

Projeto 2: Analisando dados com Spark

Gustavo Molina e Guilherme Batista

```
In [21]: df_mercedes
```

```
Out[21]:
```

palavra	frequencia
-	52872
gasolina	5488
serie	5216
jardim	4988
mecanico	4458
...	...
3.0	379
maria	376
c	374
dia	370
3	368

100 rows × 1 columns

```
In [23]: df_mercedes_fiat
```

```
Out[23]:
```

palavra	frequencia
diesel	3.853211
-	3.826981
revisão	3.407561
2p	3.369772
jardim	3.351796
...	...
6x2	2.376577
oficina	2.376577
iveco	2.369216
(2)	2.363612
jeep	2.357935

100 rows × 1 columns

```
In [22]: df_fiat
```

```
Out[22]:
```

palavra	frequencia
-	4.639954
(1)	3.525563
=	3.438701
jardim	3.402433
tv	3.366423
...	...
lopes	2.556303
(5)	2.553883
la	2.551450
câmera	2.551450
3	2.549003

100 rows × 1 columns

Análise e visualização dos resultados



Img: WorldCloud Mercedes.



Img: WorldCloud Fiat.



Img: WorldCloud Mercedes e Fiat

Com base nos resultados obtidos nessa worldCloud é razoável pensar que as palavras com frequência maior entre Mercedes e Fiat se deve a uma comparação que as pessoas estão fazendo entre os dois carros, com fatores como revisão e diesel como as palavras com maior frequência encontrada. Outras palavras que fortalecem essa *hipótese* são outras palavras com frequência considerável encontradas, como chevrolet, volkswagen, audi, etc, que indicam algum tipo de benchmark.

Para a Mercedes, no WorldCloud algumas hipóteses interessantes também podem ser feitas: a palavra Santa, João e Maria poderiam indicar que algum personagem ou pessoa famosa utiliza o carro em algum vídeo no Youtube e/ou ficou famosa por algo relacionada a marca.

Já no Fiat, a frequência de palavras mais alta está relacionada a estados como SP, MG, Belo e Horizonte (separadas). Seus usos indicam ser mais relacionados ao cotidiano do povo brasileiro, que tem uma frequência de palavras buscadas muito maior que a da Mercedes. (como pode ser visto nas tabelas acima)