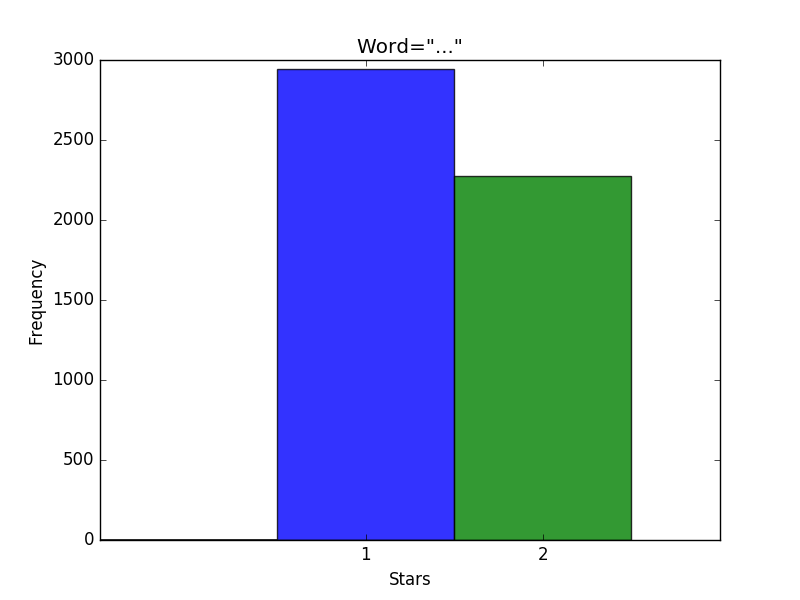
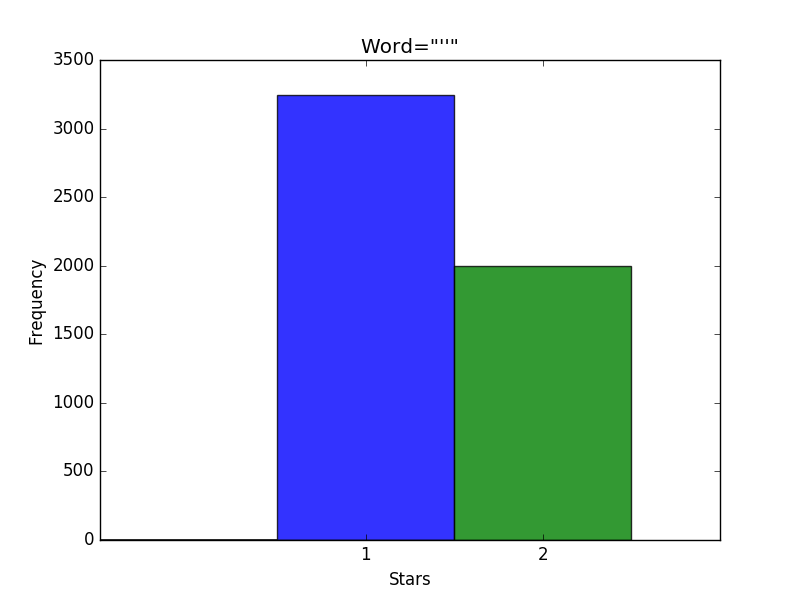
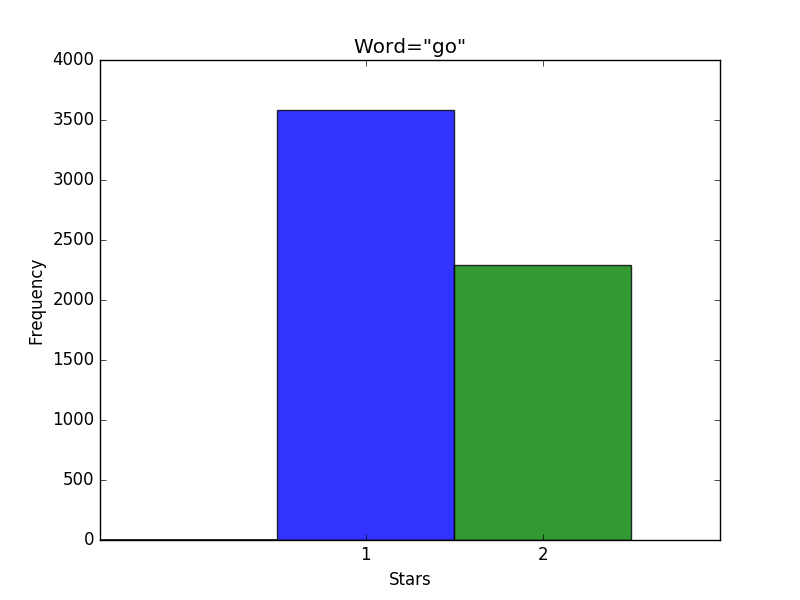
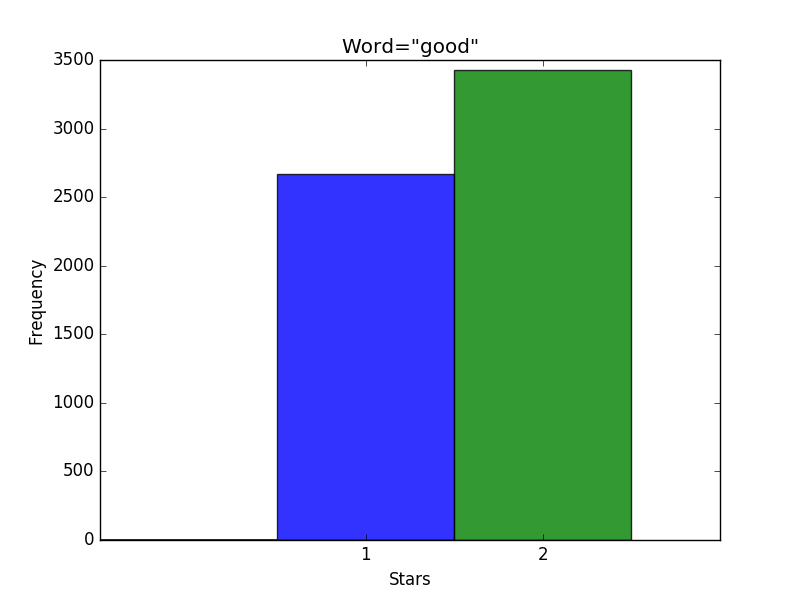
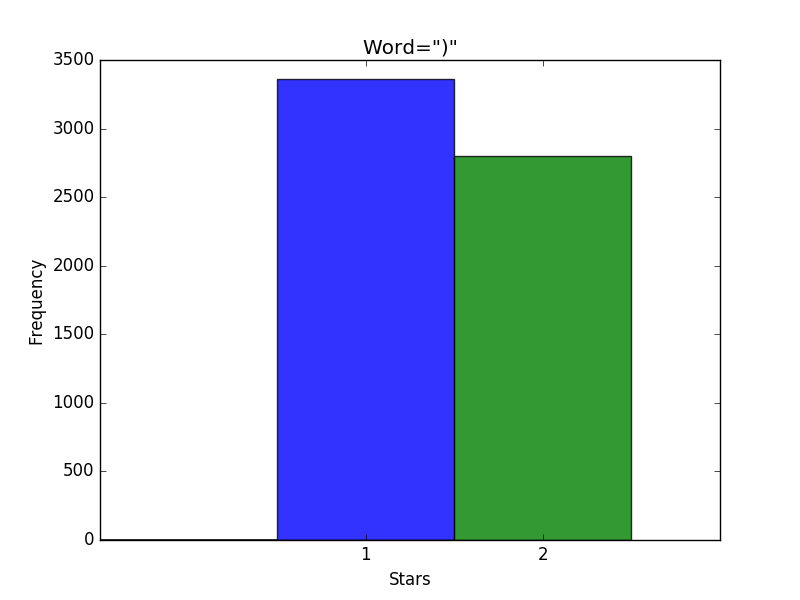
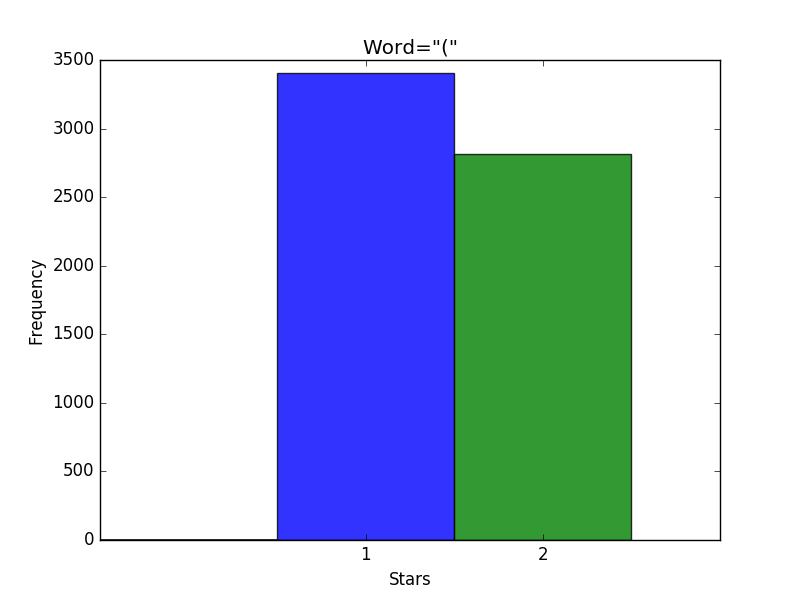
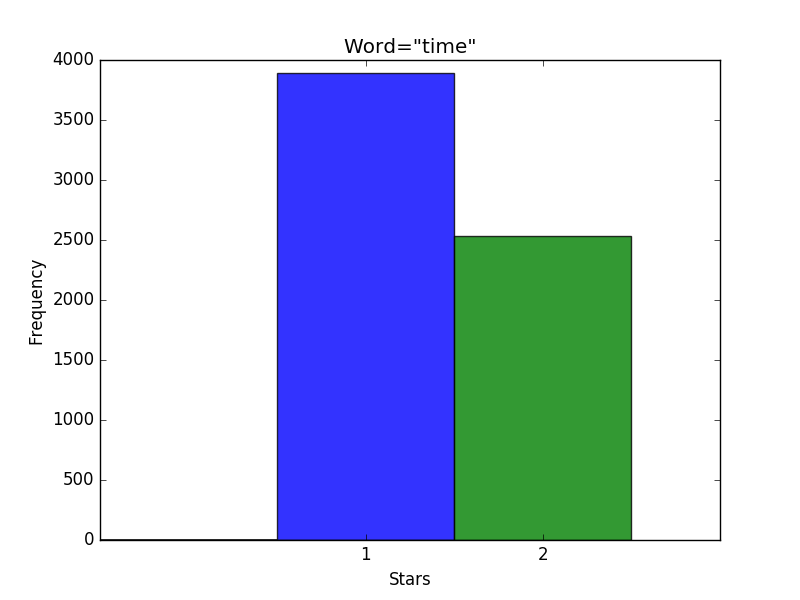
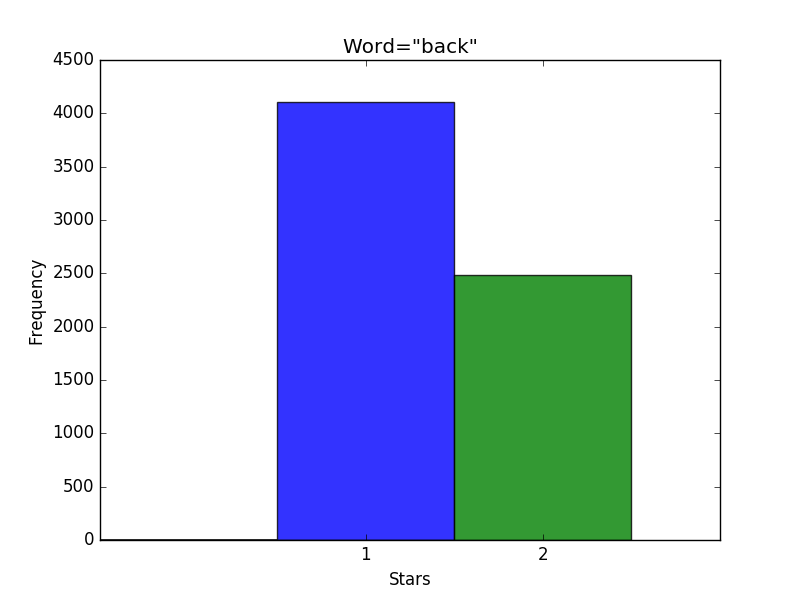
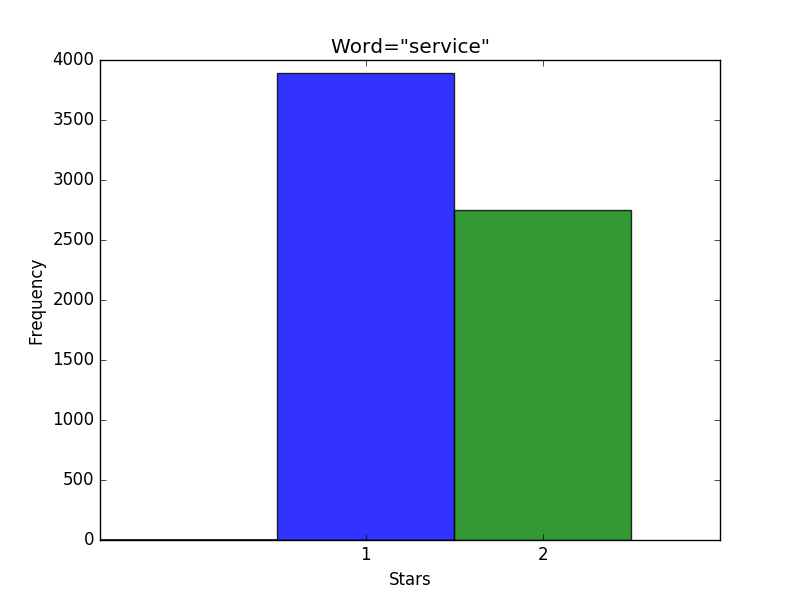
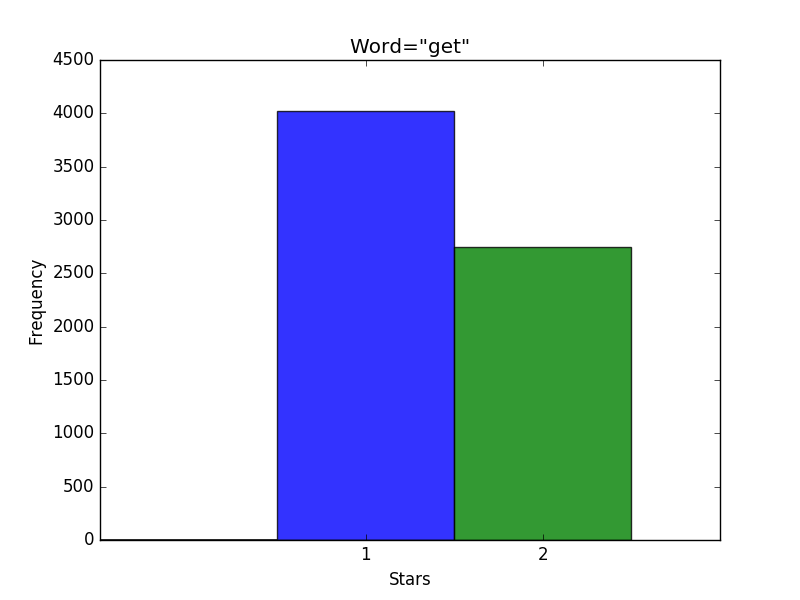
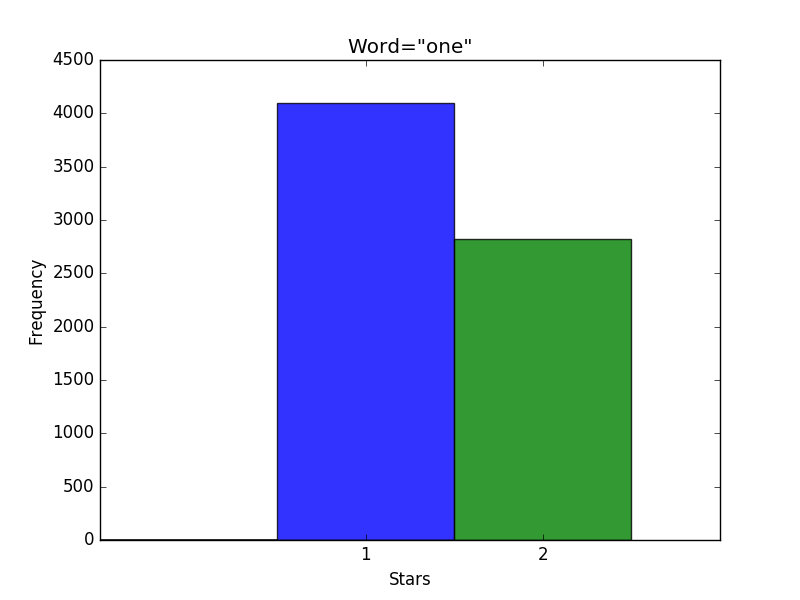
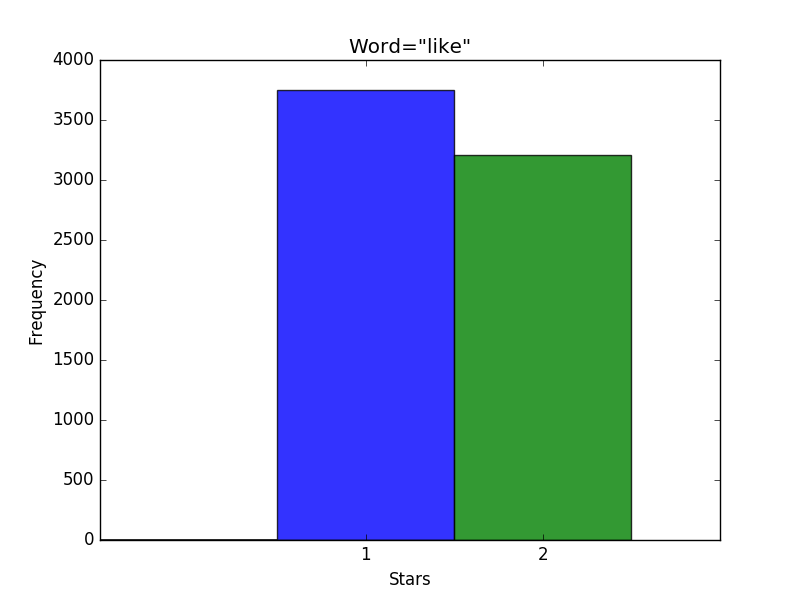
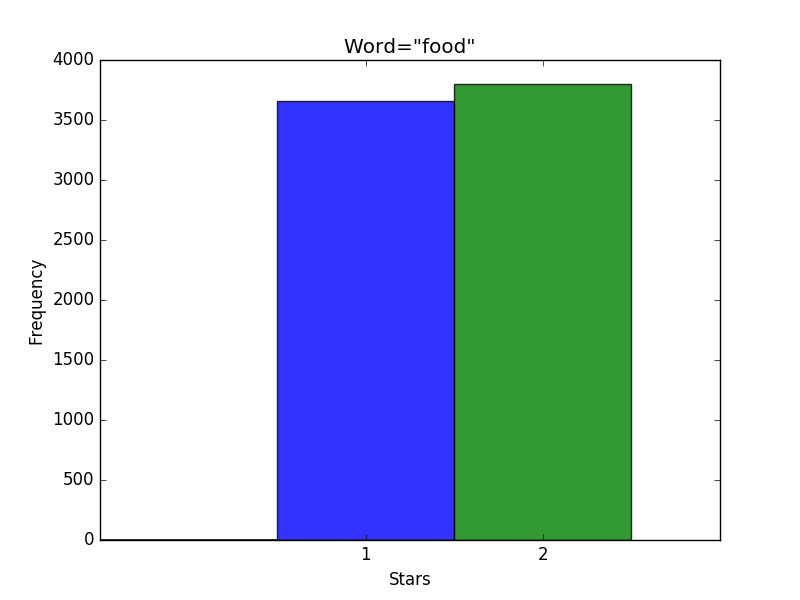
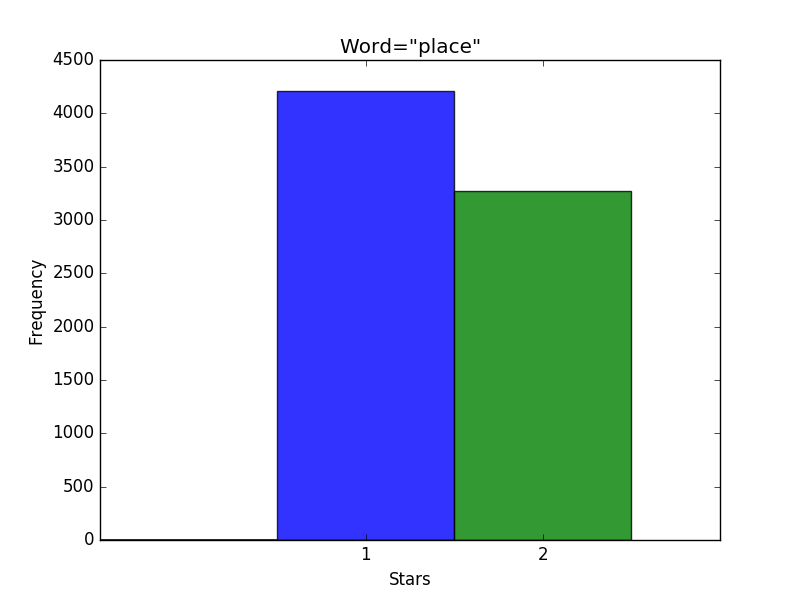
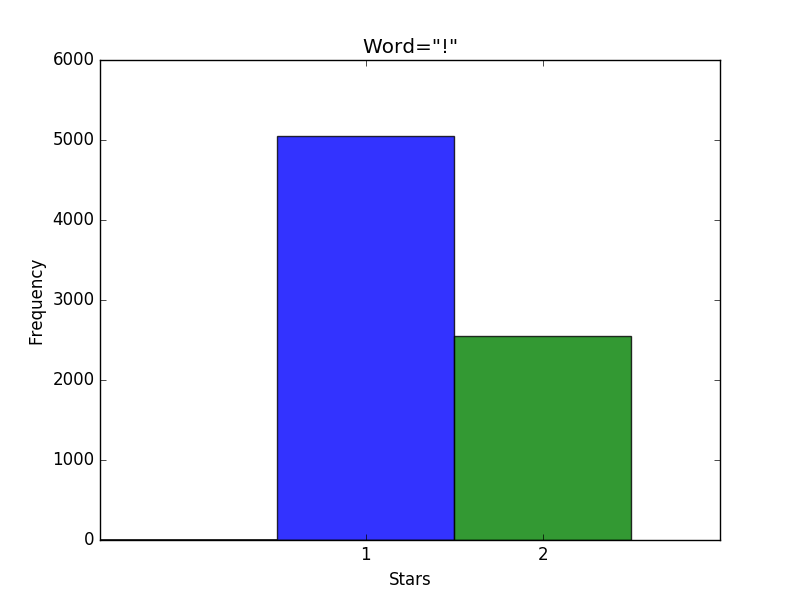
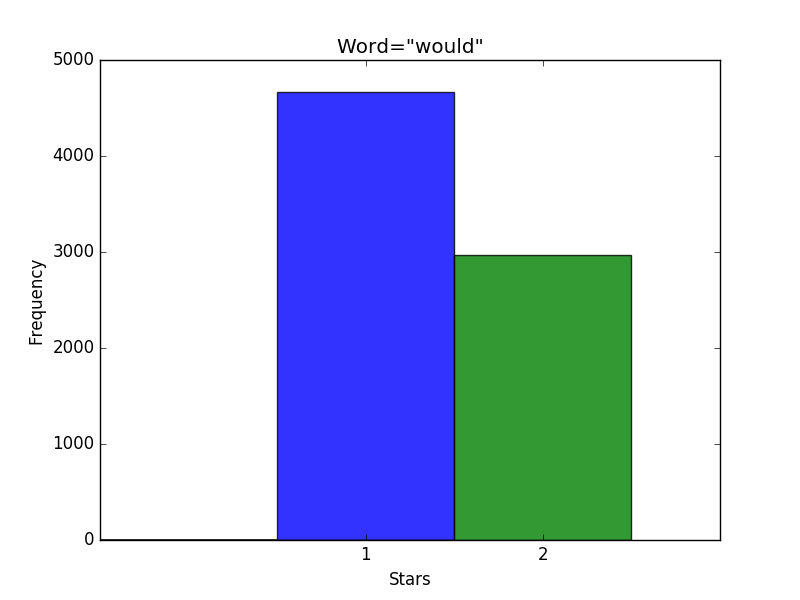
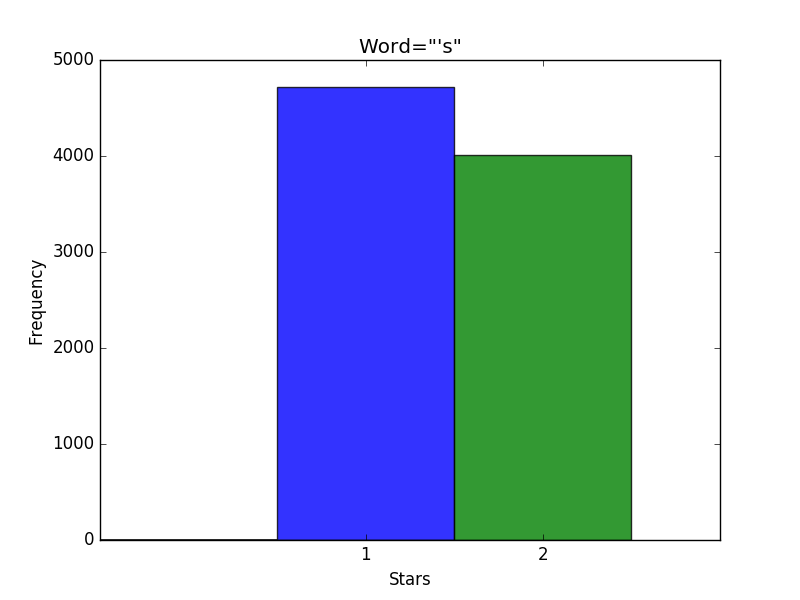
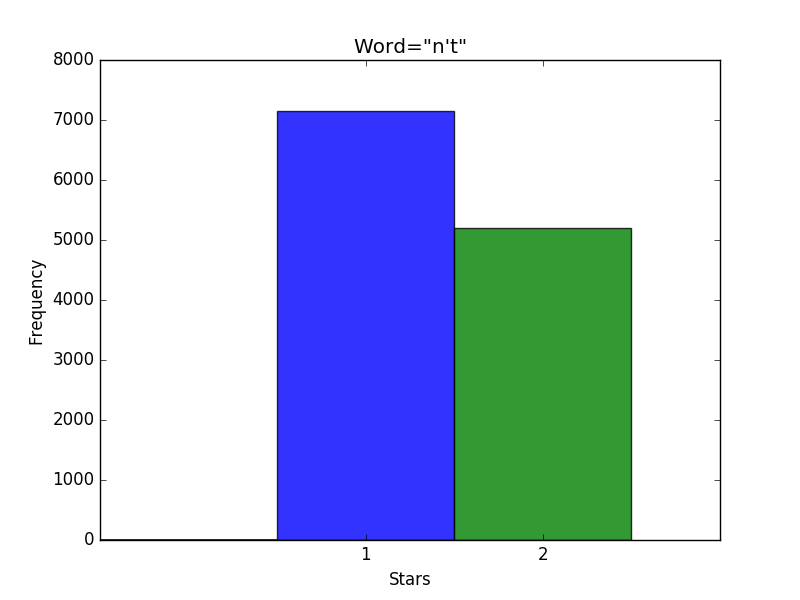
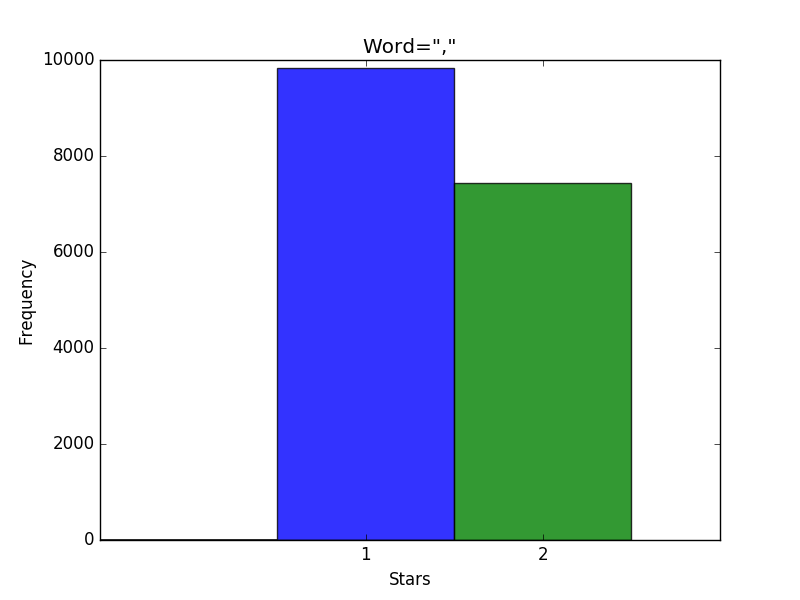
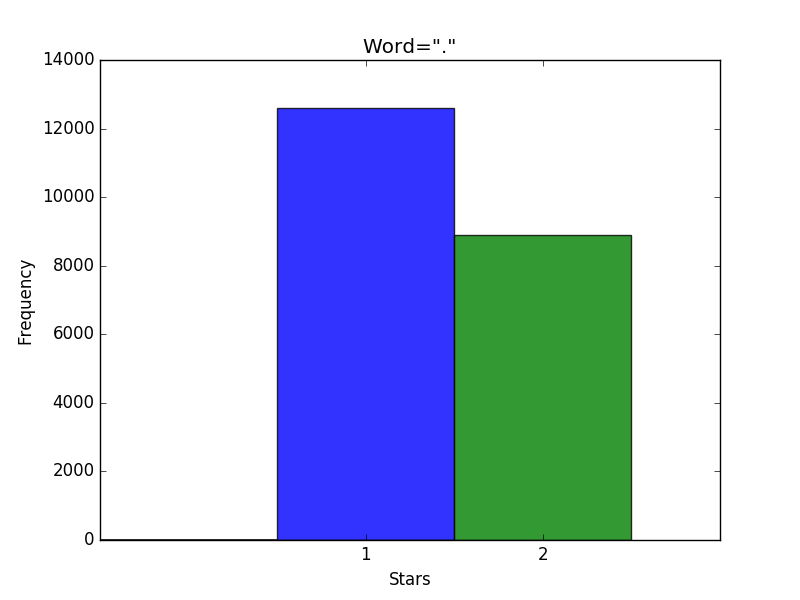
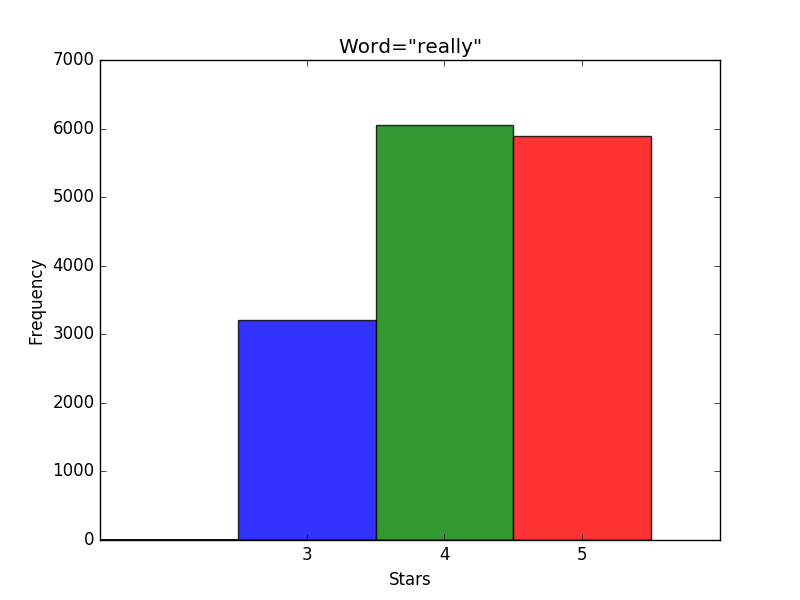
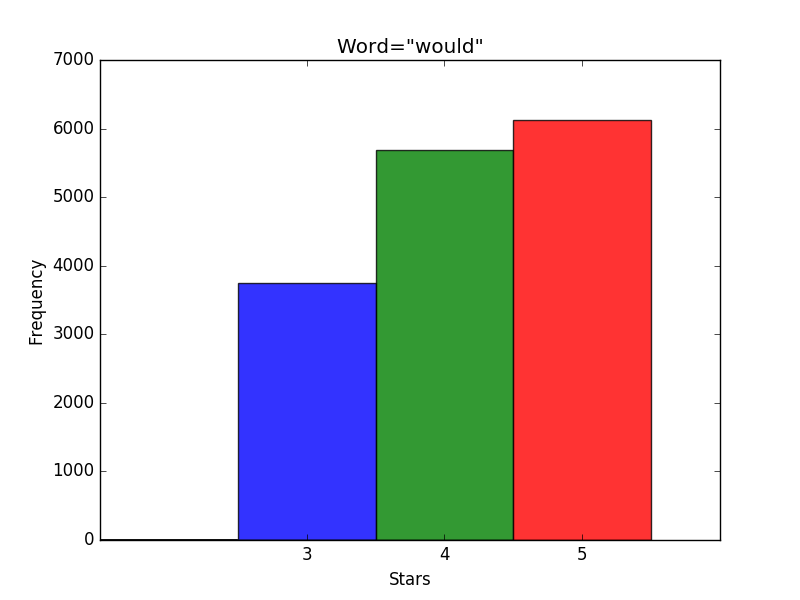
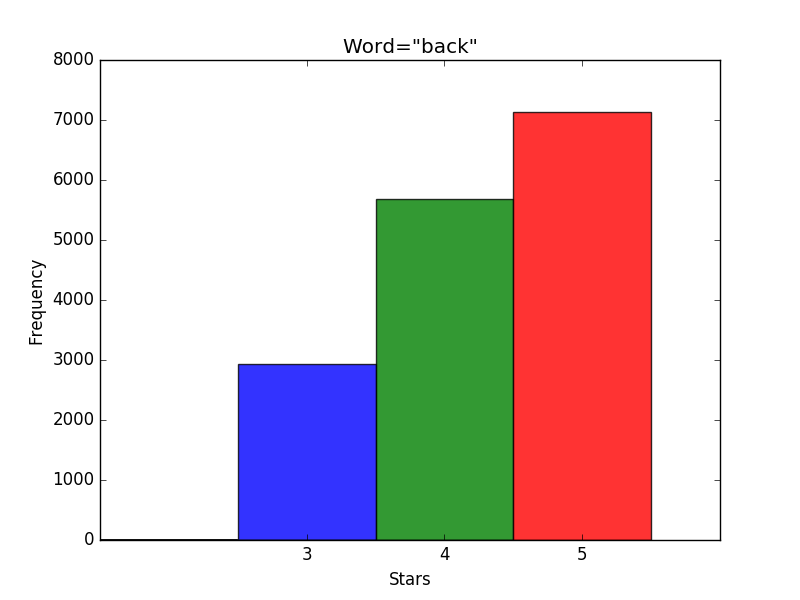
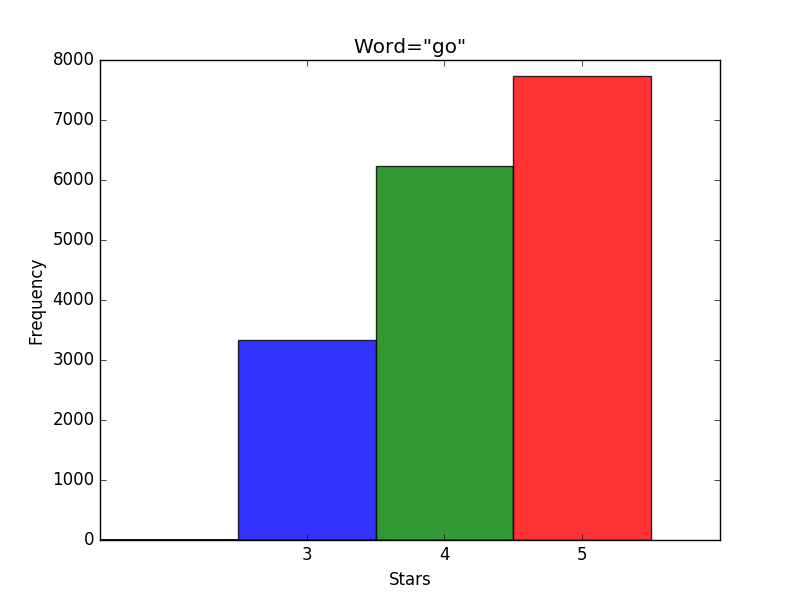
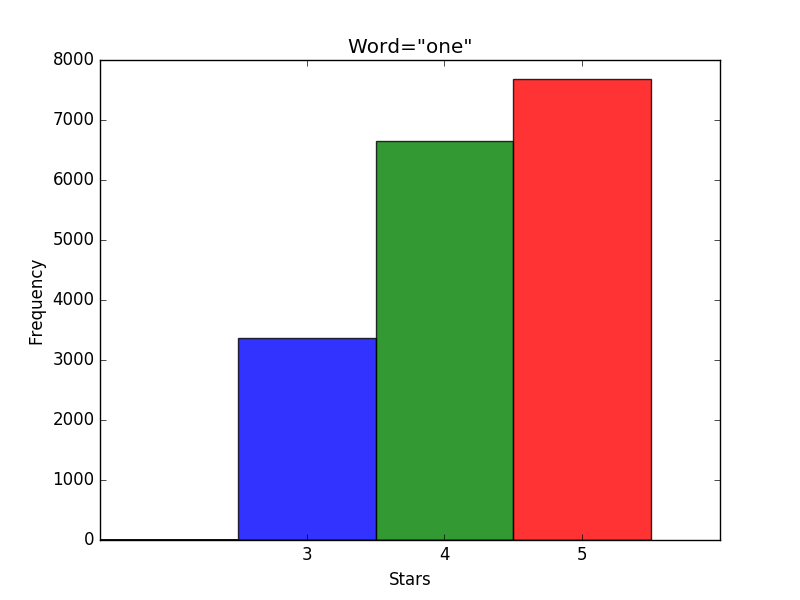
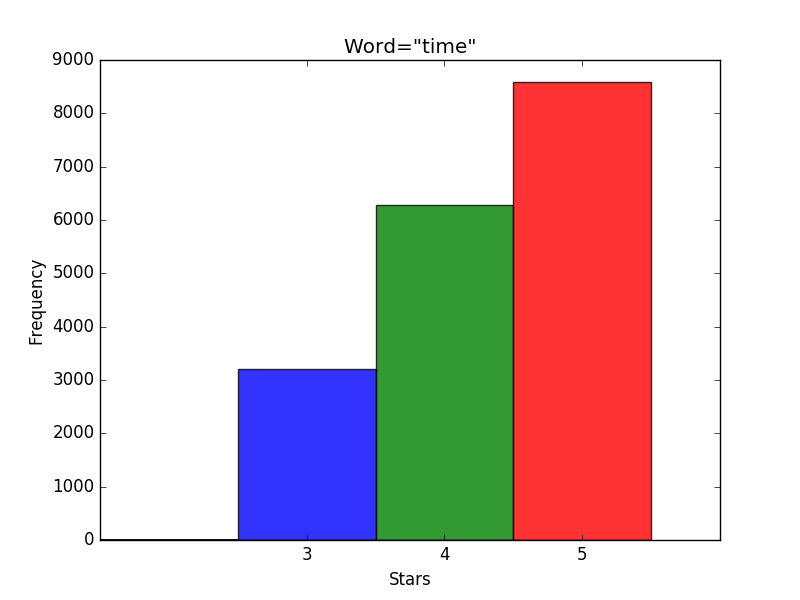
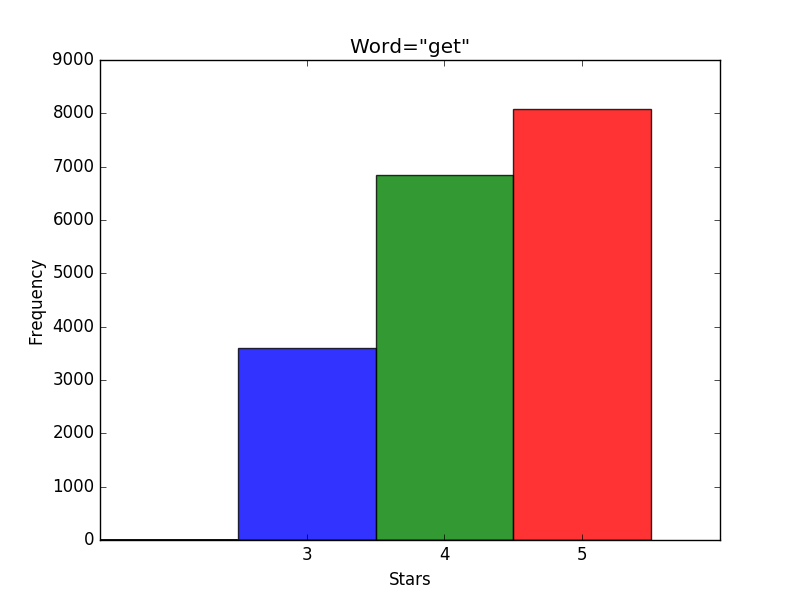
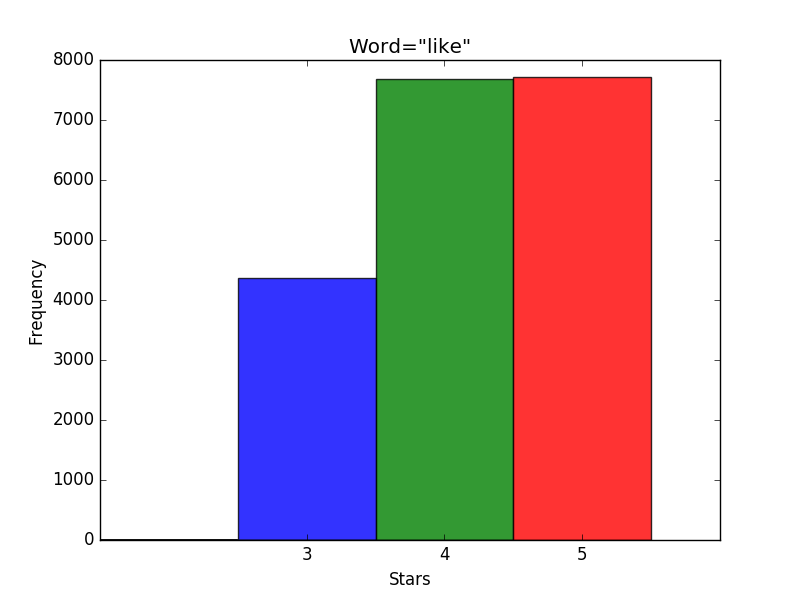
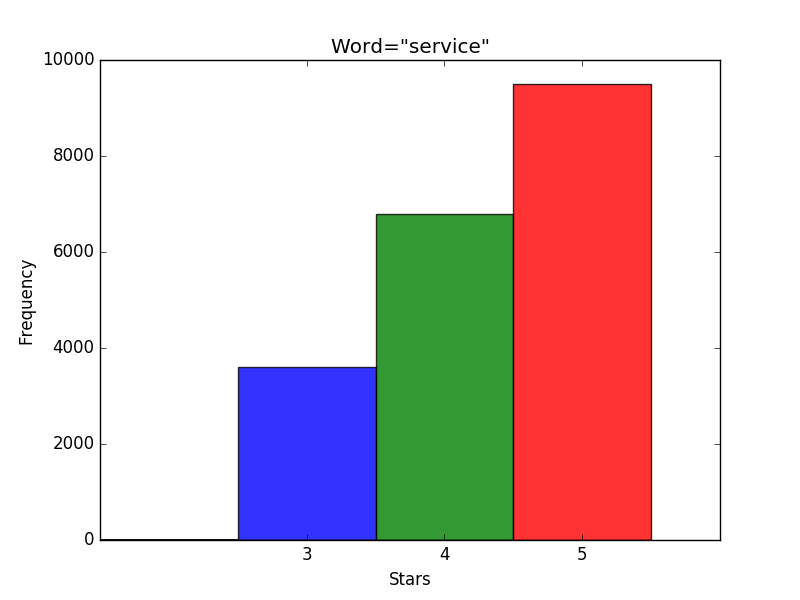
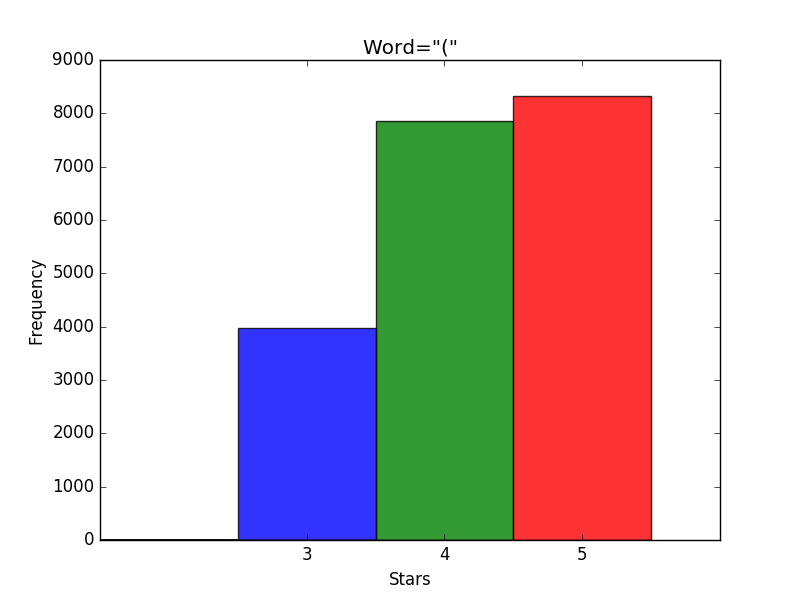
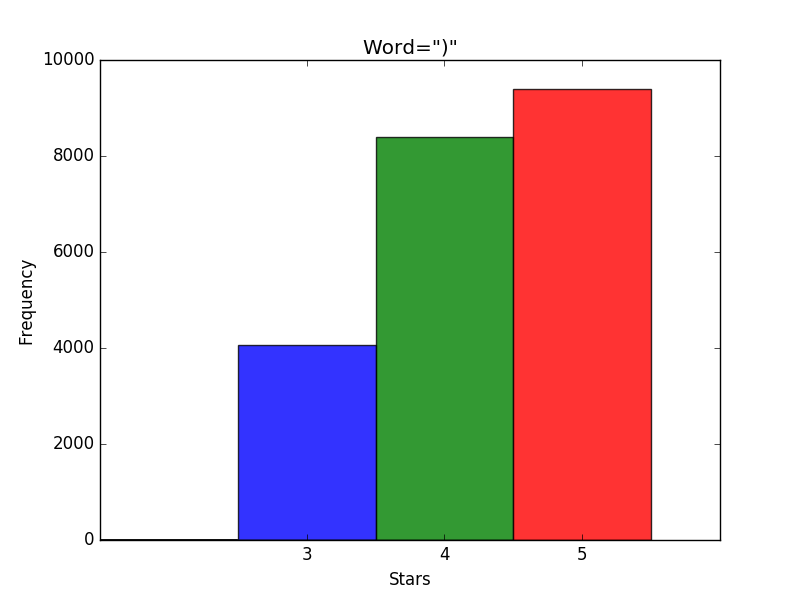
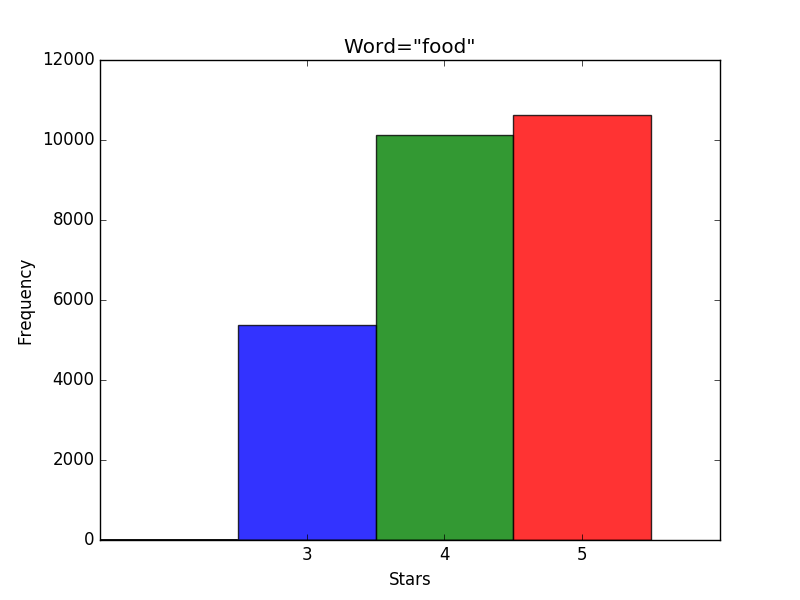
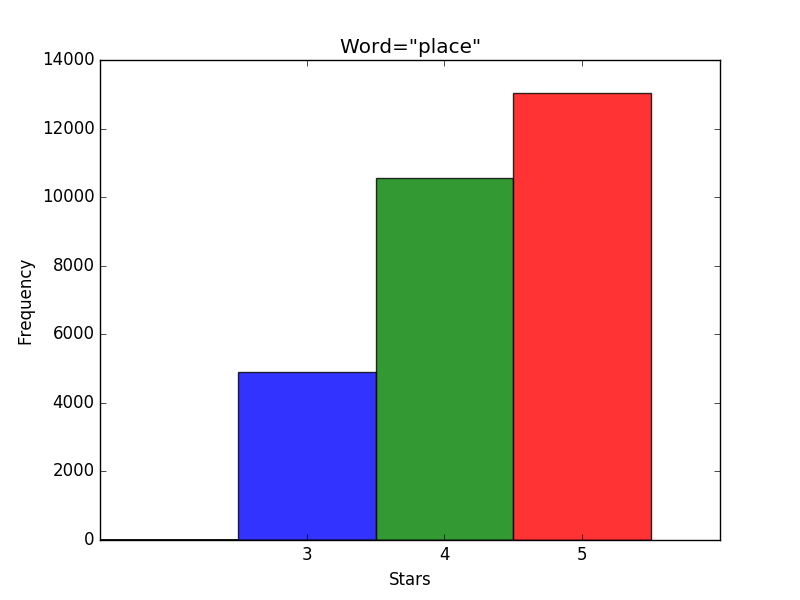
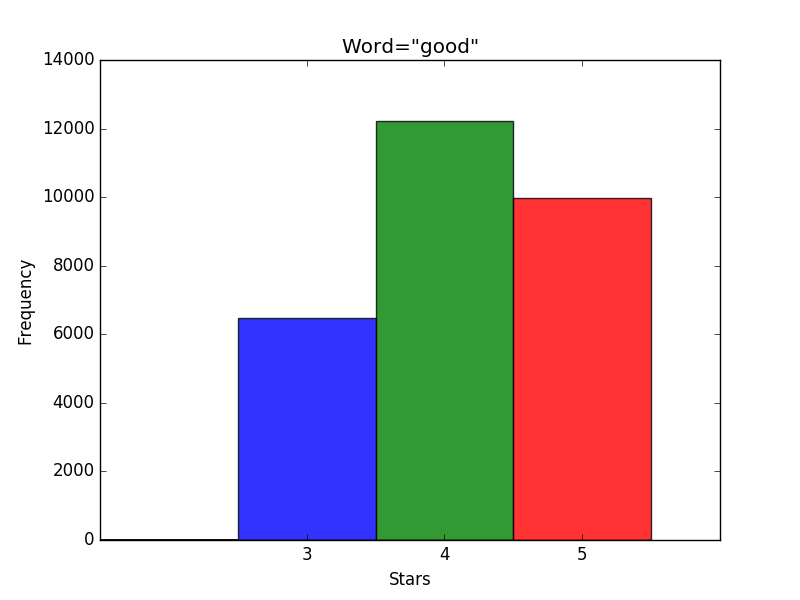
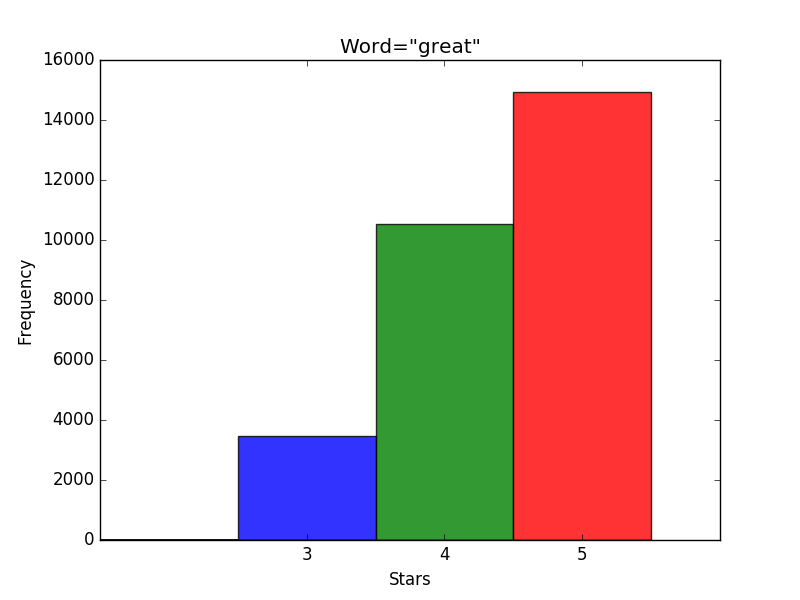
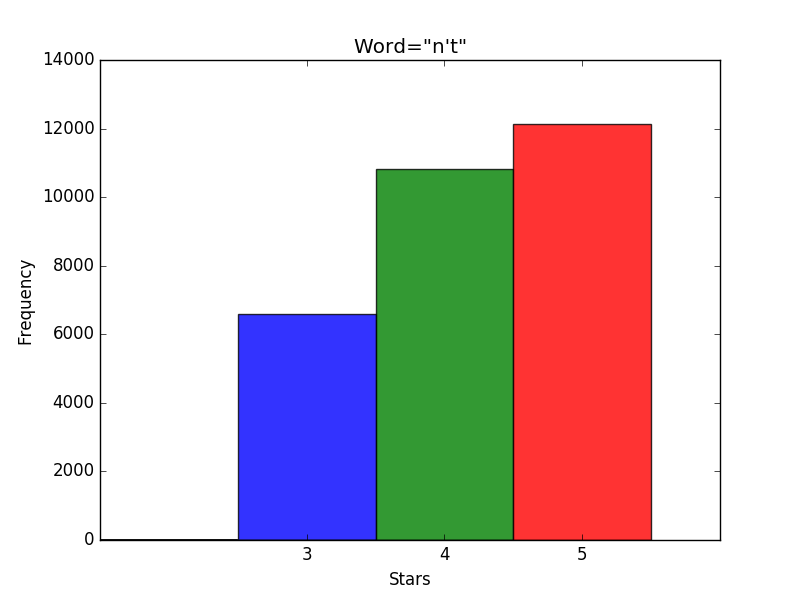
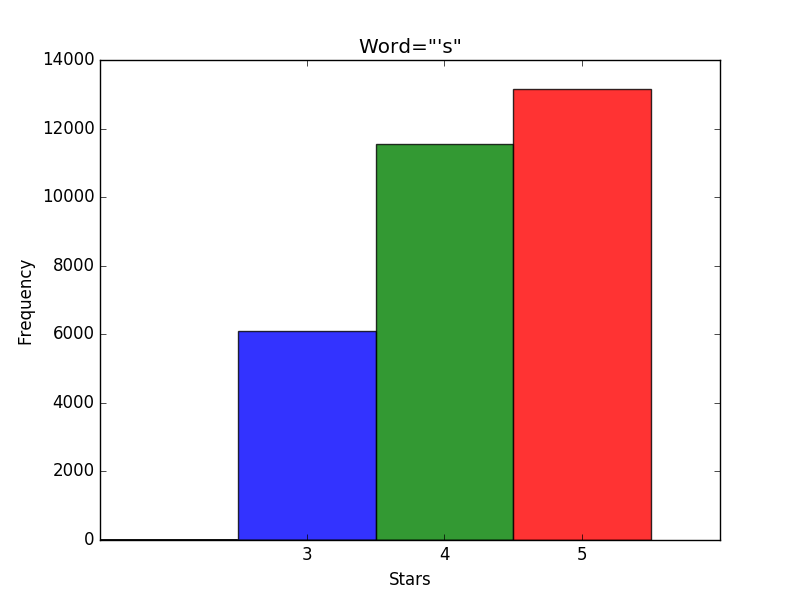
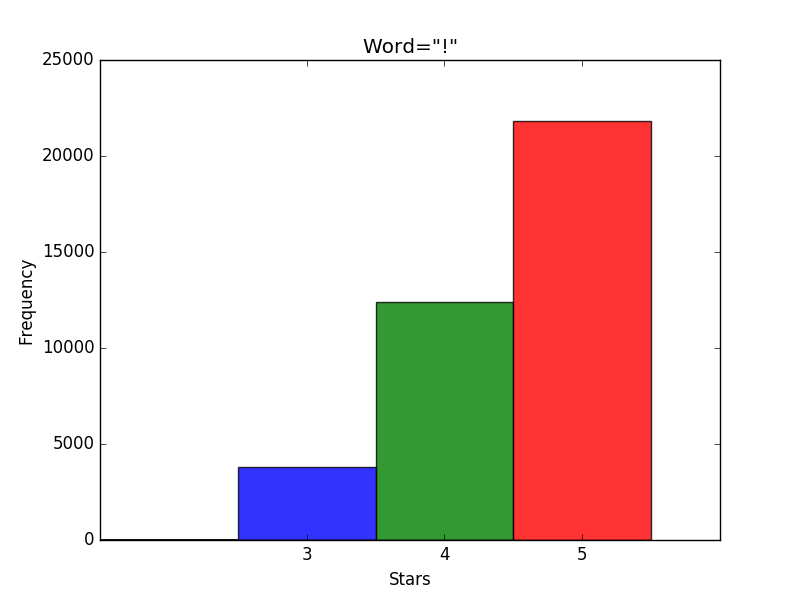
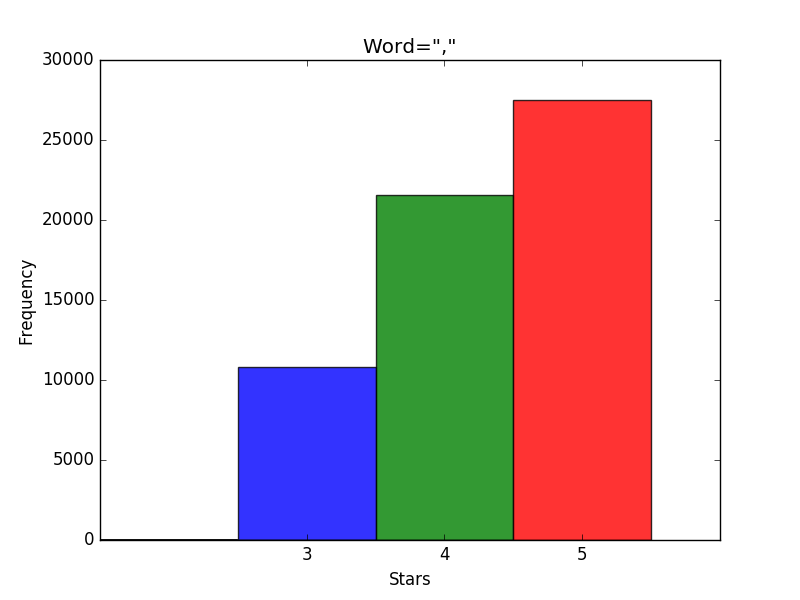
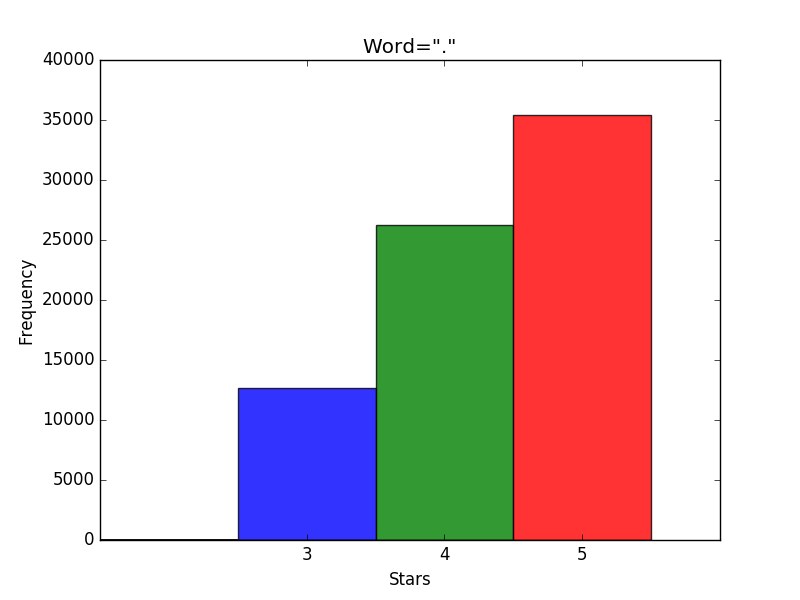
good 6091 [2667, 3424]

go 5874 [3586, 2288]

'' 5244 [3248, 1996]

... 5218 [2942, 2276]

Los gráficos obtenidos:



Con todo el dataset, se obtuvo las siguientes 20 palabras más ocupadas sin stopwords (en inglés), para cada clasificación (3 estrellas en el rating o más y 2 estrellas en el rating o menos), con sus estrellas respectivas (la interpretación es para los 3+ estrellas, el primer número del arreglo es la cantidad de veces que se dice en reviews con 3 estrellas, el siguiente es lo mismo pero para 4 estrellas y el ultimo lo mismo pero con 5 estrellas):

No se pudo.

Tiempo Dataset 10.000: 330.780605058

Tiempo Dataset 100.000: 3122.42692919

Estimación Tiempo Dataset completo: 23,415 horas

Tiempo Dataset completo: No se pudo.

La estimación del dataset completo nos muestra que, ya que luego de tokenizar las palabras en un arreglo (sin repetidas) y luego buscar las palabras dentro de ese arreglo que sean stopwords es O(n) [para todas las palabras en mi arreglo] \* O(n) [buscar dentro del arreglo de stopwords] o sea esta parte es O(n^2) y considerando que esto es para cada línea entonces el algoritmo finalmente es de O(n^3), lo cual explica porque se demora tanto.

Y debido a que se demora tanto, no podre calcularlo antes de la entrega propia de la tarea.

Como comentarios de este ejercicio, se puede ver que si bien los top 20 palabras más usadas están ensuciadas con símbolos de puntuación, las palabras que se pueden analizar es que por ejemplo en la categoría de 3+ estrellas en general se ocupaba el término “great” para indicar que era muy bueno, por ende se ocupaba más en reviews con 5 estrellas, mientras que “good” se ocupaba más en reviews con 4 estrellas. También se podía ver que en la categoría de 2- estrellas también se ocupaba la palabra “good” muchas veces, pero muy probablemente era para decir que el lugar no era bueno, también se ven términos como “…” que indica insatisfacción en este caso. En ambas categorías se ocupan las palabras “service”, “food”, “go”, “back”; lo que nos hace ver que estas palabras dependiendo de qué tipo de review, se ocuparan en su forma positiva o negativa, ejemplo: “El servicio de este restaurant es muy malo, no volvería a venir a este sitio” vs. “El servicio de este restaurant es genial, volvería a venir”. Lo que hay que destacar es que solo se contó cada palabra que apareciese, solo 1 vez por review y cantidad de estrellas, por lo que esto demuestra que todas las veces que fueron ocupadas, implica que fueron ocupadas en la misma cantidad de reviews con esa cantidad de estrellas, lo cual es algo importante de entender. También es importante mencionar que el tiempo que toma el cálculo de las mejores 20 palabras sin stopwords aumenta considerablemente.

**5. Comentarios y Conclusiones**

El manejo de grandes volúmenes de dato, implica tener que realizar varios operaciones de input y output. Las cuales exceden los límites normales de uso. Por lo que en mi caso, como mi Laptop es antiguo, el hecho de tratar de trabajar con muchos datos (texto plano), hacia que mi computador quedará al borde de un “congelamiento” completo.

Lo que más destaco de esta experiencia, fue más o menos tener una idea bastante mejor de como ocupar lo más básico de nltk en un dataset, entre algunas cosas, el entender que las “stopwords” y las puntuaciones pueden afectar en el análisis de los datos de forma importante y que las palabras pueden decir mucho.

**5. Lista de Bibliografía**

<https://docs.python.org/2/library/collections.html>

<http://www.nltk.org/book/ch01.html>

<http://www.nltk.org/book/ch02.html>

<http://www.nltk.org/book/ch03.html>

<http://stackoverflow.com/questions/6475328/read-large-text-files-in-python-line-by-line-without-loading-it-in-to-memory>

<http://stackoverflow.com/questions/7349646/sorting-a-dictionary-of-tuples-in-python>

<http://stackoverflow.com/questions/8462506/how-to-change-ticks-on-a-histogram-matplotlib>

<http://stackoverflow.com/questions/24809757/how-to-make-a-histogram-from-a-list-of-data>

<http://stackoverflow.com/questions/21489111/how-to-assign-a-plot-to-a-variable-and-use-the-variable-as-the-return-value-in-a>

<http://stackoverflow.com/questions/9622163/save-plot-to-image-file-instead-of-displaying-it-using-matplotlib-so-it-can-be>

<http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.savefig>