

## Inicializando o código

### Extraindo o corpúsculo de tweets.

```
In[*]:= Clear[fulldata]
fulldata = Get[path <> "DocNow_data\\hyd_tweets_consolidated_2021-01-31.dat"];
Dimensions[fulldata]
```

```
Out[*]:= {1544097, 11}
```

Selecionamos, dentre os metadados, o ID, data de criação, e texto dos tweets:

```
In[*]:= Clear[shortdata]
shortdata = Map[{#[ "ID"], #[ "Date"], #[ "Text"]} &, Normal[fulldata]];
Dimensions[shortdata]
```

```
Out[*]:= {1544097, 3}
```

Obtemos o intervalo de datas de criação dos tweets :

```
In[*]:= Clear[dates]
dates = shortdata[ [All, 2] ];
Max[dates]
Min[dates]
```

```
Out[*]:=  Mon 30 Nov 2020 12:55:20 GMT-3.
```

```
Out[*]:=  Wed 18 Nov 2020 03:35:26 GMT-3.
```

## Reconhecendo os idiomas

Primeiro criamos uma função que limpa os strings de construções como links e caracteres não alfanuméricos, para facilitar o reconhecimento dos idiomas.

```
In[*]:= Clear[CleanString]
CleanString[str_] := StringReplace[StringReplace[str,
  {RegularExpression["^[[:ascii:]]"] -> " ", RegularExpression["\\W"] -> " ",
    RegularExpression["[http]*[s]*[:] [\\/\\/] [a-z0-9\\.\\/] + [ ]*" -> " ",
    RegularExpression["[rt: ]*[rt ]*[\\@] [a-z0-9]*[:]*" -> " ",
    RegularExpression["@[a-z0-9]*[:]*" -> " "},
  IgnoreCase -> True], RegularExpression["\\s+"] -> " "];
```

Combinamos essa função com a função LanguageIdentify...

```
In[*]:= Clear[shortdataLang]
shortdataLang =
  Map[{#[ [1]], #[ [2]], #[ [3]], LanguageIdentify[CleanString[#[ [3]]]]} &, shortdata];

In[*]:= shortdataLang = Get[path <> "processing\\shortdata_with_Language-2021-02-20.dat"];
```

## Processando os tweets

Fazemos uma tabela com os 20 idiomas mais frequentes (dos 39 totais) nos tweets sobre Maradona reconhecidas pelo software e a quantidade de tweets em cada idioma.

```
In[*]:= Length[Tally[shortdataLang[All, 4]]]
```

```
Out[*]:= 39
```

```
In[*]:= TablaMaradona = Style[TableForm[{"Maradona Tweets"}, {Take[
  Map[{If[TrueQ[Head[#][1]] == Entity], CommonName[#][1]], #[1]], #[2]] &,
  lang[[1, All]]], 25]}], FontFamily -> "Calibri", FontSize -> 18, Bold]
```

### Maradona Tweets

English	832 276
Spanish	445 410
French	123 464
Portuguese	70 234
Italian	18 878
Malay	17 280
German	8 623
Turkish	6 897
Polish	6 397
Catalan	3 900
Dutch	3 560
Esperanto	1 939
Swedish	907
Swahili	770
Tagalog	670
Finnish	510
Bokmål Norwegian	344
Basque	340
Czech	285
Danish	270
Romanian	187
Bosnian	145
Croatian	102
Slovenian	100
Welsh	78

Selecionamos os tweets das três línguas mais relevantes no corpus (Espanhol, Português e Inglês), para fazer uma série de atividades de limpeza.

```
In[*]:= Clear[selected]
```

```
selected["en"] = Select[shortdataLang, #[[4]] == English LANGUAGE &];
```

```
selected["es"] = Select[shortdataLang, #[[4]] == Spanish LANGUAGE &];
```

```
selected["pt"] = Select[shortdataLang, #[[4]] == Portuguese LANGUAGE &];
```

```
In[ ]:= Clear[mylang]
mylang = {"en", "es", "pt"};
Map[Dimensions[selected[#]] &, mylang]

Out[ ]:= {{832 276, 4}, {445 410, 4}, {70 234, 4}}
```

Obtemos a quantidade de tweets em cada língua.

```
In[ ]:= Clear[text]
Do[text[j] = selected[j][[All, 3]], {j, mylang}]
Map[Dimensions[text[#]] &, mylang]

Out[ ]:= {{832 276}, {445 410}, {70 234}}
```

Limpamos os tweets que não incluem a palavra Maradona.

```
In[ ]:= Clear[myPcrit, myNcrit, myselect, GoodTweetsRT]
myPcrit = {"Maradona"};
myNcrit = {};
myselect[s_] := And[And @@ Map[StringContainsQ[s, #, IgnoreCase → True] &, myPcrit],
  If[Length[myNcrit] ≥ 1,
    And @@ Map[Not[StringContainsQ[s, #, IgnoreCase → True]] &, myNcrit], True]];
Do[GoodTweetsRT[j] = Select[text[j], myselect], {j, mylang}]
```

Fazemos uma tabela com os tweets resultantes

```
In[ ]:= Table[{j} → Length[GoodTweetsRT[j]], {j, mylang}] // TableForm

Out[ ]//TableForm=
{en} → 753 459
{es} → 393 399
{pt} → 63 561
```

Criamos uma função capaz de extrair os nomes de usuários mais mencionados nos tweets e fazemos tabelas com os nomes de usuários mais frequentes sobre Maradona em diferentes línguas.

```
In[ ]:= Clear[username]
username[text_] := Reverse[SortBy[Tally[Flatten[StringCases[text,
  RegularExpression["@" ( [a-z0-9] | [À-ü] | _ ) + "], IgnoreCase → True]]], Last]]];
```

```
In[*]:= Tusernames =
  Style[TableForm[{{"Maradona"}, {Take[usernames[GoodTweetsRT["pt"]], 29]}},
    FontFamily -> "Calibri", FontSize -> 18, Bold]
```

### Maradona

@mundodabola	8597
@Esp_Interativo	3786
@OficialSala12	2913
@DoentesPFutebol	2865
@Atletico	2081
@SCInternacional	1790
@FluminenseFC	1488
@futebol_info	1221
@momentostvbra	1152
@UOL	1042
@UOLEsporte	781
@muitoHumilde	630
@PCBpartidao	623
@gabinolasco	612
@juanj4oficial	607
@luanaraujo90	556
@ColunadoFla	555
@allansimon91	518
@geglobo	404
@DTransferencias	385
@CNNBrasil	368
@sportrecife	359
@FlaGalaxy	343
@lbertozzi	340
@newscolina	328
@ESPNBrasil	283
@calciopedia	282
@William_Castro	280
@LulaOficial	248

```
Out[*]:=
```

Fazemos um gráfico com os nomes de usuários mais mencionados no corpus em Inglês.

```
ln[*]:= Clear[othersUSN, us, US]
us = Take[usernames[GoodTweetsRT["en"]], 29];
othersUSN = Total[usernames[GoodTweetsRT["en"]][[All, 2]]] - Total[us[[All, 2]]];
US = us /. {"@tphoto2005" -> "@████", "@urstrulyMahesh" -> "@████",
"@RahulGandhi" -> "@████", "@imVkohli" -> "@████",
"@433" -> "@████", "@iamsrk" -> "@████", "@queeralamode" -> "@████",
"@afrorevolt" -> "@afro████", "@drkeishakhan" -> "@drkeisha████",
"@praxisnegra" -> "@praxis████", "@quilombomodern" -> "@quilombo████",
"@jonasdiandrade" -> "@jonas████", "@AndressaMDuarte" -> "@Andressa████",
"@Nailahnv" -> "@Naila████", "@jessicabatan" -> "@jessica████"};
BCUMEN = BarChart[us[[All, 2]], ImageSize -> 300,
ChartLegends -> Take[US[[All, 1]], 20], ChartStyle -> {"Pastel"}]
```

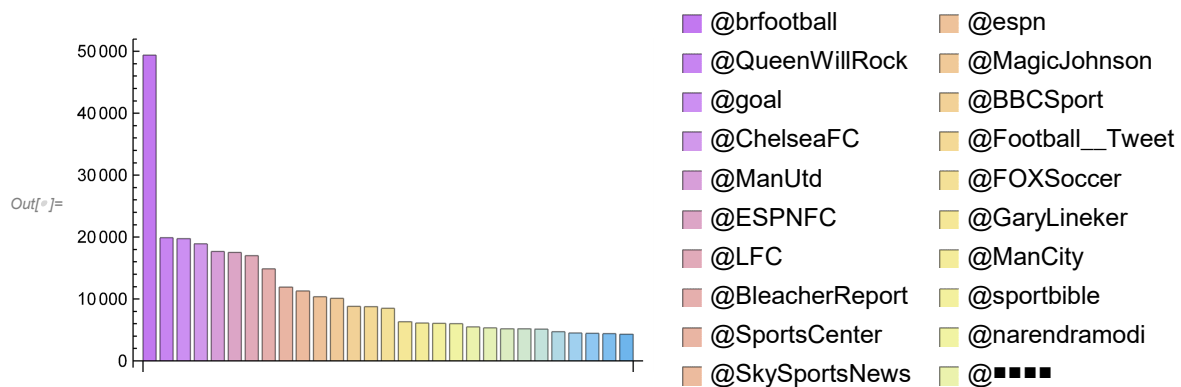


Tabela para os nomes de usuários em espanhol.

```

In[*]:= Clear[othersUSN, us, US]
us = Take[usernames[GoodTweetsRT["es"]], 29];
othersUSN = Total[usernames[GoodTweetsRT["es"]][[All, 2]]] - Total[us[[All, 2]]];
US = us /. {"@tphoto2005" -> "@████", "@urstrulyMahesh" -> "@████",
"@RahulGandhi" -> "@████", "@imVkohli" -> "@████",
"@433" -> "@████", "@iamsrk" -> "@████", "@queeralamode" -> "@████",
"@afrorevolt" -> "@afro████", "@drkeishakhan" -> "@drkeisha████",
"@praxisnegra" -> "@praxis████", "@quilombomodern" -> "@quilombo████",
"@jonasdiandrade" -> "@jonas████", "@AndressaMDuarte" -> "@Andressa████",
"@Nailahnv" -> "@Naila████", "@jessicabatan" -> "@jessica████"};
BCUMEN = BarChart[us[[All, 2]], ImageSize -> 300,
ChartLegends -> Take[US[[All, 1]], 20], ChartStyle -> {"Pastel"}]

```

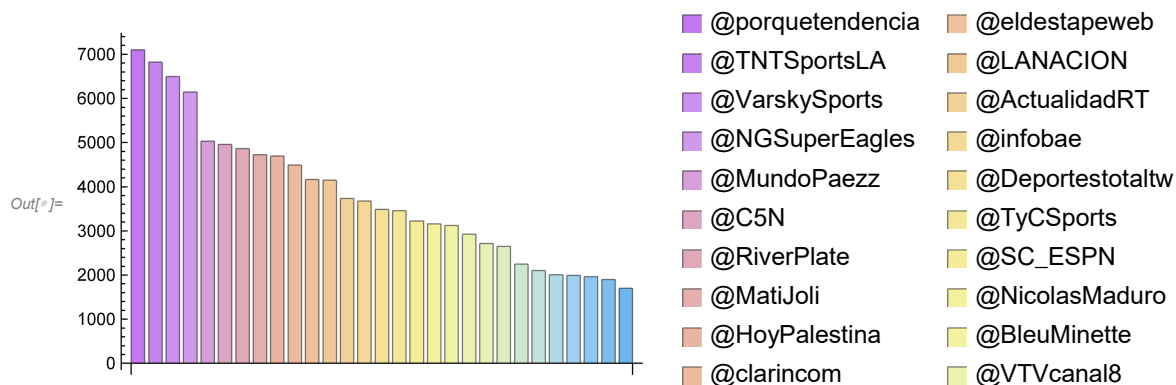
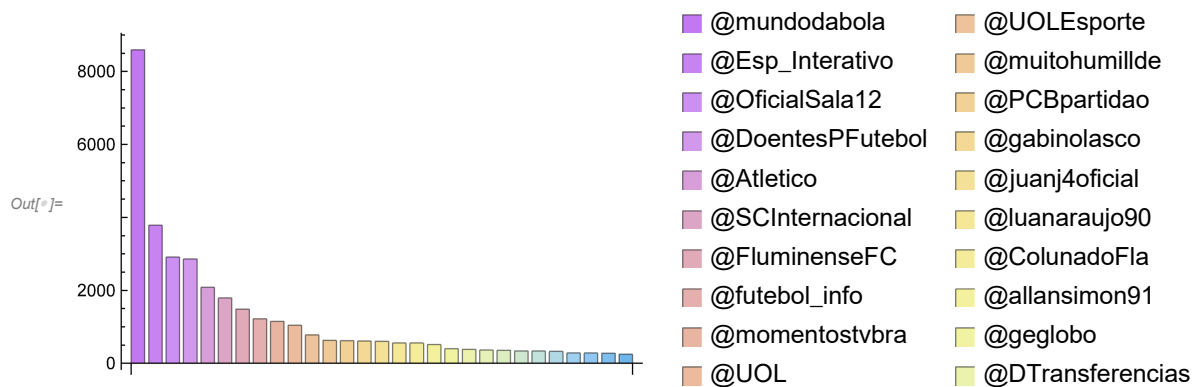


Tabela para nomes de usuários mais frequentes nos tweets em português.

```
In[ ]:= Clear[othersUSN, us, US]
us = Take[usernames[GoodTweetsRT["pt"]], 29];
othersUSN = Total[usernames[GoodTweetsRT["pt"]][[All, 2]]] - Total[us[[All, 2]]];
US = us /. {"@tphoto2005" -> "@", "@urstrulyMahesh" -> "@",
"@RahulGandhi" -> "@", "@imVkohli" -> "@",
"@433" -> "@", "@iamsrk" -> "@", "@queeralamode" -> "@",
"@afrorevolt" -> "@afro", "@drkeishakhan" -> "@drkeisha",
"@praxisnegra" -> "@praxis", "@quilombomodern" -> "@quilombo",
"@jonasdiandrade" -> "@jonas", "@AndressaMDuarte" -> "@Andressa",
"@Nailahnv" -> "@Naila", "@jessicabatan" -> "@jessica"};
BCUMEN = BarChart[us[[All, 2]], ImageSize -> 300,
ChartLegends -> Take[US[[All, 1]], 20], ChartStyle -> {"Pastel"}]
```



Vamos pesquisar os tweets sem os retweets para ter uma medida diferencial da dimensão da produção de conteúdos e da circulação de conteúdos em relação com Maradona.

```
In[ ]:= Clear[NoRT]
Do[NoRT[j] = Select[GoodTweetsRT[j],
Not[StringContainsQ[#, StartOfString ~~ "RT", IgnoreCase -> True]] &], {j, mylang}]
```

```
In[ ]:= TableForm[Table[Length[NoRT[j]], {j, mylang}], TableHeadings -> {mylang}]
```

```
Out[ ]:= TableForm=
en | 94 900
es | 104 207
pt | 12 370
```

Criamos uma série de regras para construir termos significativos e as aplicamos aos tweets com e sem retweets:

```
In[ ]:= Clear[EncodeRules, DecodeRules]
EncodeRules = {"Brazil" → "Brasil", ("Buenos Aires" | "Baires" | "BsAs") → "BuenosAires",
  ("Sao Paulo" | "São Paulo" | "San Pablo") → "SaoPaulo",
  ("Rio de Janeiro" | "RioJaneiro" | "RiodeJaneiro") → "RioDeJaneiro",
  ("NYC" | "New York City" | "NewYorkCity" | "NewYork" |
    "Nova Yorque" | "New York" | "Nueva York" | "Nova York") → "NewYork",
  ("Estados Unidos de América" | "Estados Unidos") → "EstadosUnidos",
  ("United States of America" | "United States") → "UnitedStates",
  ("Latin America" | "Latin American") → "LatinAmerica",
  ("América Latina" | "Latino América" | "Latinoamérica") → "AméricaLatina",
  ("Diego Maradona" | "DiegoMaradona" | "Dieguito Maradona") → "DiegoMaradona",
  "Casa Rosada" → "CasaRosada",
  ("Mané Garrincha" | "Manoel Garrincha") → "ManoelGarrincha",
  ("Lionel Messi" | "Leonel Messi" | "Leo Messi" | "Lio Messi") → "LionelMessi",
  "Boca Juniors" → "BocaJuniors", ("Fidel Castro" | "fidelcastro") → "FidelCastro",
  ("Nicolás Maduro" | "Nicolas Maduro") → "NicolasMaduro",
  ("Cristina Kirchner" | "Cristina C. Kirchner") → "CristinaKirchner",
  "Kobe Bryant" → "KobeBryant",
  "Chadwick Boseman" → "ChadwickBoseman", "All Blacks" → "AllBlacks"};
DecodeRules = {"BuenosAires" → "Buenos Aires", "SaoPaulo" → "São Paulo",
  "RioDeJaneiro" → "Rio de Janeiro", "NewYork" → "New York",
  "EstadosUnidos" → "Estados Unidos", "UnitedStates" → "United States",
  "LatinAmerica" → "Latin America", "AméricaLatina" → "América Latina",
  "DiegoMaradona" → "Diego Maradona", "CasaRosada" → "Casa Rosada",
  "ManoelGarrincha" → "Manoel Garrincha",
  "LionelMessi" → "Lionel Messi", "BocaJuniors" → "Boca Juniors",
  "FidelCastro" → "Fidel Castro", "NicolasMaduro" → "Nicolas Maduro",
  "CristinaKirchner" → "Cristina Kirchner", "KobeBryant" → "Kobe Bryant",
  "ChadwickBoseman" → "Chadwick Boseman", "AllBlacks" → "All Blacks"};

In[ ]:= Clear[EncRT, EncNoRT]
Do[EncRT[j] = Map[{StringReplace[#[[1]], EncodeRules, IgnoreCase → True], #[[2]]} &,
  Tally[GoodTweetsRT[j]]], {j, mylang}];
Do[EncNoRT[j] = Map[{StringReplace[#[[1]], EncodeRules, IgnoreCase → True], #[[2]]} &,
  Tally[NoRT[j]]], {j, mylang}];

In[ ]:= RandomChoice[EncNoRT["es"]]
Out[ ]:= {El Diego político: Maradona sí se mancha | https://t.co/s1q73oMc2M, 1}
```



Operações adicionais de “limpeza” dos tweets com o objetivo de construir uma nuvem de palavras com os termos mais frequentes no corpús.

Fazemos uma “limpeza” de links, nomes de usuários e outros caracteres não alfanuméricos no corpús.

```
In[ ]:= Clear[CleanText, ProperWords]
CleanText[text_] := StringReplace[text,
  {RegularExpression["[http]*[s]*[:] [\\/\"] [a-z0-9.\\/\"]+[" ]*" ]> " ",
   RegularExpression["[rt: ]*[rt ]*[\\@] [a-z0-9]*[:]*" ]> " ",
   RegularExpression["@[a-z0-9]*[:]*" ]> " "}, IgnoreCase -> True];
ProperWords[lw_] := Select[lw, And[StringFreeQ[#, RegularExpression["\\W"]],
  StringLength[#] > 4] &];

In[ ]:= Clear[ttweetsRT, ttweetsNoRT]
Do[ttweetsRT[j] =
  SortBy[Map[{ProperWords[TextWords[CleanText[#[[1]]]]], #[[2]]] &, EncRT[j]],
  -#[[2]] &], {j, mylang}];
Do[ttweetsNoRT[j] = SortBy[Map[{ProperWords[TextWords[CleanText[#[[1]]]]], #[[2]]] &,
  EncNoRT[j]], -#[[2]] &], {j, mylang}];

In[ ]:= RandomChoice[ttweetsNoRT["pt"], 3]

Out[ ]:= {{Futebol, Diego, Maradona, faleceu}, 1}, {{diego, maradona, morreu}, 1},
  {{Diego, Maradona, história, futebol, mundial, chegou, segundo, história}, 1}}

In[ ]:= Length[ttweetsRT["pt"]]

Out[ ]:= 14 590

In[ ]:= TableForm[Outer[Length[ttweetsRT[#1]] &, mylang], TableHeadings -> {mylang}]

Out[ ]//TableForm=
en | 111 450
es | 118 160
pt | 14 590
```

Partindo de uma lista de stopwords, ou palavras sem valor léxico, (artigos, pronomes, preposições) em português, inglês e espanhol (artigos, pronomes, preposições) em português, inglês e espanhol (disponíveis em <http://snowball.tartarus.org/algorithms/portuguese/stop.txt>, <http://snowball.tartarus.org/algorithms/spanish/stop.txt> e <http://snowball.tartarus.org/algorithms/english/stop.txt> e com acréscimos próprios), eliminamos todas essas palavras de nosso corpus.

```
In[ ]:= Clear[swes, swpt, swen, stopwords, RemoveStopWords, cleantweetsNoRT0, cleantweetsRT0]
swes = Get[path <> "dictionaries\\sw-es.txt"];
swpt = Get[path <> "dictionaries\\sw-pt.txt"];
swen = Get[path <> "dictionaries\\sw-en.txt"];
stopwords = Join[swes, swpt, swen];
RemoveStopWords[s_] :=
  Select[s, Not[StringMatchQ[#, Alternatives @@ stopwords, IgnoreCase -> True]] &];
Do[cleantweetsRT0[j] = Map[{RemoveStopWords[#[[1]]], #[[2]]] &, tweetsRT[j]],
  {j, mylang}];
Do[cleantweetsNoRT0[j] = Map[{RemoveStopWords[#[[1]]], #[[2]]] &, tweetsNoRT[j]],
  {j, mylang}];

In[ ]:= RandomChoice[cleantweetsRT0["es"], 4]
Out[ ]:= {{juzgan, deben, juzgados, humanidad, Diego, Maradona, bendecidos}, 1},
  {{27Nov, coincidencia, muerte, Diego, Maradona, amigo, Fidel, Castro}, 1},
  {{DIEGO, MARADONA, CUMPLO, MUERTO, PASAS, DORMIRE, CREES,
    BUSCALO, GOOGLE, DIEMGOM, MARAMDONAM, MANDA, IGNORÓ, MURIÓ}, 3},
  {{muerte, Diego, Maradona, Matías, Morla, acusó, personal, salud}, 1}}
```

Depois consolidamos nosso corpus agrupando aqueles tweets que tenham ficado iguais após a eliminação das stopwords.

```
In[ ]:= Clear[ConsolidateTally, cleantweetsNoRT, cleantweetsRT]
ConsolidateTally[tally_] :=
  Map[{#[[1, 1]], Total[#[[All, 2]]]} &, Gather[tally, SameQ[#[[1]], #[[1]]] &]];
Do[cleantweetsNoRT[j] = ConsolidateTally[cleantweetsRT0[j]], {j, mylang}];
Do[cleantweetsRT[j] = ConsolidateTally[cleantweetsNoRT0[j]], {j, mylang}];
```

A última operação de limpeza tem a ver com obter as raízes das palavras nas três línguas e agrupar as palavras de acordo com a palavra mais frequente da mesma raiz

Primeiro computamos a quantidade de repetições das palavras com uma função que considera a multiplicidade de tweets.

```
In[ ]:= Clear[DistributeList, ConsolidateTally, wordtallyNoRT, wordtallyRT]
DistributeList[{l_List, x_}] := Map[{#, x} &, l];
ConsolidateTally[tally_] :=
  Map[{#[[1, 1]], Total[#[[All, 2]]]} &, Gather[tally, SameQ[#[[1]], #[[2]]] &]];
Do[wordtallyNoRT[j] = SortBy[ConsolidateTally[
  Join@@Map[DistributeList, cleantweetsNoRT[j]]], -#[[2]] &], {j, mylang}];
Do[wordtallyRT[j] = SortBy[ConsolidateTally[
  Join@@Map[DistributeList, cleantweetsRT[j]]], -#[[2]] &], {j, mylang}];
```

Definimos um comando que obtém o termo mais comum de uma classe de termos para cada língua. Fazemos isso através das funções WordStem, WordStemES e WordStemPT. A primeira é uma função de Wolfram que obtém a raiz das palavras. As duas últimas foram implementadas por nós no Wolfram Mathematica a partir do algoritmo para obter raízes dessas línguas presente aqui: <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html> e <http://snowball.tartarus.org/algorithms/spanish/stemmer.html>.

```
In[ ]:= Clear[MeaningfulTermsRules]
MeaningfulTermsRules[ts_, "en"] := Module[{cat, f},
  cat = Gather[ts,
    SameQ[WordStem[ToLowerCase[#[[1]]]], WordStem[ToLowerCase[#[[2]]]]] &];
  f[c_] := Map[{#[[1]]} → Last[SortBy[c, Last]][[1]] &, c];
  Select[Union[Flatten[Map[f, cat]]], Not[SameQ[#[[1]], #[[2]]]] &];
MeaningfulTermsRules[ts_, "es"] := Module[{cat, f},
  cat = Gather[ts,
    SameQ[WordStemES[ToLowerCase[#[[1]]]], WordStemES[ToLowerCase[#[[2]]]]] &];
  f[c_] := Map[{#[[1]]} → Last[SortBy[c, Last]][[1]] &, c];
  Select[Union[Flatten[Map[f, cat]]], Not[SameQ[#[[1]], #[[2]]]] &];
MeaningfulTermsRules[ts_, "pt"] := Module[{cat, f},
  cat = Gather[ts,
    SameQ[WordStemPT[ToLowerCase[#[[1]]]], WordStemPT[ToLowerCase[#[[2]]]]] &];
  f[c_] := Map[{#[[1]]} → Last[SortBy[c, Last]][[1]] &, c];
  Select[Union[Flatten[Map[f, cat]]], Not[SameQ[#[[1]], #[[2]]]] &];
```

Geramos uma série de regras de transformação das palavras a partir de nosso corpus dos tweets em diferentes línguas.

```
In[ ]:= Clear[mrulesNoRT, mrulesRT]
Do[mrulesNoRT[j] = MeaningfulTermsRules[wordtallyNoRT[j], j], {j, mylang}];
Do[mrulesRT[j] = MeaningfulTermsRules[wordtallyRT[j], j], {j, mylang}];
```

```
In[ ]:= mrulesNoRT["pt"]
```

Out[ ]:=

```
{2022Presidente → 2022PRESIDENTE, 21H30 → 21h30, Abaixo → abaixo, abala → abalou,
abalada → abalou, abalado → abalou, abalados → abalou, abalar → abalou,
abalará → abalou, abandonado → abandono, abandonam → abandono, abandonar → abandono,
abandonou → abandono, abatido → abatidos, aberta → aberto, ABERTA → aberto,
abertas → aberto, ... 5732 ..., volte → volta, voltei → volta, Voltei → volta,
volto → volta, voltou → volta, votar → votou, votaram → votou, votaria → votou,
vulgar → vulgo, world → World, xingando → xingar, youtube → YouTube,
Youtube → YouTube, YOUTUBE → YouTube, zmarł → Zmarł, zueira → zueiras}
```

large output

[show less](#)

[show more](#)

[show all](#)

[set size limit...](#)

```
In[ ]:= RandomChoice[mrulesRT["en"], 20]
```

```
Out[ ]:= {FARMERS → farmers, Redemption → redemption, confirmation → confirmed, scooped → scoop,
originally → original, newscast → Newscast, SENOR → Senor, Grand → grand,
Playmaker → playmaker, spanning → spans, driving → drive, directors → director,
Studio → studio, POLITICAL → politics, Diegos → Diego, Middle → middle,
accepting → accept, NIALl → Niall, REMEMBER → remember, protecting → protect}
```

Vamos substituir as palavras pela palavra mais frequente do grupo dado pela mesma raiz, de acordo com as regras já computadas.

```
In[ ]:= Clear[mtweetsNoRT, mtweetsRT]
Do[mtweetsNoRT[j] = Map[{Sort[#[[1]]], #[[2]]} &,
  Replace[cleantweetsNoRT[j], mrulesNoRT[j], {3}]], {j, mylang}];
Do[mtweetsRT[j] = Map[{Sort[#[[1]]], #[[2]]} &,
  Replace[cleantweetsRT[j], mrulesRT[j], {3}]], {j, mylang}];
```

```
In[ ]:= RandomChoice[mtweetsRT["pt"], 4]
```

```
Out[ ]:= {{atenção, atleta, braço, conservadores, Diego,
fenomenal, Fidel, filho, generoso, gigante, Guevara, homenagear,
Maradona, Plantão, Saiba, sentir, simpatia, social, tatuagem}, 1},
{{argentino, Diego, imprensa, Maradona, morte}, 1},
{{argentino, Diego, espera, feriado, Maradona, morte, viram}, 1},
{{argentino, brabo, Diego, futebol, ídolo, Maradona, perda}, 1}}
```

Vamos computar as palavras que só aparecem uma vez em cada conjunto de tweets (por língua), que consideraremos triviais para obtenção dos temas principais.

```
In[ ]:= Clear[trivialwordsNoRT, trivialwordsRT]
Do[trivialwordsNoRT[j] =
  Select[Tally[Flatten[mtweetsNoRT[j][[All, 1]]]], #[[2]] < 2 &][[All, 1]], {j,
  mylang}];
Do[trivialwordsRT[j] = Select[Tally[Flatten[mtweetsRT[j][[All, 1]]]], #[[2]] < 2 &][[
  All, 1]], {j, mylang}];
```

Computamos os tweets significativos eliminando as palavras que aparecem só uma vez, as quais consideraremos triviais para extrair os temas principais.

```
In[ ]:= Clear[MtweetsNoRT, MtweetsRT]
Do[MtweetsNoRT[j] = ConsolidateTally[
  Select[Map[{Sort[Complement#[[1]], trivialwordsNoRT[j]]], #[[2]]} &,
    mtweetsRT[j]], Length[#] ≥ 1 &], {j, mylang}];
Do[MtweetsRT[j] = ConsolidateTally[Select[
  Map[{Sort[Complement#[[1]], trivialwordsRT[j]]], #[[2]]} &, mtweetsRT[j]],
  Length[#] ≥ 1 &], {j, mylang}];
```

```
In[ ]:= Clear[mwordsNoRT, mwordsRT]
Do[mwordsNoRT[j] = SortBy[ConsolidateTally[
  Join@@Map[DistributeList, MtweetsNoRT[j]]], -#[[2]] &], {j, mylang}];
Do[mwordsRT[j] = SortBy[ConsolidateTally[Join@@Map[DistributeList, MtweetsRT[j]]],
  -#[[2]] &], {j, mylang}];
```

Guardando nossos processamentos feitos até agora.

```
In[ ]:= Do[
  Put[{ttweetsRT[j], cleantweetsRT0[j], cleantweetsRT[j], wordtallyRT[j], mrulesRT[j],
    trivialwordsRT[j], MtweetsRT[j], mtweetsRT[j], mwordsRT[j], ttweetsNoRT[j],
    cleantweetsNoRT0[j], cleantweetsNoRT[j], wordtallyNoRT[j], mrulesNoRT[j],
    trivialwordsNoRT[j], MtweetsNoRT[j], mtweetsNoRT[j], mwordsNoRT[j]},
    path <> "output\\Maradona_processed_tweets-" <> ToString[j] <> ".dat"], {j, mylang}];
```

Recuperando nossos processamentos feitos até agora.

```
In[ ]:= Clear[mylang, ttweetsRT, cleantweetsRT0, cleantweetsRT,
  wordtallyRT, mrulesRT, trivialwordsRT, MtweetsRT, mtweetsRT, mwordsRT,
  ttweetsNoRT, cleantweetsNoRT0, cleantweetsNoRT, wordtallyNoRT,
  mrulesNoRT, trivialwordsNoRT, MtweetsNoRT, mtweetsNoRT, mwordsNoRT];
mylang = {"en", "es", "pt"};
Do[{ttweetsRT[j], cleantweetsRT0[j], cleantweetsRT[j], wordtallyRT[j], mrulesRT[j],
  trivialwordsRT[j], MtweetsRT[j], mtweetsRT[j], mwordsRT[j], ttweetsNoRT[j],
  cleantweetsNoRT0[j], cleantweetsNoRT[j], wordtallyNoRT[j], mrulesNoRT[j],
  trivialwordsNoRT[j], MtweetsNoRT[j], mtweetsNoRT[j], mwordsNoRT[j]} =
  Get[path <> "output\\Maradona_processed_tweets-" <> ToString[j] <> ".dat"], {j,
  mylang}];
```

Tabelas para o número de tweets significativos únicos, com ou sem retweets:

```
In[ ]:= TableForm[Table[Length[MtweetsNoRT[j]], {j, mylang}], TableHeadings -> {mylang}]
TableForm[Table[Length[MtweetsRT[j]], {j, mylang}], TableHeadings -> {mylang}]
```

```
Out[ ]:= TableForm=
en | 48 397
es | 59 535
pt | 8798
```

```
Out[ ]:= TableForm=
en | 48 345
es | 59 480
pt | 8791
```

Com os processamentos resultantes, elaboramos Nuvens de Palavras.

```
In[ ]:= Do[mywcRT[j] = WordCloud[mwordsRT[j][[3 ;; 400]], PlotTheme -> "Web",
      MaxItems -> 400, WordOrientation -> "HorizontalVertical", ImageSize -> 500];
      Put[mywcRT[j], path <> "output\\mywcRT-" <> "-" <> j <> ".dat"];
      Print["Done RT: " <> j];, {j, mylang}]

Done RT: en
Done RT: es
Done RT: pt

In[ ]:= Do[mywcRT[j] = Get[path <> "output\\mywcRT-" <> "-" <> j <> ".dat"], {j, mylang}]

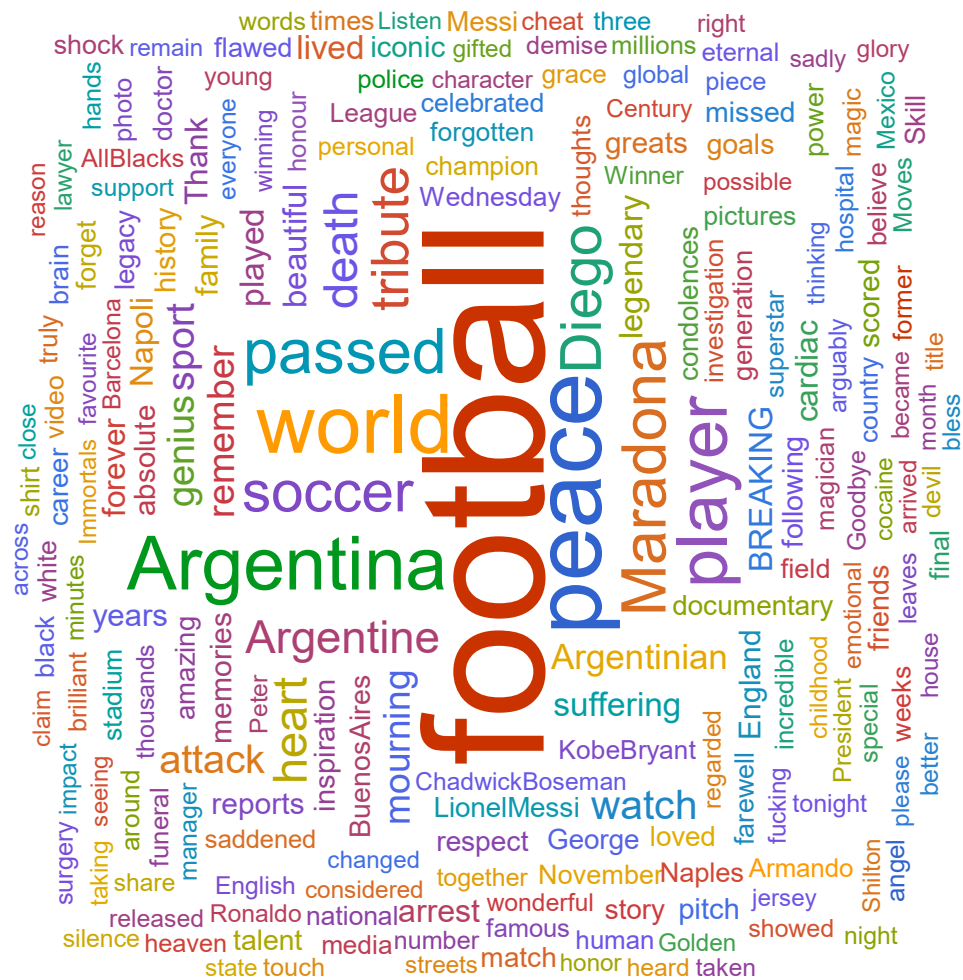
In[ ]:= Do[Export[path <> "output\\mywcRTpicture-" <> j <> ".png", mywcRT[j]], {j, mylang}];

Clear[mywcRTimage]
Do[
  mywcRTimage[j] = Import[path <> "output\\mywcRTpicture-" <> j <> ".png"], {j, mylang}];
```

Nuvem de palavras para os tweets sobre Maradona em inglês

```
In[ ]:= WordCloud[mwordsRT["en"][[3 ;; 400]], PlotTheme -> "Web",
MaxItems -> 200, WordOrientation -> "HorizontalVertical", ImageSize -> 500]
```

Out[ ]:=



Nuvem de palavras para os tweets sobre Maradona em espanhol

```
ln[5] := WordCloud[mwordsRT["es"][[3 ;; 400]], PlotTheme -> "Web",
  MaxItems -> 200, WordOrientation -> "HorizontalVertical", ImageSize -> 500]
```





## Nuvem de palavras para os tweets sobre Maradona em português

```
In[ ] := WordCloud[mwordsRT["pt"]][[3 ;; 400]], PlotTheme -> "Web",
  MaxItems -> 200, WordOrientation -> "HorizontalVertical", ImageSize -> 500]
```

