PageRank avec PIG

Links: `notebook` (../_downloads/2015_page_rank.ipynb), `html` (../_downloads/2015_page_rank.html), `PDF` (../_downloads/2015_page_rank.pdf), `python` (../_downloads/2015_page_rank.py), `slides` (../_downloads/2015_page_rank.slides.html)

auteurs : *M. Amestoy M., A. Auffret*

L'algorithme PageRank (https://en.wikipedia.org/wiki/PageRank) propose une mesure de la pertinence d'un site. Il fut inventé par les fondateurs de google. L'implémentation proposée ici s'est appuyée sur celle proposée dans Data-Intensive Text Processing with MapReduce (http://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf), page 106. L'algorithme est d'abord appliqué sur un jeu de test (plus petit et permettant un développement rapide) puis à un jeu plus consistent : Google web graph (http://snap.stanford.edu/data/web-Google.html).

```
import pyensae
%nb_menu
```

Plan

- Connexion au cluster
- Création d'un petit jeu de données
- Récupération de données réelles
- Algorithme Page Rank
- Avec les données Google

Connexion au cluster

```
import pyquickhelper
params={"blob_storage":"",
        "password1":"",
        "hadoop_server":"",
        "password2":"",
        "username":""}
pyquickhelper.ipythonhelper.open_html_form(params=params,title="server + hadoop +
 credentials", key_save="blobhp")
```

server + hadoop + credentials

blob_storage

hadoop_server

password1

password2

username

Ok

```
import pyensae
blobstorage = blobhp["blob_storage"]
blobpassword = blobhp["password1"]
hadoop_server = blobhp["hadoop_server"]
hadoop_password = blobhp["password2"]
username = blobhp["username"]
client, bs =  %hd_open
client, bs
```

```
(<pyensae.remote.azure_connection.AzureClient at 0x86b2550>,
 <azure.storage.blobservice.BlobService at 0x86b2518>)
```

Création d’un petit jeu de données

On crée un set de données pour tester l’algorithme. (en reprenant celui présenté dans l’article)

```
with open("DataTEST.txt", "w") as f :
    f.write("1"+"\t"+"2"+"\n"+"1"+"\t"+"4"+"\n"+"2"+"\t"+"3"+"\n"+"2"+"\t"+"5"+"\
n"+"3"+"\t"+"4"+"\n"+"4"+"\t"+"5"+"\n"+"5"+"\t"+"3"+"\n"+"5"+"\t"+"1"+"\n"+"5"+"\
t"+"2")
```

```
import pandas
df = pandas.read_csv("DataTEST.txt", sep="\t",names=["Frm","To"])
df
```

	Frm	To
0	1	2
1	1	4
2	2	3
3	2	5
4	3	4
5	4	5
6	5	3
7	5	1
8	5	2

On importe ce graphe:

```
%blob_up DataTEST.txt /$PSEUDO/Data/DataTEST.txt
```

```
'$PSEUDO/Data/DataTEST.txt'
```

On vérifie que les données ont bien été chargées:

```
%blob_ls /$PSEUDO/Data/
```

	name	last_modified	content_type	content_length	blob_type
0	xavier/Data/DataTEST.txt	Tue, 14 Jul 2015 20:41:01 GMT	application/octet-stream	43	BlockBlob

Récupération de données réelles

On fait de même avec les données réelles : Google web graph (http://snap.stanford.edu/data/web-Google.html)

```
pyensae.download_data("web-Google.txt.gz", url="http://snap.stanford.edu/data/")
```

```
downloading of  http://snap.stanford.edu/data/web-Google.txt.gz  to  web-Google.t
xt.gz
```

```
['.\web-Google.txt']
```

```
%head web-Google.txt
```

```
# Directed graph (each unordered pair of nodes is saved once): web-Google.txt
# Webgraph from the Google programming contest, 2002
# Nodes: 875713 Edges: 5105039
# FromNodeId        ToNodeId
0   11342
0   824020
0   867923
0   891835
11342       0
11342       27469
```

On filtre les premières lignes.

```
with open("web-Google.txt", "r") as f:
    with open("DataGoogle.txt", "w") as g:
        for line in f:
            if not line.startswith("#"):
                g.write(line)
```

```
%head DataGoogle.txt
```

```
0   11342
0   824020
0   867923
0   891835
11342       0
11342       27469
11342       38716
11342       309564
11342       322178
11342       387543
```

```
%blob_up DataGoogle.txt /$PSEUDO/Data/DataGoogle.txt
```

```
'$PSEUDO/Data/DataGoogle.txt'
```

```
%blob_ls /$PSEUDO/Data/
```

	name	last_modified	content_type	content_length	blob_type
0	xavier/Data/DataGoogle.txt	Tue, 14 Jul 2015 21:10:13 GMT	application/octet-stream	75379926	BlockBlob
1	xavier/Data/DataTEST.txt	Tue, 14 Jul 2015 20:41:01 GMT	application/octet-stream	43	BlockBlob

Algorithme Page Rank

Initialisation de la table

```
%%PIG Creation_Graph.pig
Arcs = LOAD '$CONTAINER/$PSEUDO/Data/$path'
       USING PigStorage('\t')
       AS (frm:int,to:int);
GrSort = GROUP Arcs BY frm;
deg_sort = FOREACH GrSort
           GENERATE COUNT(Arcs) AS degs, Arcs , group AS ID;
GrEntr = GROUP Arcs BY to;
GrFin= JOIN deg_sort BY ID,
            GrEntr BY group;
N = FOREACH (group GrSort ALL)
    GENERATE COUNT(GrSort);
Pr = FOREACH GrFin
     GENERATE deg_sort::ID AS ID , (float) 1 / (float)N.$0 AS PageRank;
PageRank = JOIN GrFin BY deg_sort::ID,
                Pr BY ID;
STORE PageRank
INTO '$CONTAINER/$PSEUDO/Projet/SortTest.txt'
USING PigStorage('\t') ;
```

```
client.pig_submit(bs,
                  client.account_name,
                  "Creation_Graph.pig",
                  params=dict(path="DataTEST.txt"),
                  stop_on_failure=True)
```

```
{'id': 'job_1435385350894_0001'}
```

```
st = %hd_job_status job_1435385350894_0001
st["id"],st["percentComplete"],st["status"]["jobComplete"]
```

```
('job_1435385350894_0001', '100% complete', True)
```

```
%tail_stderr job_1435385350894_0001 10
```

```
2015-07-14 21:15:00,763 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:15:00,935 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:15:01,013 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:15:01,858 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:15:01,936 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:15:02,155 [main] WARN  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - No FileSystem for scheme: wasb. Not creating
 success file
2015-07-14 21:15:02,155 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:15:02,233 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:15:02,452 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
```

Itérations

On crée une macro pour répéter les iterations.

```
%%PIG iteration.pig
gr = LOAD '$CONTAINER/$PSEUDO/Projet/SortTest.txt'
     USING PigStorage('\t')
     AS (DegS:long,Asort:{(frm: int,to: int)},Noeud:int,Noeud2:int,Aent:{(frm: int,to: int)},ID: int,PageRank: float);
Arcs = LOAD '$CONTAINER/$PSEUDO/Data/DataTEST.txt'
       USING PigStorage('\t')
       AS (frm:int,to:int);
Graph = FOREACH gr
        GENERATE Noeud , DegS, PageRank AS Pinit, PageRank,  PageRank/ (float) DegS AS Ratio;

DEFINE my_macro(G,A,ALP) RETURNS S {
    Gi= FOREACH $G GENERATE Noeud , Ratio;
    GrEntr = JOIN $A BY frm , Gi BY Noeud ;
    Te = GROUP GrEntr BY to;
    so = FOREACH Te GENERATE SUM(GrEntr.Ratio) AS Pr, group AS ID;
    tu = JOIN $G BY Noeud, so BY ID;
    sort = FOREACH tu GENERATE Noeud , DegS, Pinit, $ALP*Pinit+(1-$ALP)*Pr AS PageRank;
    $S = FOREACH sort GENERATE Noeud , DegS, Pinit, PageRank, PageRank/ (float) DegS AS Ratio;
}
Ite1 = my_macro(Graph,Arcs,$alpha);
Ite2 = my_macro(Ite1,Arcs,$alpha);
Ite3 = my_macro(Ite2,Arcs,$alpha);
Ite4 = my_macro(Ite3,Arcs,$alpha);
Ite5 = my_macro(Ite4,Arcs,$alpha);
Ite6 = my_macro(Ite5,Arcs,$alpha);
Ite7 = my_macro(Ite6,Arcs,$alpha);
Ite8 = my_macro(Ite7,Arcs,$alpha);
Dump Ite1;
dump Ite8;
```

```
jid = client.pig_submit(bs,
                  client.account_name,
                  "iteration.pig",
                  params=dict(alpha="0"),
                  stop_on_failure=True )
jid
```

```
{'id': 'job_1435385350894_0006'}
```

```
st = %hd_job_status job_1435385350894_0006
st["id"],st["percentComplete"],st["status"]["jobComplete"]
```

```
('job_1435385350894_0006', '100% complete', True)
```

```
%tail_stderr job_1435385350894_0006 20
```

```
2015-07-14 21:39:13,398 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:39:13,572 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:39:13,650 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:39:13,822 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:39:13,884 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:39:14,575 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:39:14,653 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:39:14,809 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:39:14,903 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:39:15,512 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:39:15,590 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:39:15,746 [main] WARN  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - No FileSystem for scheme: wasb. Not creating
 success file
2015-07-14 21:39:15,746 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:39:15,825 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:39:15,965 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
2015-07-14 21:39:15,965 [main] INFO  org.apache.hadoop.conf.Configuration.depreca
tion - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-07-14 21:39:15,965 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key
 [pig.schematuple] was not set... will not generate code.
2015-07-14 21:39:16,012 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileIn
putFormat - Total input paths to process : 1
2015-07-14 21:39:16,012 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.util.MapRedUtil - Total input paths to process : 1
```

OUT:

```
(1,2,0.2,0.06666667014360428,0.03333333507180214)
(2,2,0.2,0.1666666716337204,0.0833333358168602)
(3,1,0.2,0.1666666716337204,0.1666666716337204)
(4,1,0.2,0.30000000447034836,0.30000000447034836)
(5,3,0.2,0.30000000447034836,0.10000000149011612)
(1,2,0.2,0.10671296522573188,0.05335648261286594)
(2,2,0.2,0.15640432448816244,0.07820216224408122)
(3,1,0.2,0.18429784359479393,0.18429784359479393)
(4,1,0.2,0.23171296837325728,0.23171296837325728)
(5,3,0.2,0.3208719206697963,0.1069573068899321)
```

On peut alors s’intéresser aux vraies données !

Avec les données Google

initialisation

```
%%PIG Creation_Graph2.pig
Arcs = LOAD '$CONTAINER/$PSEUDO/Data/$path'
       USING PigStorage('\t')
       AS (frm:int,to:int);
GrSort = GROUP Arcs BY frm;
deg_sort = FOREACH GrSort
           GENERATE COUNT(Arcs) AS degs, Arcs , group AS ID;
GrEntr = GROUP Arcs BY to;
GrFin = JOIN deg_sort BY ID,
        GrEntr BY group;
N = FOREACH (GROUP GrSort ALL)
    GENERATE COUNT(GrSort);
Pr = FOREACH GrFin
     GENERATE deg_sort::ID AS ID , (float) 1 / (float)N.$0 AS PageRank;
PageRank = JOIN GrFin BY deg_sort::ID, Pr BY ID;
STORE PageRank
INTO '$CONTAINER/$PSEUDO/Projet/SortGoogle.txt'
USING PigStorage('\t') ;
```

```
client.pig_submit(bs, client.account_name, "Creation_Graph2.pig", params=dict(pat
h="DataGoogle.txt"), stop_on_failure=True )
```

```
{'id': 'job_1435385350894_0037'}
```

```
st = %hd_job_status job_1435385350894_0037
st["id"],st["percentComplete"],st["status"]["jobComplete"]
```

```
('job_1435385350894_0037', '100% complete', True)
```

```
%tail_stderr job_1435385350894_0037 20
```

```
Total records proactively spilled: 0

Job DAG:
job_1435385350894_0038      ->      job_1435385350894_0039,job_1435385350894_0040
,
job_1435385350894_0039      ->      job_1435385350894_0041,
job_1435385350894_0040      ->      job_1435385350894_0041,
job_1435385350894_0041



2015-07-14 21:50:12,957 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:50:13,066 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:50:13,223 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:50:13,285 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:50:13,457 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:50:13,535 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:50:13,756 [main] WARN  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - No FileSystem for scheme: wasb. Not creating
 success file
2015-07-14 21:50:13,756 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 21:50:13,832 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 21:50:14,301 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
```

```
%%PIG iteration2.pig
gr = LOAD '$CONTAINER/$PSEUDO/Projet/SortGoogle.txt'
    USING PigStorage('\t')
    AS (DegS:long,Asort:{(frm: int,to: int)},Noeud:int,Noeud2:int,Aent:{(frm: int
,to: int)},ID: int,PageRank: float);
Arcs = LOAD '$CONTAINER/$PSEUDO/Data/DataGoogle.txt'
       USING PigStorage('\t')
       AS (frm:int,to:int);
Graph = FOREACH gr
        GENERATE Noeud , DegS, PageRank AS Pinit, PageRank,  PageRank/ (float) De
gS AS Ratio;
DEFINE my_macro(G,A,ALP) RETURNS S {
    Gi= FOREACH $G GENERATE Noeud , Ratio;
    GrEntr = JOIN $A by frm , Gi by Noeud ;
    Te = GROUP GrEntr by to;
    so = FOREACH Te generate SUM(GrEntr.Ratio) AS Pr, group AS ID;
    tu = JOIN $G by Noeud, so by ID;
    sort = FOREACH tu GENERATE Noeud , DegS, Pinit, $ALP*Pinit+(1-$ALP)*Pr AS Pag
eRank;
    $S = FOREACH sort GENERATE Noeud , DegS, Pinit, PageRank, PageRank/ (float) D
egS AS Ratio;
}
Ite1 = my_macro(Graph,Arcs,$alpha);
Ite2 = my_macro(Ite1,Arcs,$alpha);
Ite3 = my_macro(Ite2,Arcs,$alpha);
Ite4 = my_macro(Ite3,Arcs,$alpha);
Ite5 = my_macro(Ite4,Arcs,$alpha);
Ite6 = my_macro(Ite5,Arcs,$alpha);
Ite7 = my_macro(Ite6,Arcs,$alpha);
Ite8 = my_macro(Ite7,Arcs,$alpha);
Dump Ite1;
dump Ite8;
STORE Ite8 INTO '$CONTAINER/$PSEUDO/Projet/PageRank.txt' USING PigStorage('\t') ;
```

```
client.pig_submit(bs,
                  client.account_name,
                  "iteration2.pig",
                  params=dict(alpha="0.5"),
                  stop_on_failure=True )
```

```
{'id': 'job_1435385350894_0042'}
```

```
st = %hd_job_status job_1435385350894_0042
st["id"],st["percentComplete"],st["status"]["jobComplete"]
```

```
('job_1435385350894_0042', '100% complete', True)
```

```
%tail_stderr job_1435385350894_0042 20
```

```
2015-07-14 23:14:59,966 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:00,122 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:00,200 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:00,341 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:00,419 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:00,575 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:00,638 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:00,794 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:00,872 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:01,044 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:01,107 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:01,278 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:01,357 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:01,513 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:01,591 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:02,162 [main] WARN  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - No FileSystem for scheme: wasb. Not creating
 success file
2015-07-14 23:15:02,162 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Conn
ecting to ResourceManager at headnodehost/100.74.20.101:9010
2015-07-14 23:15:02,225 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelega
te - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirectin
g to job history server
2015-07-14 23:15:02,381 [main] INFO  org.apache.pig.backend.hadoop.executionengin
e.mapReduceLayer.MapReduceLauncher - Success!
```

OUT:

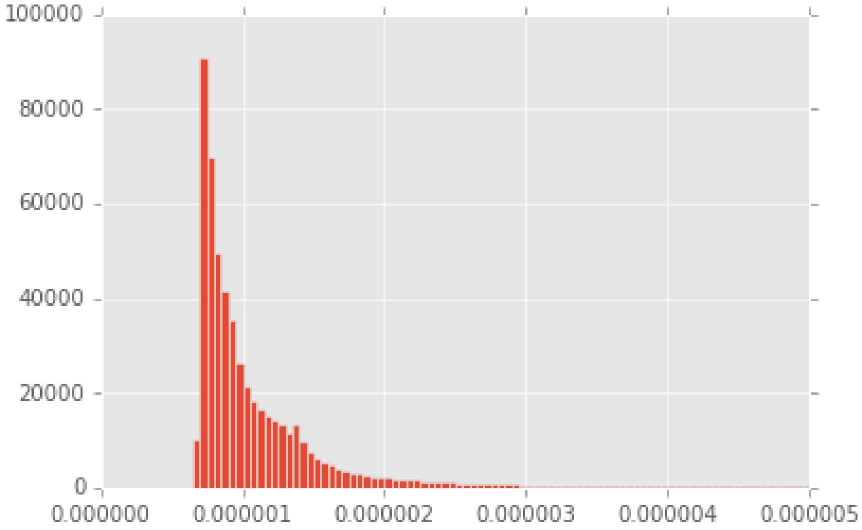
```
(916395,10,1.3523492E-6,1.061582549461105E-6,1.0615825494611051E-7)
(916397,4,1.3523492E-6,7.852355363200487E-7,1.9630888408001218E-7)
(916399,4,1.3523492E-6,7.042943628481174E-7,1.7607359071202936E-7)
(916400,4,1.3523492E-6,1.0821261274423128E-6,2.705315318605782E-7)
(916403,20,1.3523492E-6,7.202791281919268E-7,3.601395640959634E-8)
(916406,4,1.3523492E-6,7.00859985440611E-7,1.7521499636015275E-7)
(916407,18,1.3523492E-6,9.929362758159058E-7,5.516312643421699E-8)
(916408,1,1.3523492E-6,7.059236966261702E-7,7.059236966261702E-7)
(916409,1,1.3523492E-6,8.217128076315342E-7,8.217128076315342E-7)
(916415,8,1.3523492E-6,1.4456263606943261E-6,1.8070329508679077E-7)
(916416,3,1.3523492E-6,1.4650362621399523E-6,4.883454207133174E-7)
(916417,26,1.3523492E-6,4.652020819408753E-6,1.7892387766956742E-7)
(916418,13,1.3523492E-6,1.9168401786443277E-6,1.4744924451110214E-7)
(916419,13,1.3523492E-6,1.0262355087925307E-6,7.894119298404082E-8)
(916420,15,1.3523492E-6,1.0198868854927004E-6,6.79924590328467E-8)
(916421,3,1.3523492E-6,8.879735835276162E-7,2.959911945092054E-7)
(916423,7,1.3523492E-6,1.0184723897973568E-6,1.4549605568533667E-7)
(916424,4,1.3523492E-6,7.739901456530845E-7,1.9349753641327112E-7)
(916427,10,1.3523492E-6,1.133338628852173E-6,1.1333386288521729E-7)
```

```
%blob_downmerge /$PSEUDO/Projet/PageRank.txt PageRank.txt
```

```
'PageRank.txt'
```

```python
import pandas
import matplotlib as plt
plt.style.use('ggplot')
df = pandas.read_csv("PageRank.txt", sep="\t",names=["Node","OutDeg","Pinit", "PageRank", "k"])
df
df['PageRank'].hist(bins=100, range=(0,0.000005))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x8cbfda0>
```



```
df.sort("PageRank",ascending=False).head()
```

	Node	OutDeg	Pinit	PageRank	k
438534	751384	68	0.000001	0.000397	0.000006
351091	605856	22	0.000001	0.000381	0.000017
290922	504140	19	0.000001	0.000378	0.000020
310377	537039	27	0.000001	0.000373	0.000014
346122	597621	22	0.000001	0.000371	0.000017

```
%blob_close
```

```
True
```

Mis à jour le 2016-01-31.

Back to top