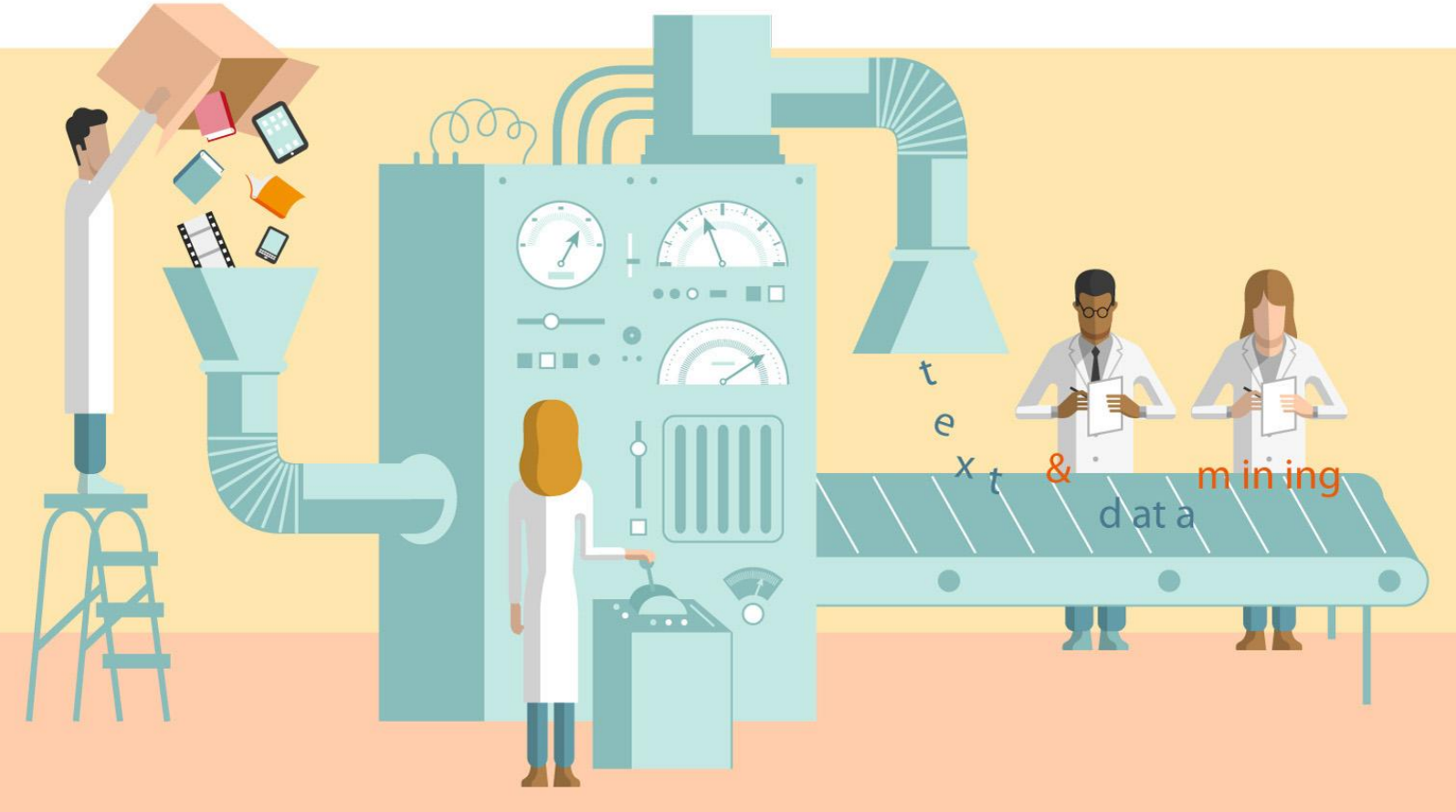
Conceitos de Data Mining e o uso da linguagem Python para a mineração de dados

Thiago Medeiros de Souza
Valmei Abreu Junior
João Paulo de Lima Barbosa

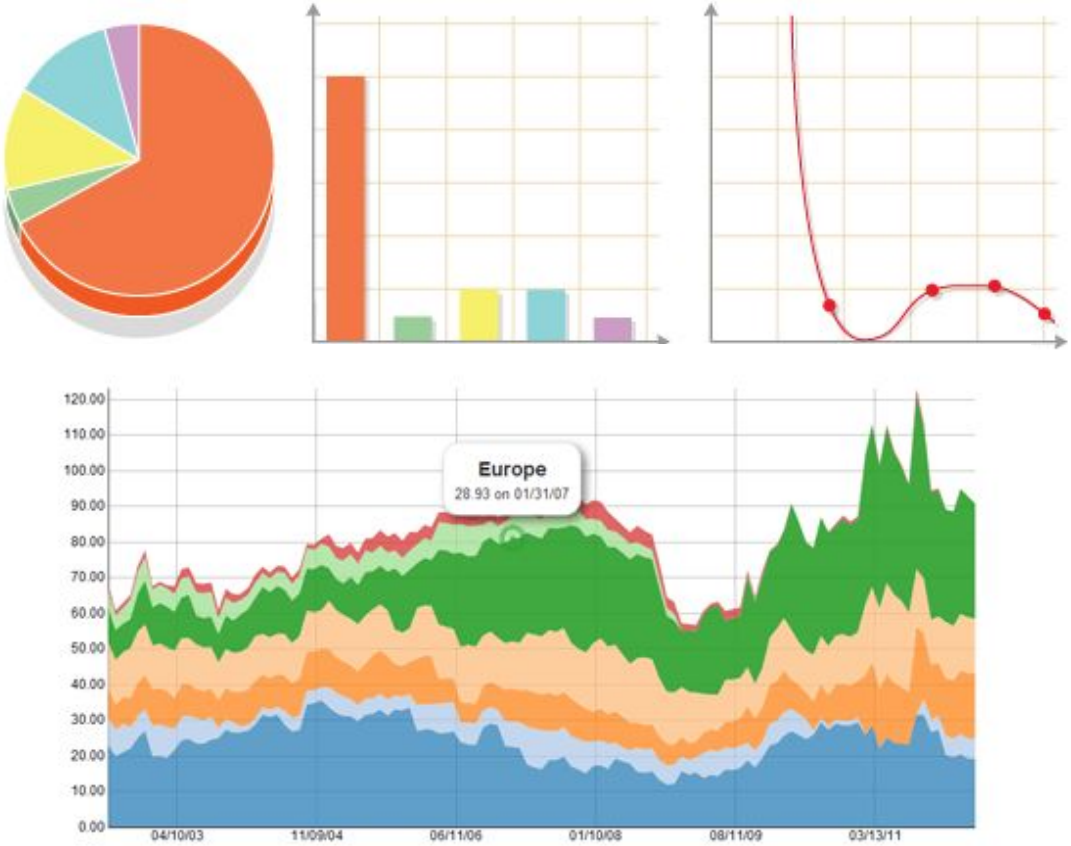
Conteúdo

- Introdução
- KDD e Data Mining
- KDD
- Data Mining
- Python
- Casos de Uso
- Considerações Finais

Introdução

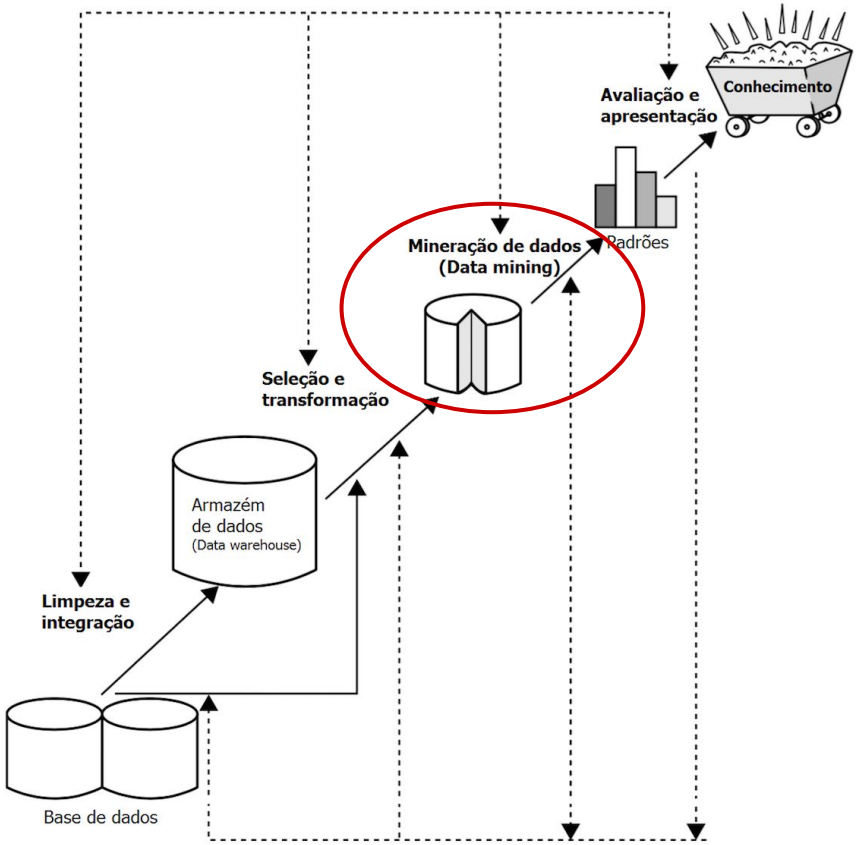


Introdução


Europe
28.93 on 01/31/07
120.00
110.00
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
04/10/03
11/09/04
06/11/06
01/10/08
08/11/09
03/13/11

KDD e Data Mining

KDD - Knowledge Discovery from Data



Data Mining

Métodos utilizados em Data Mining:

- Classificação
- Modelos de Relacionamento
- Análise de Agrupamento (Cluster)
- Sumarização
- Modelo de Dependência
- Regras de Associação
- Análise de Séries Temporais

Python

Python

- pandas


```
[{u'created_at': u'2015-11-06T13:23:50Z', u'trends': [{u'url': u'http://twitter.com/search?q=%23PSYPagyanig', u'query': u'%2
3PSYPagyanig', u'name': u'#PSYPagyanig', u'promoted_content': None}, {u'url': u'http://twitter.com/search?q=%237DaysUntilMIT
AM', u'query': u'%237DaysUntilMITAM', u'name': u'#7DaysUntilMITAM', u'promoted_content': None}, {u'url': u'http://twitter.co
m/search?q=%23BuenViernes', u'query': u'%23BuenViernes', u'name': u'#BuenViernes', u'promoted_content': None}, {u'url': u'ht
tp://twitter.com/search?q=%23%EB%8B%89%EB%84%A4%EC%9E%84%EC%9D%84_00%EC%9D%B4%EB%9D%BC_%EC%A7%80%EC%9D%80_%EC%9D%B4%EC%9C%A
0', u'query': u'%23%EB%8B%89%EB%84%A4%EC%9E%84%EC%9D%84_00%EC%9D%B4%EB%9D%BC_%EC%A7%80%EC%9D%80_%EC%9D%B4%EC%9C%A0', u'nam
e': u'#\ub2c9\ub124\uc784\uc744_00\uc774\ub77c_\uc9c0\uc740_\uc774\uc720', u'promoted_content': None}, {u'url': u'http://twi
tter.com/search?q=%23ManOnTheMoon', u'query': u'%23ManOnTheMoon', u'name': u'#ManOnTheMoon', u'promoted_content': None},
{u'url': u'http://twitter.com/search?q=%E3%82%8B%E3%82%8D%E5%89%A3', u'query': u'%E3%82%8B%E3%82%8D%E5%89%A3', u'name': u'\u
308b\u308d\u5263', u'promoted_content': None}, {u'url': u'http://twitter.com/search?q=Coldplay', u'query': u'Coldplay', u'na
me': u'Coldplay', u'promoted_content': None}, {u'url': u'http://twitter.com/search?q=%22DIA+13+%C3%89+DIA+DE+JUSTIN+BIEBER%2
2', u'query': u'%22DIA+13+%C3%89+DIA+DE+JUSTIN+BIEBER%22', u'name': u'DIA 13 \xc9 DIA DE JUSTIN BIEBER', u'promoted_conten
t': None}, {u'url': u'http://twitter.com/search?q=%E6%9D%BE%E3%81%95%E3%82%93%E8%A8%BA%E6%96%AD', u'query': u'%E6%9D%BE%E3%8
1%95%E3%82%93%E8%A8%BA%E6%96%AD', u'name': u'\u677e\u3055\u3093\u8a3a\u65ad', u'promoted_content': None}, {u'url': u'htt
p://twitter.com/search?q=%E3%82%A2%E3%83%8B%E3%83%A1%E3%82%AD%E3%83%A3%E3%83%A9%E5%8C%96', u'query': u'%E3%82%A2%E3%83%8B%E
3%83%A1%E3%82%AD%E3%83%A3%E3%83%A9%E5%8C%96', u'name': u'\u30a2\u30cb\u30e1\u30ad\u30e3\u30e9\u5316', u'promoted_content': N
one}], u'as_of': u'2015-11-06T13:30:05Z', u'locations': [{u'woeid': 1, u'name': u'Worldwide'}]}]

[{u'created_at': u'2015-11-06T13:28:41Z', u'trends': [{u'url': u'http://twitter.com/search?q=%23CarsonOnCNN', u'query': u'%2
3CarsonOnCNN', u'name': u'#CarsonOnCNN', u'promoted_content': None}, {u'url': u'http://twitter.com/search?q=%237DaysUntilMIT
AM', u'query': u'%237DaysUntilMITAM', u'name': u'#7DaysUntilMITAM', u'promoted_content': None}, {u'url': u'http://twitter.co
m/search?q=%23GetWeirdOutNow', u'query': u'%23GetWeirdOutNow', u'name': u'#GetWeirdOutNow', u'promoted_content': None}, {u'u
rl': u'http://twitter.com/search?q=%23FlavorAFilm', u'query': u'%23FlavorAFilm', u'name': u'#FlavorAFilm', u'promoted_conten
t': None}, {u'url': u'http://twitter.com/search?q=%23FridayFeeling', u'query': u'%23FridayFeeling', u'name': u'#FridayFeelin
g', u'promoted_content': None}, {u'url': u'http://twitter.com/search?q=%22Sagamore+Bridge%22', u'query': u'%22Sagamore+Bridg
e%22', u'name': u'Sagamore Bridge', u'promoted_content': None}, {u'url': u'http://twitter.com/search?q=%22Finally+Friday%2
2', u'query': u'%22Finally+Friday%22', u'name': u'Finally Friday', u'promoted_content': None}, {u'url': u'http://twitter.co
m/search?q=Coldplay', u'query': u'Coldplay', u'name': u'Coldplay', u'promoted_content': None}, {u'url': u'http://twitter.co
m/search?q=Chaffee', u'query': u'Chaffee', u'name': u'Chaffee', u'promoted_content': None}, {u'url': u'http://twitter.com/se
arch?q=%22Art+Modell%22', u'query': u'%22Art+Modell%22', u'name': u'Art Modell', u'promoted_content': None}], u'as_of': u'20
15-11-06T13:30:05Z', u'locations': [{u'woeid': 23424977, u'name': u'United States'}]}]
```

Python

- pandas


```
[
 {
  "created_at": "2015-11-06T13:23:50Z",
  "trends": [
   {
    "url": "http://twitter.com/search?q=%23PSYPagyanig",
    "query": "%23PSYPagyanig",
    "name": "#PSYPagyanig",
    "promoted_content": null
   },
   {
    "url": "http://twitter.com/search?q=%237DaysUntilMITAM",
    "query": "%237DaysUntilMITAM",
    "name": "#7DaysUntilMITAM",
    "promoted_content": null
   },
   {
    "url": "http://twitter.com/search?q=%23BuenViernes",
    "query": "%23BuenViernes",
    "name": "#BuenViernes",
    "promoted_content": null
   },
   {
    "url": "http://twitter.com/search?q=%23%EB%8B%89%EB%84%A4%EC%9E%84%EC%9D%84_00%EC%9D%B4%EB%9D%BC_%EC%A7%80%EC%9D%80_%E
C%9D%B4%EC%9C%A0",
    "query": "%23%EB%8B%89%EB%84%A4%EC%9E%84%EC%9D%84_00%EC%9D%B4%EB%9D%BC_%EC%A7%80%EC%9D%80_%EC%9D%B4%EC%9C%A0",
    "name": "#\ub2c9\ub124\uc784\uc744_00\uc774\ub77c_\uc9c0\uc740_\uc774\uc720",
    "promoted_content": null
   },
   {
    "url": "http://twitter.com/search?q=%23ManOnTheMoon",
    "query": "%23ManOnTheMoon",
    "name": "#ManOnTheMoon",
    "promoted_content": null
   },
   {
```

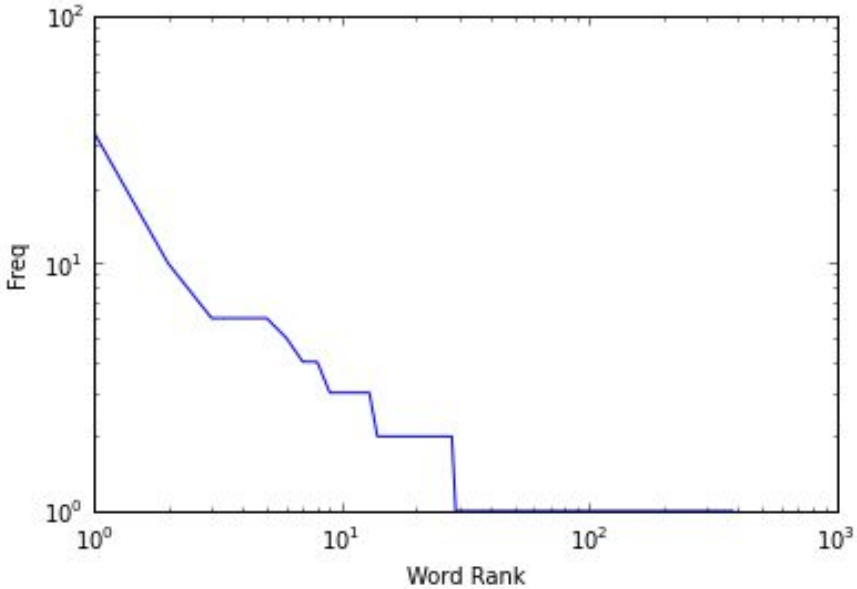
Python

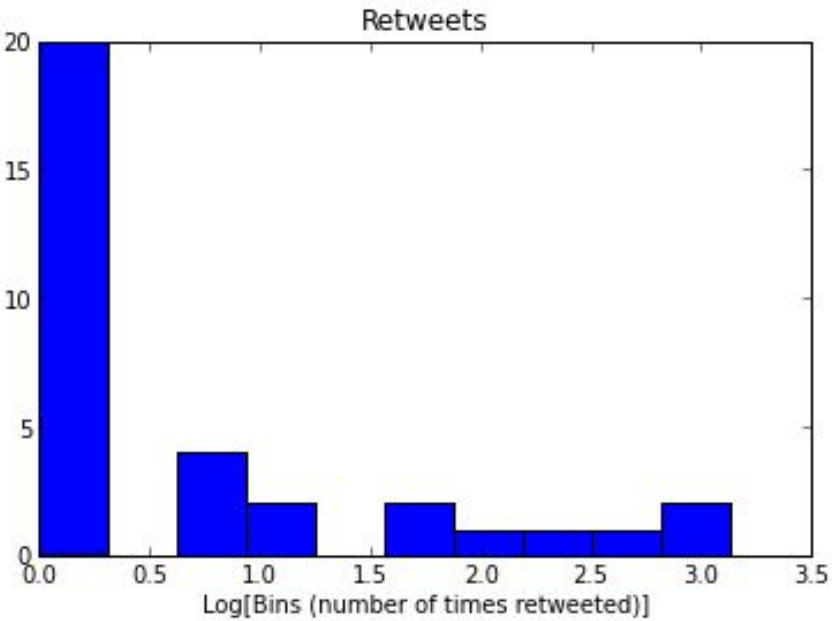
- NumPy

```
>>> a = np.ones((3, 3))  # reminder: (3, 3) is a tuple
>>> a
array([[ 1.,  1.,  1.],
       [ 1.,  1.,  1.],
       [ 1.,  1.,  1.]])
>>> b = np.zeros((2, 2))
>>> b
array([[ 0.,  0.],
       [ 0.,  0.]])
>>> c = np.eye(3)
>>> a = np.random.rand(4)       # uniform in [0, 1]
>>> a
array([ 0.95799151,  0.14222247,  0.08777354,  0.51887998])

>>> b = np.random.randn(4)      # Gaussian
>>> b
array([ 0.37544699, -0.11425369, -0.47616538,  1.79664113])

>>> np.random.seed(1234)        # Setting the random seed
```

Python

- matplotlib





Python

- IPython

Issues · ptwobrussell/Minin | Chef Fails on VMWare Prov | IPy IPython Dashboard | IPy Chapter 1 - Mining Twitter | mining-my-conclusion | Tw

Thiago

localhost:8888/37101b74-94a5-4aec-a02b-b7831ac5be4b

IP[y]: Notebook Chapter 1 - Mining Twitter (autosaved)

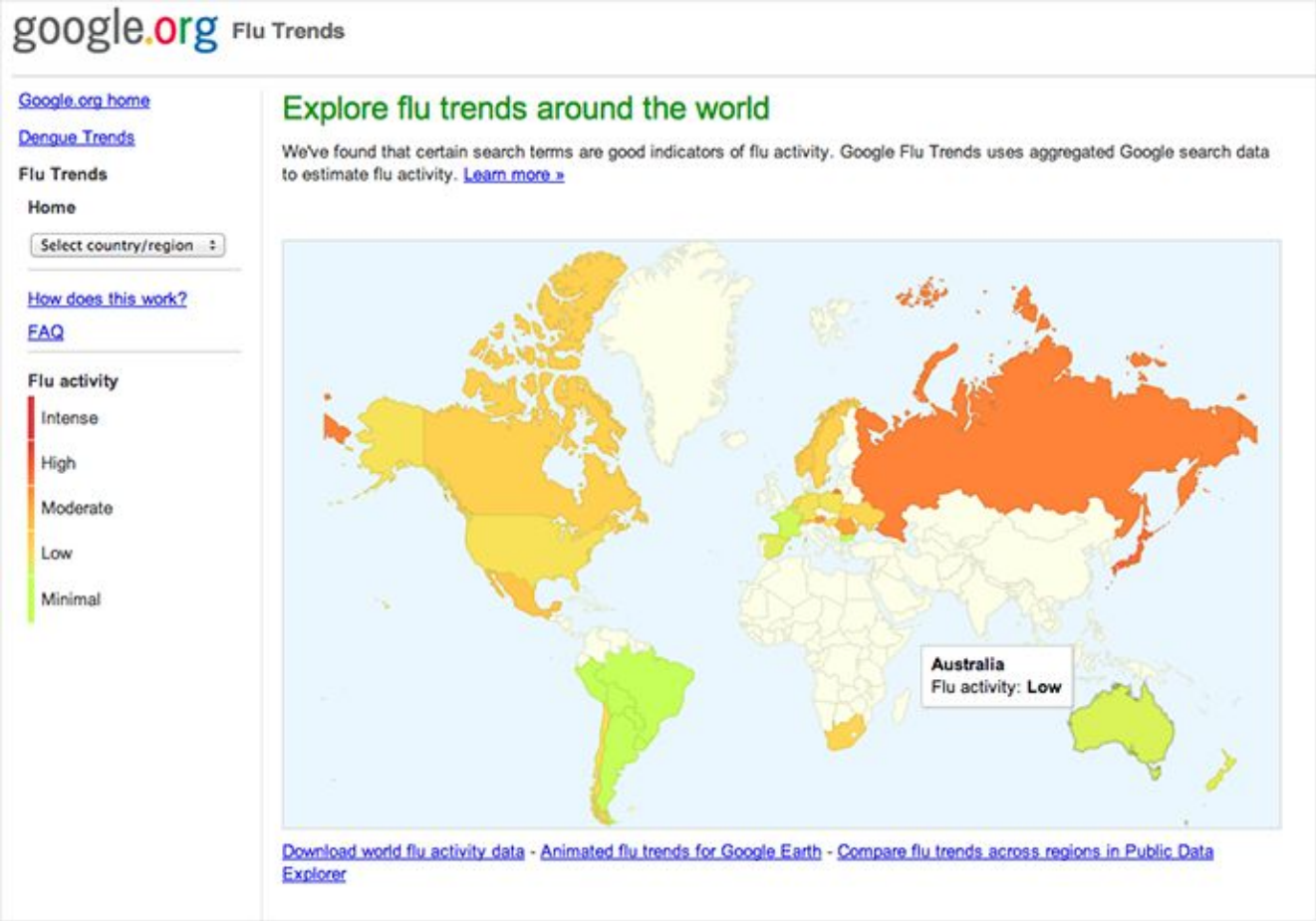
File Edit View Insert Cell Kernel Help

Code Cell Toolbar: None

```
    # Create a dictionary from next_results, which has the following form:
    # ?max_id=313519052523986943&q=NCAA&include_entities=1
    kwargs = dict([ kv.split('=') for kv in unquote(next_results[1:]).split("&") ])

    search_results = twitter_api.search.tweets(**kwargs)
    statuses += search_results['statuses']

# Show one sample search result by slicing the list...
print json.dumps(statuses[0], indent=1)
```

```
Length of statuses 100
Length of statuses 200
Length of statuses 300
Length of statuses 400
Length of statuses 500
{
 "contributors": null,
 "truncated": false,
 "text": "THE NEW COLDPLAY SONG. FUCK ME UP.",
 "is_quote_status": false,
 "in_reply_to_status_id": null,
 "id": 662623229723316224,
 "favorite_count": 0,
 "source": "<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
 "retweeted": false,
 "coordinates": null,
 "entities": {
  "symbols": [],
  "user_mentions": [],
  "hashtags": [],
  "urls": []
 },
 "in_reply_to_screen_name": null,
 "in_reply_to_user_id": null,
 "retweet_count": 0,
 "id_str": "662623229723316224",
```

Casos de Uso


google.org Flu Trends
Google.org home
Dengue Trends
Flu Trends
Home
Select country/region
How does this work?
FAQ
Flu activity
Intense
High
Moderate
Low
Minimal
Explore flu trends around the world
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »
Australia
Flu activity: Low
Download world flu activity data - Animated flu trends for Google Earth - Compare flu trends across regions in Public Data Explorer

Considerações Finais