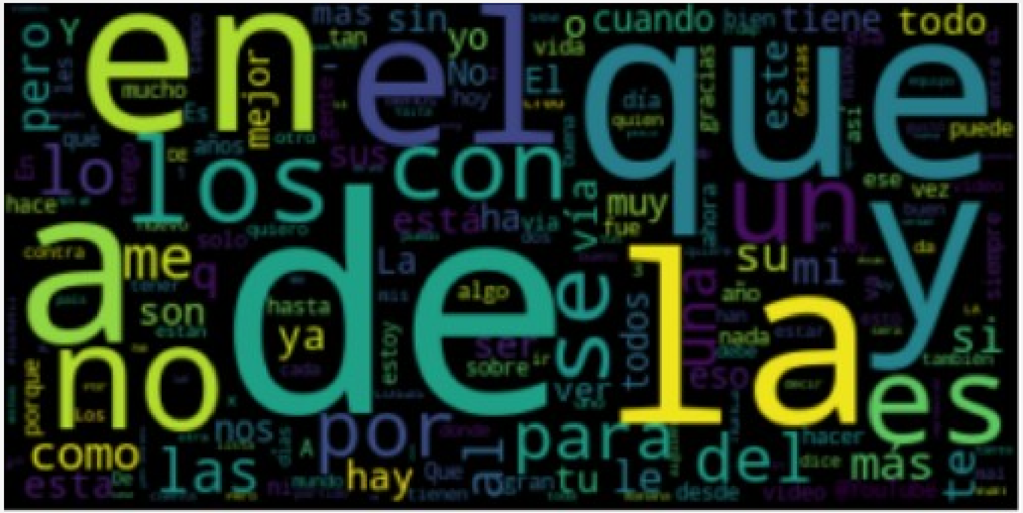
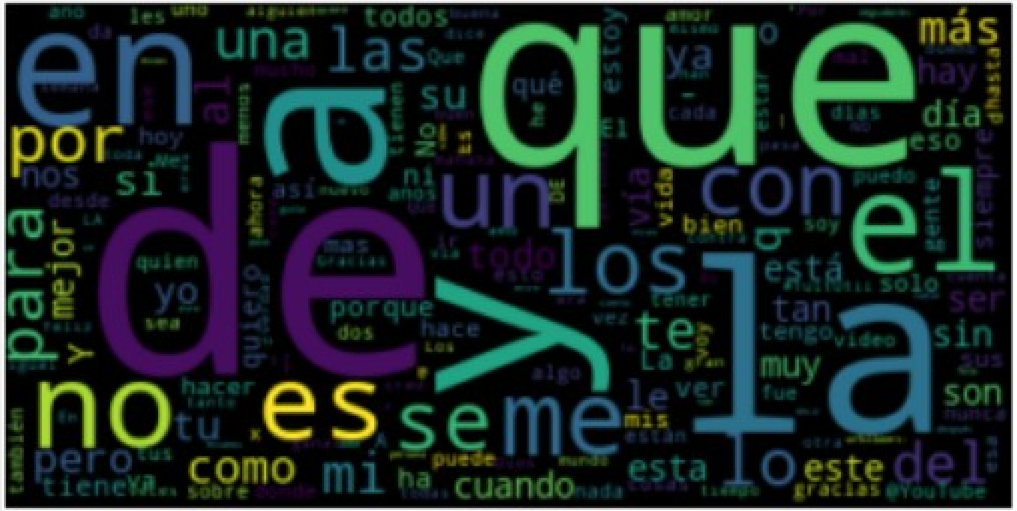
CLASIFICACIÓN POR GÉNERO Y VARIEDAD

JOSE JOAQUÍN RODRIGUEZ
RICARDO CANCAR
ROGER MONT
ALBERTO FERRER
ALEXANDER MARCO
ARACELI TERUEL

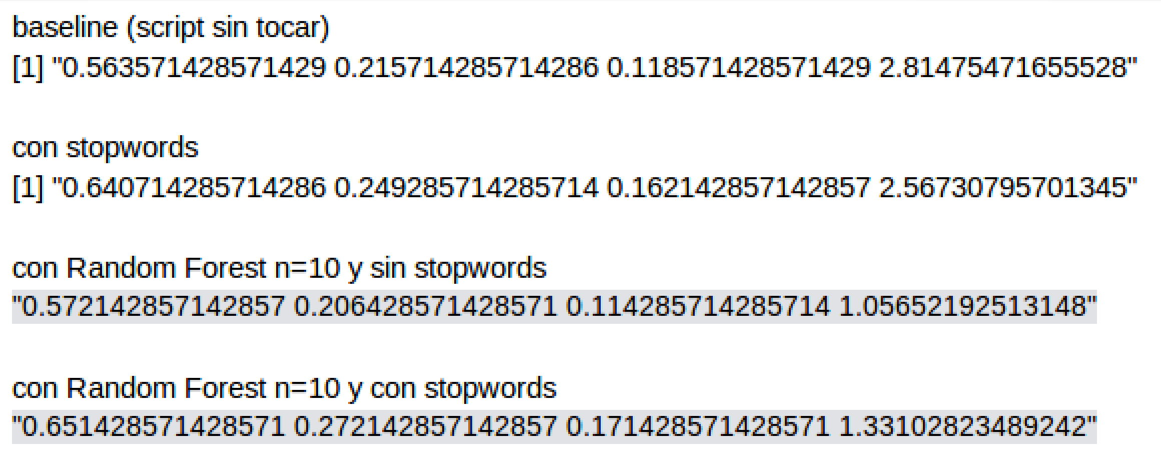
Primeras Pruebas con R





Primeras Pruebas con R

- 10 Palabras más usadas

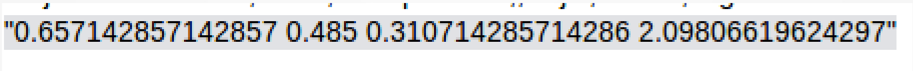


```
baseline (script sin tocar)
[1] "0.563571428571429 0.215714285714286 0.118571428571429 2.81475471655528"

con stopwords
[1] "0.640714285714286 0.249285714285714 0.162142857142857 2.56730795701345"

con Random Forest n=10 y sin stopwords
"0.572142857142857 0.206428571428571 0.114285714285714 1.05652192513148"

con Random Forest n=10 y con stopwords
"0.651428571428571 0.272142857142857 0.171428571428571 1.33102823489242"
```

Buscando palabras más usadas

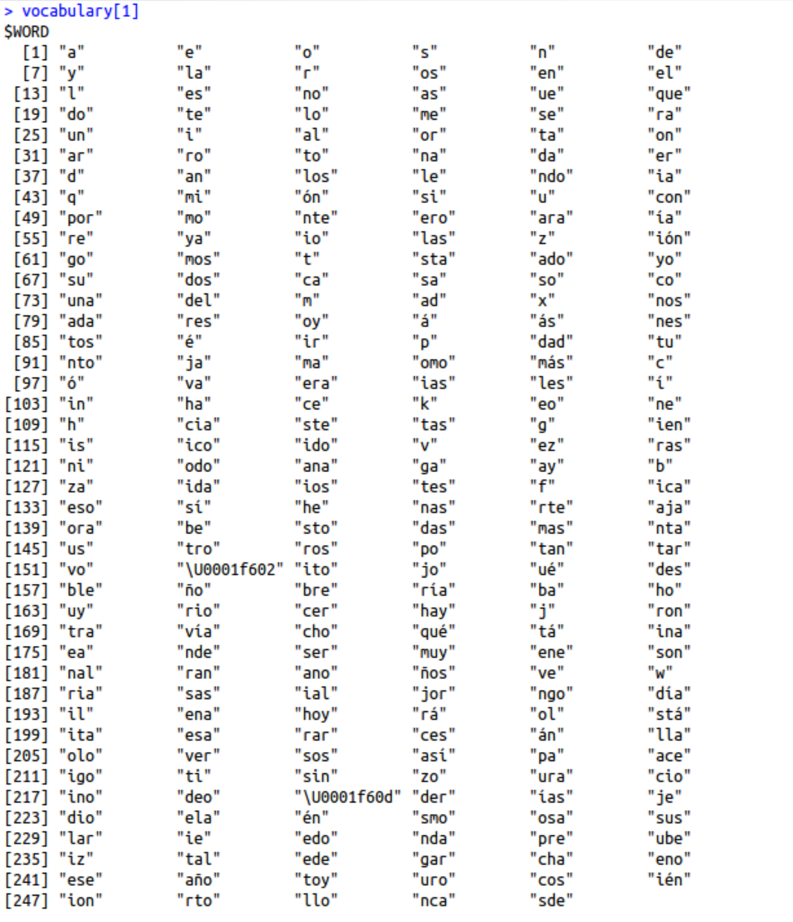
- Hombres: 'enemigo', 'libertad', 'ganar', 'perder', 'batalla'
- Mujeres: 'maravilloso', 'feliz', 'cumpleaños', 'nerviosa', 'hija', 'bebé', 'agradecida'

Random Forest y 100 palabras



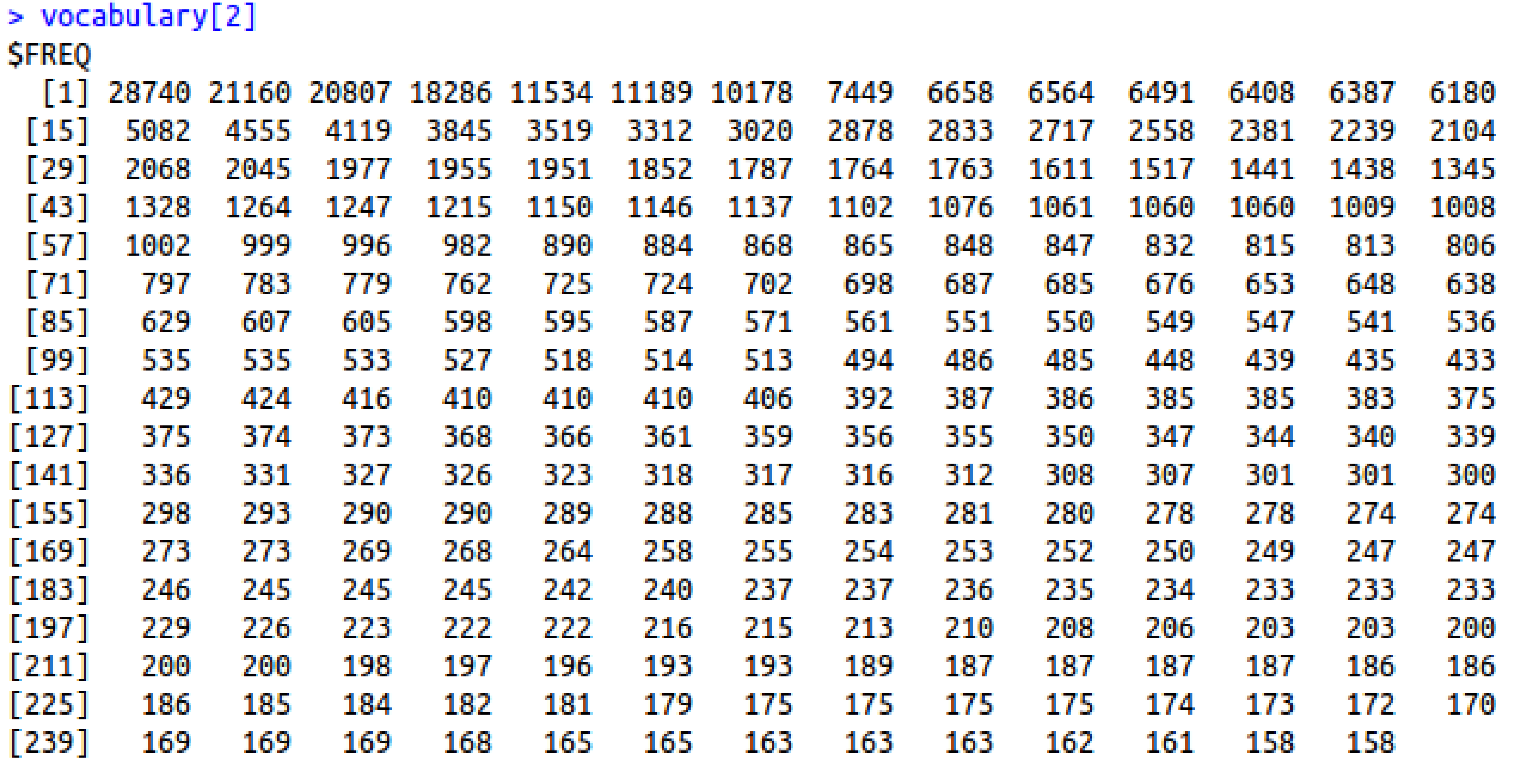
```
"0.657142857142857 0.485 0.310714285714286 2.09806619624297"
```

Jugando con Palabras



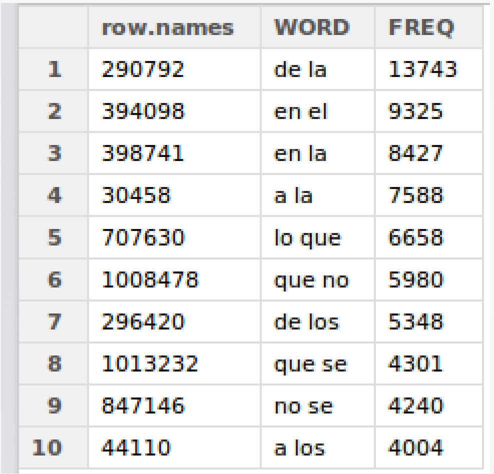
```
> vocabulary[1]
$WORD
  [1] "a"          "e"          "o"          "s"          "n"          "de"
  [7] "y"          "la"         "r"          "os"         "en"         "el"
 [13] "l"          "es"         "no"         "as"         "ue"         "que"
 [19] "do"         "te"         "lo"         "me"         "se"         "ra"
 [25] "un"         "i"          "al"         "or"         "ta"         "on"
 [31] "ar"         "ro"         "to"         "na"         "da"         "er"
 [37] "d"          "an"         "los"        "le"         "ndo"        "ia"
 [43] "q"          "mi"         "ón"         "si"         "u"          "con"
 [49] "por"        "mo"         "nte"        "ero"        "ara"        "ía"
 [55] "re"         "ya"         "io"         "las"        "z"          "ión"
 [61] "go"         "mos"        "t"          "sta"        "ado"        "yo"
 [67] "su"         "dos"        "ca"         "sa"         "so"         "co"
 [73] "una"        "del"        "m"          "ad"         "x"          "nos"
 [79] "ada"        "res"        "oy"         "á"          "ás"         "nes"
 [85] "tos"        "é"          "ir"         "p"          "dad"        "tu"
 [91] "nto"        "ja"         "ma"         "omo"        "más"        "c"
 [97] "ó"          "va"         "era"        "ias"        "les"        "í"
[103] "in"         "ha"         "ce"         "k"          "eo"         "ne"
[109] "h"          "cia"        "ste"        "tas"        "g"          "ien"
[115] "is"         "ico"        "ido"        "v"          "ez"         "ras"
[121] "ni"         "odo"        "ana"        "ga"         "ay"         "b"
[127] "za"         "ida"        "ios"        "tes"        "f"          "ica"
[133] "eso"        "sí"         "he"         "nas"        "rte"        "aja"
[139] "ora"        "be"         "sto"        "das"        "mas"        "nta"
[145] "us"         "tro"        "ros"        "po"         "tan"        "tar"
[151] "vo"         "\U0001f602" "ito"        "jo"         "ué"         "des"
[157] "ble"        "ño"         "bre"        "ría"        "ba"         "ho"
[163] "uy"         "rio"        "cer"        "hay"        "j"          "ron"
[169] "tra"        "vía"        "cho"        "qué"        "tá"         "ina"
[175] "ea"         "nde"        "ser"        "muy"        "ene"        "son"
[181] "nal"        "ran"        "ano"        "ños"        "ve"         "w"
[187] "ria"        "sas"        "ial"        "jor"        "ngo"        "día"
[193] "il"         "ena"        "hoy"        "rá"         "ol"         "stá"
[199] "ita"        "esa"        "rar"        "ces"        "án"         "lla"
[205] "olo"        "ver"        "sos"        "así"        "pa"         "ace"
[211] "igo"        "ti"         "sin"        "zo"         "ura"        "cio"
[217] "ino"        "deo"        "\U0001f60d" "der"        "ías"        "je"
[223] "dio"        "ela"        "én"         "smo"        "osa"        "sus"
[229] "lar"        "ie"         "edo"        "nda"        "pre"        "ube"
[235] "iz"         "tal"        "ede"        "gar"        "cha"        "eno"
[241] "ese"        "año"        "toy"        "uro"        "cos"        "ién"
[247] "ion"        "rto"        "llo"        "nca"        "sde"
```

Jugando con Palabras



```
> vocabulary[2]
$FREQ
  [1] 28740 21160 20807 18286 11534 11189 10178  7449  6658  6564  6491  6408  6387  6180
 [15]  5082  4555  4119  3845  3519  3312  3020  2878  2833  2717  2558  2381  2239  2104
 [29]  2068  2045  1977  1955  1951  1852  1787  1764  1763  1611  1517  1441  1438  1345
 [43]  1328  1264  1247  1215  1150  1146  1137  1102  1076  1061  1060  1060  1009  1008
 [57]  1002   999   996   982   890   884   868   865   848   847   832   815   813   806
 [71]   797   783   779   762   725   724   702   698   687   685   676   653   648   638
 [85]   629   607   605   598   595   587   571   561   551   550   549   547   541   536
 [99]   535   535   533   527   518   514   513   494   486   485   448   439   435   433
[113]   429   424   416   410   410   410   406   392   387   386   385   385   383   375
[127]   375   374   373   368   366   361   359   356   355   350   347   344   340   339
[141]   336   331   327   326   323   318   317   316   312   308   307   301   301   300
[155]   298   293   290   290   289   288   285   283   281   280   278   278   274   274
[169]   273   273   269   268   264   258   255   254   253   252   250   249   247   247
[183]   246   245   245   245   242   240   237   237   236   235   234   233   233   233
[197]   229   226   223   222   222   216   215   213   210   208   206   203   203   200
[211]   200   200   198   197   196   193   193   189   187   187   187   187   186   186
[225]   186   185   184   182   181   179   175   175   175   175   174   173   172   170
[239]   169   169   169   168   165   165   163   163   163   162   161   158   158
```

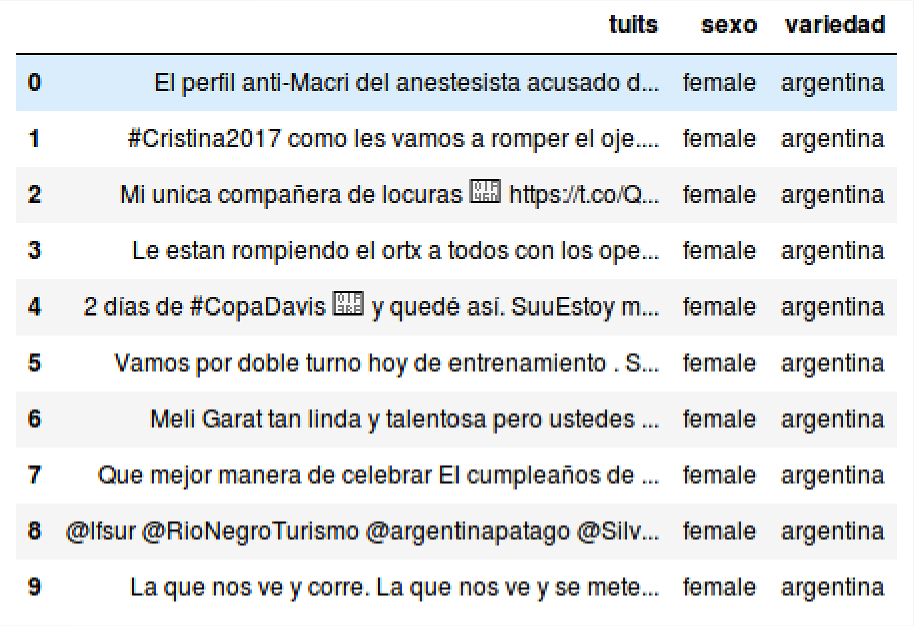
Usando N Gramas



	row.names	WORD	FREQ
1	290792	de la	13743
2	394098	en el	9325
3	398741	en la	8427
4	30458	a la	7588
5	707630	lo que	6658
6	1008478	que no	5980
7	296420	de los	5348
8	1013232	que se	4301
9	847146	no se	4240
10	44110	a los	4004

Pasando a Python

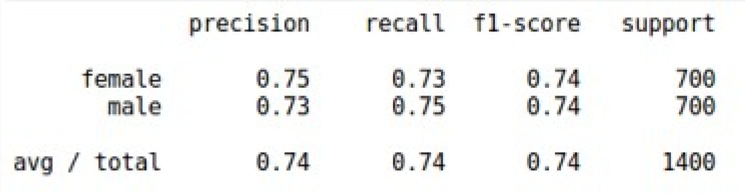
- Recogiendo Datos



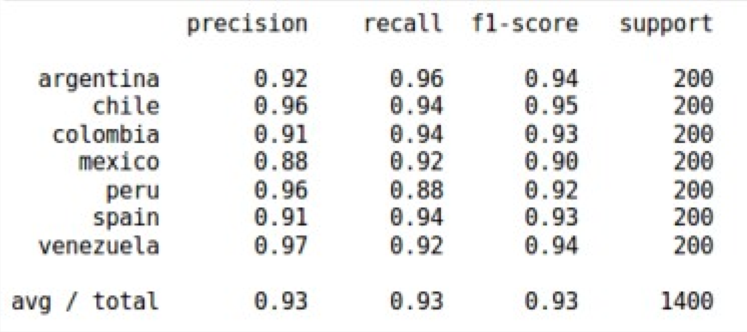
	tuits	sexo	variedad
0	El perfil anti-Macri del anestesista acusado d...	female	argentina
1	#Cristina2017 como les vamos a romper el oje....	female	argentina
2	Mi unica compañera de locuras □ https://t.co/Q...	female	argentina
3	Le estan rompiendo el ortx a todos con los ope...	female	argentina
4	2 días de #CopaDavis □ y quedé así. SuuEstoy m...	female	argentina
5	Vamos por doble turno hoy de entrenamiento . S...	female	argentina
6	Meli Garat tan linda y talentosa pero ustedes ...	female	argentina
7	Que mejor manera de celebrar El cumpleaños de ...	female	argentina
8	@lfsur @RioNegroTurismo @argentinapatago @Silv...	female	argentina
9	La que nos ve y corre. La que nos ve y se mete...	female	argentina

Pasando a Python

- Aplicando un Tf_idf



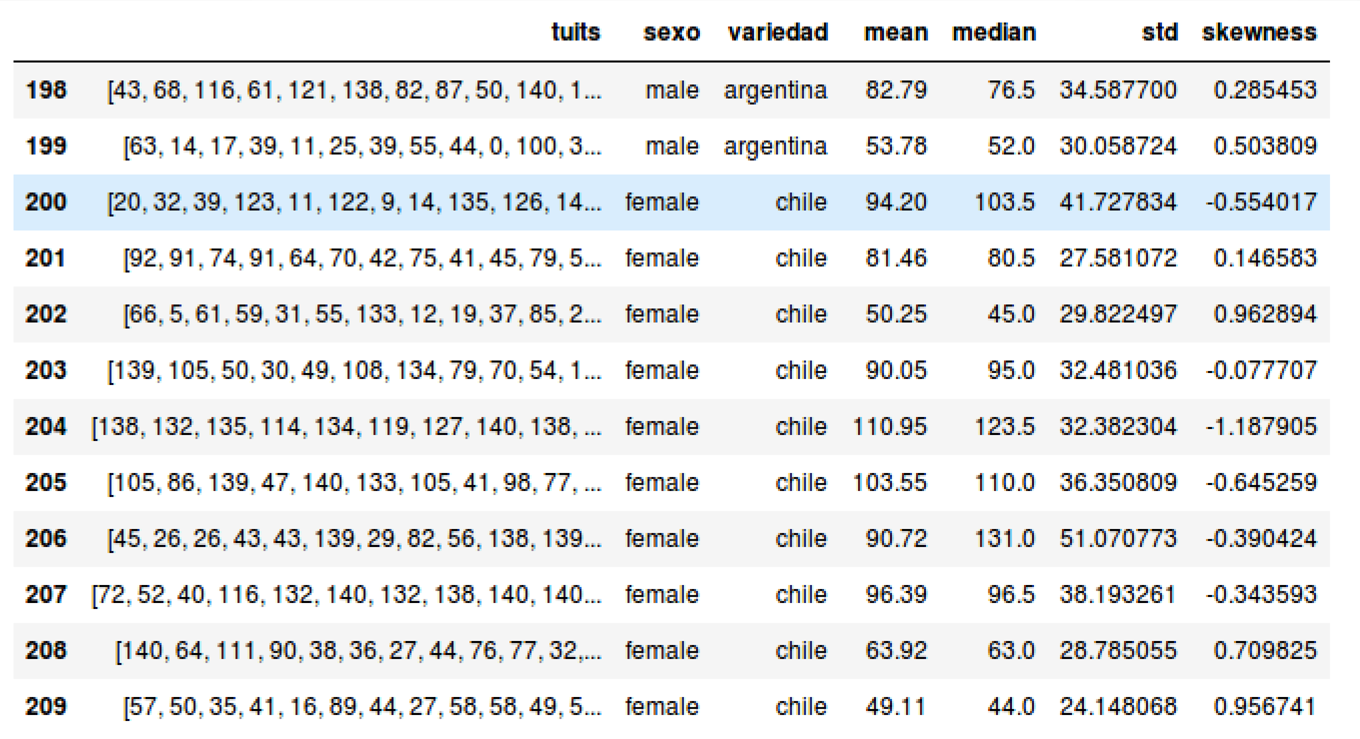
```
             precision    recall  f1-score   support

     female       0.75      0.73      0.74       700
       male       0.73      0.75      0.74       700

avg / total       0.74      0.74      0.74      1400
```



```
             precision    recall  f1-score   support

  argentina       0.92      0.96      0.94       200
      chile       0.96      0.94      0.95       200
   colombia       0.91      0.94      0.93       200
     mexico       0.88      0.92      0.90       200
       peru       0.96      0.88      0.92       200
      spain       0.91      0.94      0.93       200
  venezuela       0.97      0.92      0.94       200

avg / total       0.93      0.93      0.93      1400
```

Longitud Tweets

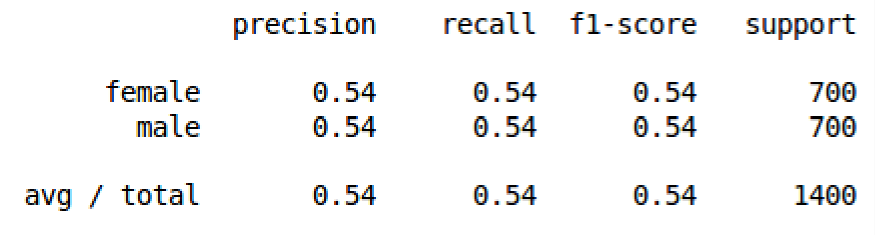
- Datos



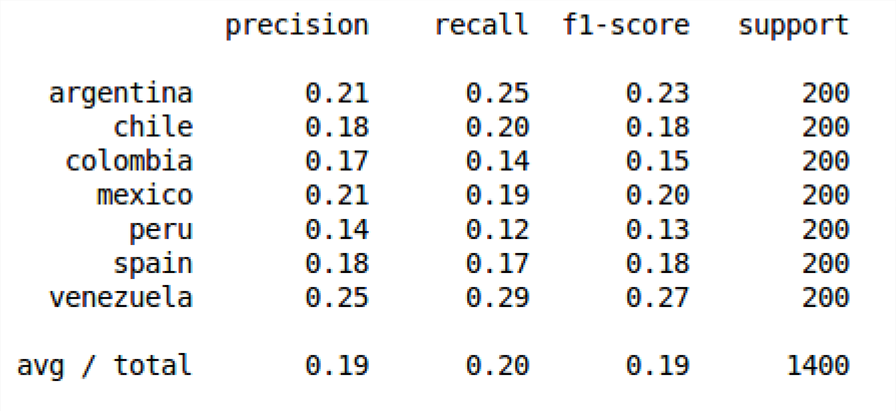
	tuits	sexo	variedad	mean	median	std	skewness
198	[43, 68, 116, 61, 121, 138, 82, 87, 50, 140, 1...	male	argentina	82.79	76.5	34.587700	0.285453
199	[63, 14, 17, 39, 11, 25, 39, 55, 44, 0, 100, 3...	male	argentina	53.78	52.0	30.058724	0.503809
200	[20, 32, 39, 123, 11, 122, 9, 14, 135, 126, 14...	female	chile	94.20	103.5	41.727834	-0.554017
201	[92, 91, 74, 91, 64, 70, 42, 75, 41, 45, 79, 5...	female	chile	81.46	80.5	27.581072	0.146583
202	[66, 5, 61, 59, 31, 55, 133, 12, 19, 37, 85, 2...	female	chile	50.25	45.0	29.822497	0.962894
203	[139, 105, 50, 30, 49, 108, 134, 79, 70, 54, 1...	female	chile	90.05	95.0	32.481036	-0.077707
204	[138, 132, 135, 114, 134, 119, 127, 140, 138, ...	female	chile	110.95	123.5	32.382304	-1.187905
205	[105, 86, 139, 47, 140, 133, 105, 41, 98, 77, ...	female	chile	103.55	110.0	36.350809	-0.645259
206	[45, 26, 26, 43, 43, 139, 29, 82, 56, 138, 139...	female	chile	90.72	131.0	51.070773	-0.390424
207	[72, 52, 40, 116, 132, 140, 132, 138, 140, 140...	female	chile	96.39	96.5	38.193261	-0.343593
208	[140, 64, 111, 90, 38, 36, 27, 44, 76, 77, 32,...	female	chile	63.92	63.0	28.785055	0.709825
209	[57, 50, 35, 41, 16, 89, 44, 27, 58, 58, 49, 5...	female	chile	49.11	44.0	24.148068	0.956741

Longitud Palabras

- Resultados



```
             precision    recall  f1-score   support

     female       0.54      0.54      0.54       700
       male       0.54      0.54      0.54       700

avg / total       0.54      0.54      0.54      1400
```



```
             precision    recall  f1-score   support

  argentina       0.21      0.25      0.23       200
      chile       0.18      0.20      0.18       200
   colombia       0.17      0.14      0.15       200
     mexico       0.21      0.19      0.20       200
       peru       0.14      0.12      0.13       200
      spain       0.18      0.17      0.18       200
  venezuela       0.25      0.29      0.27       200

avg / total       0.19      0.20      0.19      1400
```

Futuras mejoras

- Unir Características:
  - Tf_idf
  - Bolsa palabras hombres/mujeres
  - Bolsa palabras deporte
  - Longitud tweets
  - Signos de puntuación
- Añadir Características:
  - Negación
  - Pronombres
  - Tiempos verbales
  - Terminación de palabras.