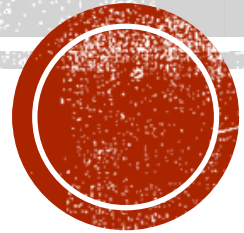
PROJETO FRAMEWORKS DE BIG DATA



MAPREDUCE 01

idDoc_termo

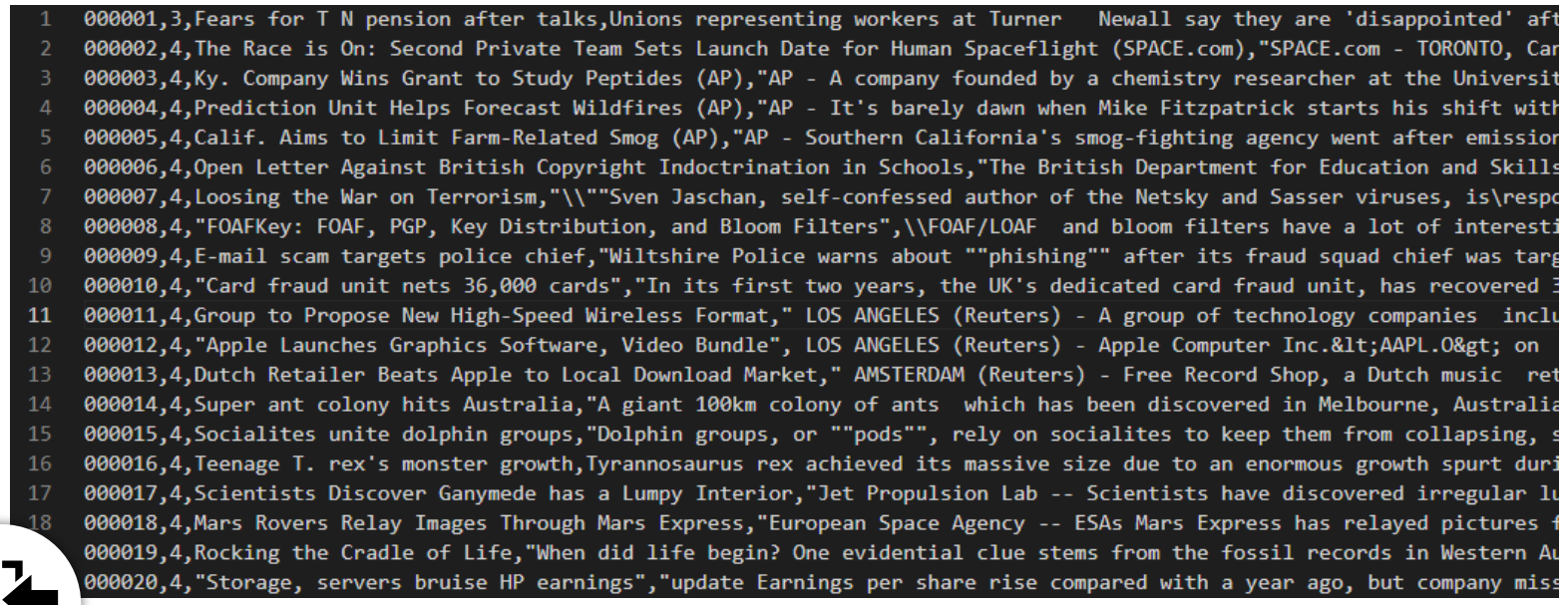
term_frequency

Processamentos:

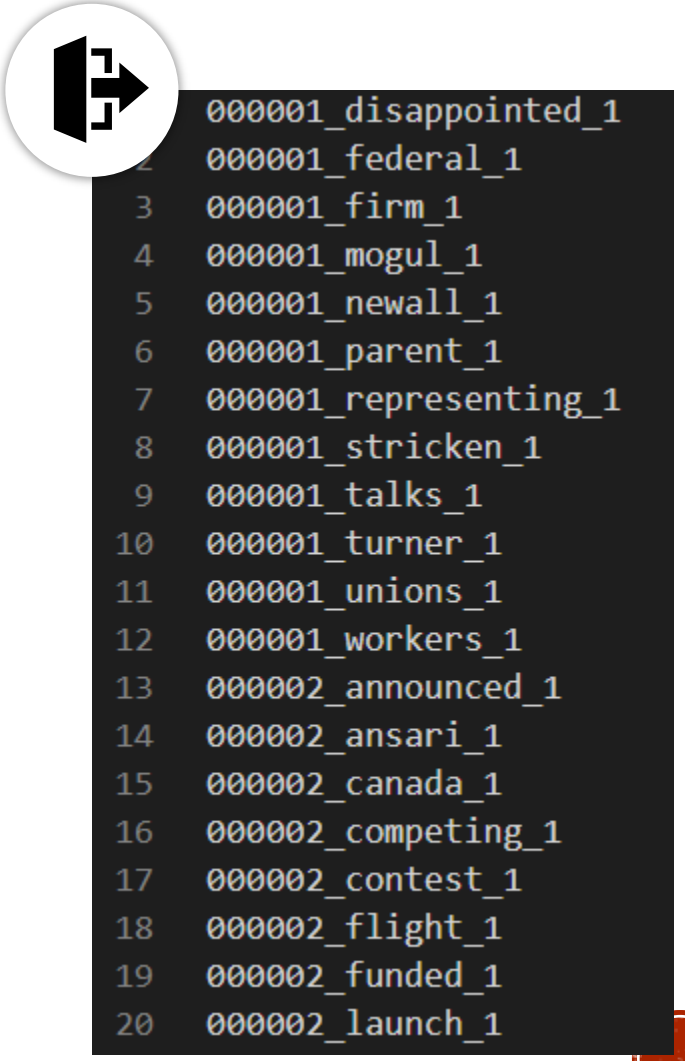
- Remoção de stopwords;
- Remoção de caracteres especiais;
- Remoção de números;
- Remoção de múltiplos espaços;
- Remoção de letras sozinhas.

- Recebemos o arquivo original e convertemos ele em um formato de chave e valor.
- Após isso, fazemos a contagem do term_frequency (frequência de cada termo no documento) para cada termo.
- Termos são as palavras do texto, e “idDoc” é o id do documento.
- Cada linha se refere a um documento diferente.

MR01 - INPUT X OUTPUT



```
000001,3,Fears for T N pension after talks,Unions representing workers at Turner   Newall say they are 'disappointed' aft
000002,4,The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com),"SPACE.com - TORONTO, Car
000003,4,Ky. Company Wins Grant to Study Peptides (AP),"AP - A company founded by a chemistry researcher at the Universit
000004,4,Prediction Unit Helps Forecast Wildfires (AP),"AP - It's barely dawn when Mike Fitzpatrick starts his shift with
000005,4,Calif. Aims to Limit Farm-Related Smog (AP),"AP - Southern California's smog-fighting agency went after emissior
000006,4,Open Letter Against British Copyright Indoctrination in Schools,"The British Department for Education and Skills
000007,4,Loosing the War on Terrorism,"\\""Sven Jaschan, self-confessed author of the Netsky and Sasser viruses, is\respo
000008,4,"FOAFKey: FOAF, PGP, Key Distribution, and Bloom Filters",\\FOAF/LOAF  and bloom filters have a lot of interesti
000009,4,E-mail scam targets police chief,"Wiltshire Police warns about ""phishing"" after its fraud squad chief was targ
000010,4,"Card fraud unit nets 36,000 cards","In its first two years, the UK's dedicated card fraud unit, has recovered 3
000011,4,Group to Propose New High-Speed Wireless Format," LOS ANGELES (Reuters) - A group of technology companies  inclu
000012,4,"Apple Launches Graphics Software, Video Bundle", LOS ANGELES (Reuters) - Apple Computer Inc.&lt;AAPL.O&gt; on
000013,4,Dutch Retailer Beats Apple to Local Download Market," AMSTERDAM (Reuters) - Free Record Shop, a Dutch music  ret
000014,4,Super ant colony hits Australia,"A giant 100km colony of ants  which has been discovered in Melbourne, Australia
000015,4,Socialites unite dolphin groups,"Dolphin groups, or ""pods"", rely on socialites to keep them from collapsing, s
000016,4,Teenage T. rex's monster growth,Tyrannosaurus rex achieved its massive size due to an enormous growth spurt duri
000017,4,Scientists Discover Ganymede has a Lumpy Interior,"Jet Propulsion Lab -- Scientists have discovered irregular lu
000018,4,Mars Rovers Relay Images Through Mars Express,"European Space Agency -- ESAs Mars Express has relayed pictures f
000019,4,Rocking the Cradle of Life,"When did life begin? One evidential clue stems from the fossil records in Western Au
000020,4,"Storage, servers bruise HP earnings","update Earnings per share rise compared with a year ago, but company miss
```



```
000001_disappointed_1
000001_federal_1
000001_firm_1
000001_mogul_1
000001_newall_1
000001_parent_1
000001_representing_1
000001_stricken_1
000001_talks_1
000001_turner_1
000001_unions_1
000001_workers_1
000002_announced_1
000002_ansari_1
000002_canada_1
000002_competing_1
000002_contest_1
000002_flight_1
000002_funded_1
000002_launch_1
```

MAPREDUCE 02

termo

idDoc_termFrequency_documentFrequency

- O termo se torna chave para podermos realizar a contagem do document_frequency, que será igual ao número de elementos da lista gerada pelo shuffle.
- Os resultados das variáveis precisam ser preservados para a próxima etapa.

MR02 - INPUT X OUTPUT

```
000001_disappointed_1
000001_federal_1
000001_firm_1
000001_mogul_1
000001_newall_1
000001_parent_1
000001_representing_1
000001_stricken_1
000001_talks_1
000001_turner_1
000001_unions_1
000001_workers_1
000002_announced_1
000002_ansari_1
000002_canada_1
000002_competing_1
000002_contest_1
000002_flight_1
000002_funded_1
000002_launch_1
```

```
aakash_004896_1_1
aapl_006273_2_5
aapl_004664_1_5
aapl_001080_1_5
aapl_000012_1_5
aapl_004195_1_5
aaron_000261_1_5
aaron_006460_1_5
aaron_002222_1_5
aaron_006754_1_5
aaron_006434_1_5
ab_001380_1_2
ab_001675_1_2
ababa_004372_1_2
ababa_006852_1_2
abandon_003333_1_7
abandon_001710_1_7
abandon_007269_1_7
abandon_001279_1_7
abandon_002277_1_7
```

MAPREDUCE 03 – CONTAGEM DE DOCS

“numeroDeDocs”

qtdDocs

- Recebemos o arquivo original e fazemos a contagem de documentos existentes no arquivo.

MR03 - INPUT X OUTPUT

```
000001,3,Fears for T N pension after talks,Unions representing workers at Turner   Newall say they are 'disappointed' aft
000002,4,The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com),"SPACE.com - TORONTO, Car
000003,4,Ky. Company Wins Grant to Study Peptides (AP),"AP - A company founded by a chemistry researcher at the Universit
000004,4,Prediction Unit Helps Forecast Wildfires (AP),"AP - It's barely dawn when Mike Fitzpatrick starts his shift with
000005,4,Calif. Aims to Limit Farm-Related Smog (AP),"AP - Southern California's smog-fighting agency went after emissior
000006,4,Open Letter Against British Copyright Indoctrination in Schools,"The British Department for Education and Skills
000007,4,Loosing the War on Terrorism,"\\""Sven Jaschan, self-confessed author of the Netsky and Sasser viruses, is\respo
000008,4,"FOAFKey: FOAF, PGP, Key Distribution, and Bloom Filters",\\FOAF/LOAF  and bloom filters have a lot of interesti
000009,4,E-mail scam targets police chief,"Wiltshire Police warns about ""phishing"" after its fraud squad chief was targ
000010,4,"Card fraud unit nets 36,000 cards","In its first two years, the UK's dedicated card fraud unit, has recovered
000011,4,Group to Propose New High-Speed Wireless Format," LOS ANGELES (Reuters) - A group of technology companies  inclu
000012,4,"Apple Launches Graphics Software, Video Bundle", LOS ANGELES (Reuters) - Apple Computer Inc.&lt;AAPL.O&gt; on
000013,4,Dutch Retailer Beats Apple to Local Download Market," AMSTERDAM (Reuters) - Free Record Shop, a Dutch music  ret
000014,4,Super ant colony hits Australia,"A giant 100km colony of ants  which has been discovered in Melbourne, Australia
000015,4,Socialites unite dolphin groups,"Dolphin groups, or ""pods"", rely on socialites to keep them from collapsing,
000016,4,Teenage T. rex's monster growth,Tyrannosaurus rex achieved its massive size due to an enormous growth spurt duri
000017,4,Scientists Discover Ganymede has a Lumpy Interior,"Jet Propulsion Lab -- Scientists have discovered irregular lu
000018,4,Mars Rovers Relay Images Through Mars Express,"European Space Agency -- ESAs Mars Express has relayed pictures f
000019,4,Rocking the Cradle of Life,"When did life begin? One evidential clue stems from the fossil records in Western Au
000020,4,"Storage, servers bruise HP earnings","update Earnings per share rise compared with a year ago, but company miss
```

```
007590,2,The Newest Hope ; Marriage of Necessity Just Might Work Out,"NEW YORK - The TV lights were on, the cameras rolle
007591,2,Saban hiring on hold,"DAVIE - The Dolphins want Nick Saban, and the LSU coach could be on his way. Although LSU
007592,1,Bosnian-Serb prime minister resigns in protest against U.S. sanctions (Canadian Press),"Canadian Press - BANJA LU
007593,1,Historic Turkey-EU deal welcomed,The European Union's decision to hold entry talks with Turkey receives a widespr
007594,2,Mortaza strikes to lead superb Bangladesh rally,"Paceman Mashrafe Mortaza claimed two prize scalps, including Sac
007595,1,Powell pushes diplomacy for N. Korea,"WASHINGTON -- Outgoing Secretary of State Colin L. Powell said yesterday he
007596,1,Around the world,"Ukrainian presidential candidate Viktor Yushchenko was poisoned with the most harmful known dio
007597,2,Void is filled with Clement,"With the supply of attractive pitching options dwindling daily -- they lost Pedro Ma
007598,2,Martinez leaves bitter,"Like Roger Clemens did almost exactly eight years earlier, Pedro Martinez has left the Re
007599,3,5 of arthritis patients in Singapore take Bextra or Celebrex &lt;b&gt;...&lt;/b&gt;,SINGAPORE : Doctors in the Un
007600,3,EBay gets into rentals,"EBay plans to buy the apartment and home rental service Rent.com for \$415 million, addi
```

```
1  numeroDeDocs  7600
```

MAPREDUCE 04

idDoc_termo

TFIDF

- Realizamos a leitura do output do mapreduce anterior que contém a quantidade de documentos do arquivo original.
- De posse de todos os dados, calculamos o TFIDF.

$$TFIDF_{t,d} = tf_d \times idf_t \quad (1)$$

tf_d (term_frequency) = número de vezes que o termo t aparece no documento d

$$idf_t \text{ (inverse document frequency)} = \log_{10}\left(\frac{N}{(1 + df_t)}\right) \quad (2)$$

em que df_t = número de documentos em que o termo t aparece

e N é o número total de documentos.

MR04 - INPUT X OUTPUT

```
aakash_004896_1_1
aapl_006273_2_5
aapl_004664_1_5
aapl_001080_1_5
aapl_000012_1_5
aapl_004195_1_5
aaron_000261_1_5
aaron_006460_1_5
aaron_002222_1_5
aaron_006754_1_5
aaron_006434_1_5
ab_001380_1_2
ab_001675_1_2
ababa_004372_1_2
ababa_006852_1_2
abandon_003333_1_7
abandon_001710_1_7
abandon_007269_1_7
abandon_001279_1_7
abandon_002277_1_7
```

```
000001_disappointed;3.102662341897148
000001_federal;1.6849139398715576
000001_firm;1.98318650099035
000001_mogul;3.1818435879447726
000001_newall;3.4036923375611288
000001_parent;2.734685556602553
000001_representing;2.8394209071225665
000001_stricken;3.57978359661681
000001_talks;1.7770098713248346
000001_turner;3.0357155522665344
000001_unions;2.57978359661681
000001_workers;1.9723285734021416
000002_announced;1.5172016123886471
000002_ansari;2.9265710828414666
000002_canada;2.1818435879447726
000002_competing;2.8808135922807914
000002_contest;2.734685556602553
000002_flight;2.3894518984465187
000002_funded;3.1818435879447726
000002_launch;1.9943228671083095
```

MAPREDUCE 05 – ORDENAÇÃO POR TFIDF

idDoc_termo

TFIDF

- Com a lista de TFIDFs, convertemos o TFIDF em chave.
- Na fase do shuffle, a lista de TFIDFs será ordenada de forma crescente.
- Salvamos da mesma maneira que recebemos o dado, mas de forma ordenada.

MR05 - INPUT X OUTPUT

```
1   000001_disappointed;3.102662341897148
2   000001_federal;1.6849139398715576
3   000001_firm;1.98318650099035
4   000001_mogul;3.1818435879447726
5   000001_newall;3.4036923375611288
6   000001_parent;2.734685556602553
7   000001_representing;2.8394209071225665
8   000001_stricken;3.579783596616 81
9   000001_talks;1.7770098713248346
10  000001_turner;3.0357155522665344
11  000001_unions;2.57978359661681
12  000001_workers;1.9723285734021416
13  000002_announced;1.5172016123886471
14  000002_ansari;2.9265710828414666
15  000002_canada;2.1818435879447726
16  000002_competing;2.8808135922807914
17  000002_contest;2.734685556602553
18  000002_flight;2.3894518984465187
    000002_funded;3.1818435879447726
    000002_launch;1.9943228671083095
```

```
1   007227_reuters;0.9842873747912361
2   004398_reuters;0.9842873747912361
3   002673_reuters;0.9842873747912361
4   001367_reuters;0.9842873747912361
5   000287_reuters;0.9842873747912361
6   001894_reuters;0.9842873747912361
7   004025_reuters;0.9842873747912361
8   002618_reuters;0.9842873747912361
9   001483_reuters;0.9842873747912361
10  005951_reuters;0.9842873747912361
11  000692_reuters;0.9842873747912361
12  000259_reuters;0.9842873747912361
13  000384_reuters;0.9842873747912361
14  000164_reuters;0.9842873747912361
15  003580_reuters;0.9842873747912361
16  003709_reuters;0.9842873747912361
17  003147_reuters;0.9842873747912361
18  005921_reuters;0.9842873747912361
19  003334_reuters;0.9842873747912361
20  003378_reuters;0.9842873747912361
```

```
125661  004352_treo;11.91089442115539
125662  004918_psp;12.142862209066138
125663  001785_apple;12.572902543620145
125664  006369_quot;12.659375953840826
125665  005182_gt;12.744094898804523
125666  002587_gt;12.744094898804523
125667  005325_traffic;12.792971487734361
125668  005182_lt;12.853245944136892
125669  002587_lt;12.853245944136892
125670  005371_cvs;13.115014403811315
125671  006576_mamma;13.614769350244515
125672  002755_supernova;13.614769350244515
125673  000790_terrorists;13.834351199869772
125674  006307_blog;14.197104535612832
125675  006595_blog;14.197104535612832
125676  000008_pgp;14.31913438646724
125677  000986_skype;16.393768004764144
125678  000072_girafa;17.89891798308405
125679  000008_foaf;17.89891798308405
125680  006307_costello;32.21805236955129
125681  006307_abbott;34.03692337561129
```

CONCLUSÃO

- Neste projeto, realizamos o pré-processamento de uma base de textos utilizando o framework Hadoop e a abordagem de *MapReduce* para calcular o TFIDF de cada palavra.

OBRIGADO!