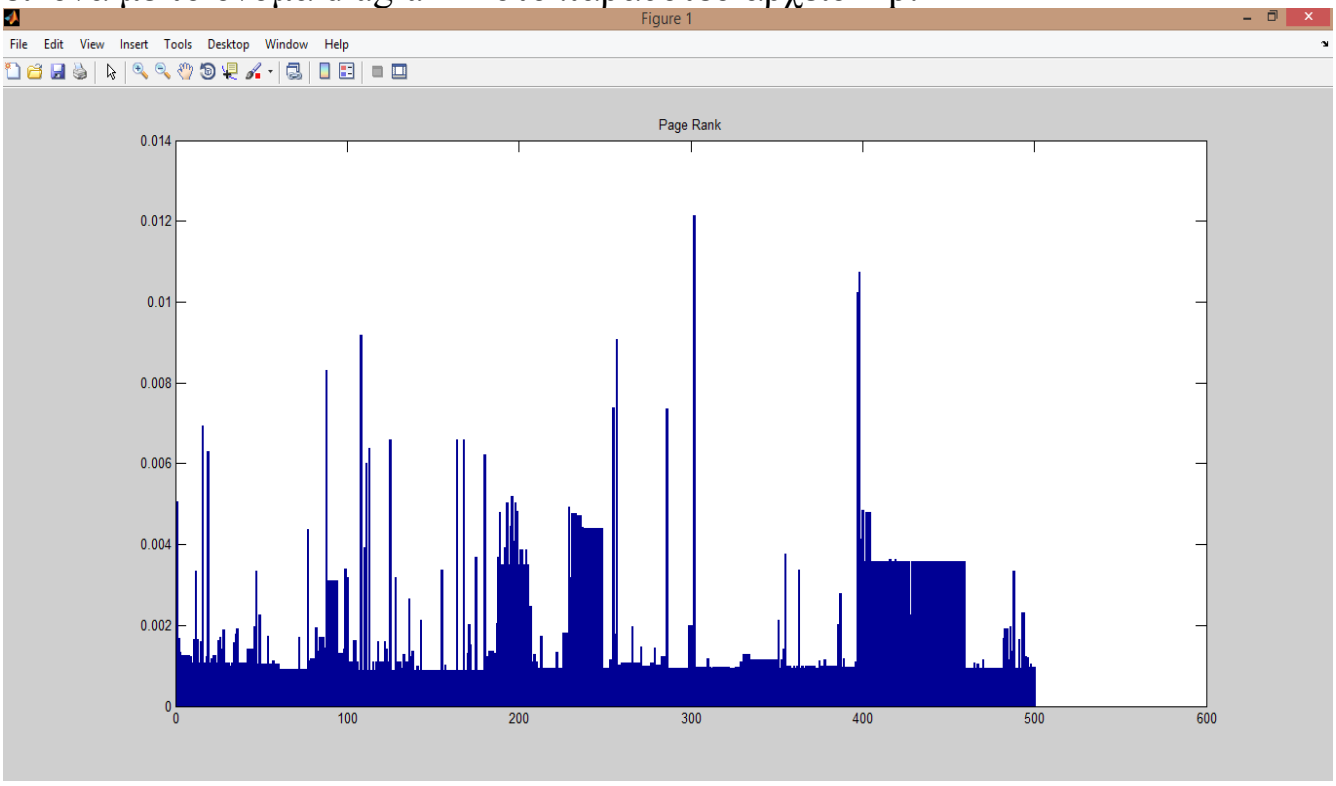
ΣΥΣΤΗΜΑΤΑ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΩΝ

Άσκηση για το σπίτι 3

Χατζηηλία Σοφία
Α.Μ.: 3100202

1. Για τον αλγόριθμο του PageRank έχω χρησιμοποιήσει έτοιμο κώδικα. Τον κώδικα αυτον τον βρήκα στην παρακάτω διευθυνση:
http://www.math.iit.edu/~fass/matlab/pagerank.m
Επέλεξα τον συγκεκριμένο κωδικα, γιατί κάνει ακριβώς αυτό που ζητάει το πρώτο ερώτημα. Δηλαδή, χρησιμοποιεί τον πίνακα με τα URLs U και τον πίνακα γειτνίασης G που παράγει η surfer, την οποία ήδη έχουμε από την 'Ασκηση για το σπίτι 1, για να υπολογίσει τα PageRank όλων των ιστοσελίδων, να τα επιστρέψει ταξινομημένα και να φτιάξει το διάγραμμα τους. Ο κώδικας για αυτό το ερώτημα βρίσκεται στο αρχείο ergasia3.m και το PageRank στο αρχείο pagerank.m. Στο παραδοτέο περιέχεται επίσης και η συνάρτηση surfer στο αρχείο surfer.m.
Για τα 500 URLs, έχουμε το παρακάτω διάγραμμα, το οποίο θα περιέχεται και σαν εικόνα με το όνομα diagramA στο παραδοτέο αρχείο zip:



Έχουμε βάλει να μην μας επιστρέφεται ο πίνακας x, δηλαδή ο πίνακας με τα PageRank, γιατι τα επόμενα ερωτήματα υλοποιούνται στο αρχείο pagerank.m, με την εντολή pagerank(U, G);. Αν θέλαμε να επιστρέφεται, τότε θα

χρησιμοποιούσαμε την εντολή x = pagerank(U, G);.
Τελικά, τυπωνεται το παρακάτω στην οθόνη μας:

```
    page-rank  in  out  url
 302   0.0121    7    0  http:
 398   0.0107    3    1  http://www.webdesign365.gr
 397   0.0102    2    1  http://www.condomlab.gr
 108   0.0092   51    0  http://images.iagora.net/js/meep-min-14.js
 257   0.0091    7    0  http://ogp.me/ns/fb#
  88   0.0083   25    1  http://ogp.me/ns#
 255   0.0074    6    0  http://www.
 286   0.0073   13    0  http://gmpg.org/xfn/11
  16   0.0069    7   23  http://www.directemploi.com
 164   0.0066   51    0  http://iagora.wordpress.com
 168   0.0066   51    2  http://i-hotels.iagora.com
 125   0.0066   51    2  http://studentsearch.iagora.com/login.html
 113   0.0064   50    0  http://images.iagora.net/images/registration/icecream-premium.png
  19   0.0063   22   24  http://www.best-masters.com
 180   0.0062    4    0  http://www.google-analytics.com/urchin.js
 111   0.0060   47    1  http://pagead2.googlesyndication.com/pagead/show_ads.js
 196   0.0052   20   18  http://www.directemploi.com/promo/19_offre-decouverte-offres-d-emploi-de-stage-et-d-alternance-i
   1   0.0051   19   30  http://www.aueb.gr
 193   0.0050   20    0  http://www.groupe-direct-performance.fr/la-societe.html
 198   0.0050   20    1  http://fr.viadeo.com/fr/company/direct-emploi
>>
```

2. Ο κώδικας για το 2ο ερώτημα είναι στο αρχειο pageank.m και είναι αυτός που φαίνεται παρακάτω.

```
   if nargout < 1
   [ignore,q] = sort(-x);
      ...
   %erotima 2o
   pos= randi(n); %position of A
   A = U(pos) %url
   x(pos) %pagerank of A
   k=1; m=1;
   disp('     page-rank  in  out  url        pagerank of A')
   while(k <= 10) & (x(q(k)) >= .005)
       if(c(k)~=0)
           prank = x(pos)+x(k)/(c(k)+1);
       else
           prank = x(pos) + x(k);
       end
       disp(sprintf(' %3.0f %8.4f %4.0f %4.0f  %s  %8.4f', ...
         k,x(k),r(k),c(k),U{k},prank))
       k=k+1;
   end
      ...
end
```

Αρχικά, βρίσκω την ιστοσελίδα A τυχαια από τον πίνακα με τα URLs U. Στην συνέχεια, για να βρω πόσο αυξάνεται το PageRank του A, αν εξασφαλίσω ένα σύνδεσμο προς τη σελίδα A, πέρνω τα πρώτα 10 PageRank από τον πίνακα x, με τα ταξινομημένα PageRank. Για κάθε στοιχείο i, προσθέτω το πηλίκο της διαίρεσης του PageRank του i με το out-degree του στοιχείου αυτού +1, καθώς τώρα αυτό το στοιχείο δειχνει και προς τον κόμβο A, με το PageRank που έχει ήδη ο κόμβος A. Τέλος, τυπώνεται το URL του τυχαίου κόμβου A και το PageRank του, η λίστα με τα δέκα URLs, τα in-degree τους, τα out-degree τους, τα PageRank τους και το καινούριο PageRank, που θα είχε ο A, αν ο αντίστοιχος κόμβος έδειχνε προς αυτόν. Το αποτέλεσμα φαίνεται παρακάτω.

```
A =

    'http://www.newyorker.de/en/company/community-projects/foundation'


ans =

    0.0035

     page-rank  in  out  url        pagerank of A
   1   0.0050   20   30  http://www.aueb.gr    0.0036
   2   0.0016    6    0  http://html5shiv.googlecode.com/svn/trunk/html5.js    0.0051
   3   0.0013    6   11  http://irakleitos.aueb.gr    0.0036
   4   0.0012    5    0  http://eclass.aueb.gr    0.0047
   5   0.0012    5    2  http://e-grammateia.aueb.gr/unistudent    0.0039
   6   0.0012    5    2  http://dias.aueb.gr:8100    0.0039
   7   0.0012    5    1  http://e-grammateia.aueb.gr:8080/cem-el.htm    0.0041
   8   0.0012    5   14  http://www.esnathens.gr    0.0035
   9   0.0012    4   28  http://www.aueb.gr/pages/organoseis/#AIESEC    0.0035
  10   0.0010    2    3  http://jobs.aueb.gr    0.0037
fx >>
```

Εδώ, η τυχαία σελίδα είναι αυτή που φαίνεται πάνω πάνω και το PageRank της φαίνεται αμέσως από κάτω. Ακολουθεί η λίστα με τα URL των 10 ιστοσελίδων με τις αντίστοιχες τιμές του PageRank της σελίδας A.

3. Ο κώδικας για αυτό το ερώτημα βρίσκεται στο αρχείο pagerank.m και είναι αυτος που φαίνεται παρακάτω.

```
if nargout < 1
   [ignore,q] = sort(-x);
      . . .
%erotima 3o
   %trim A
   Adomain = regexp(A, ...
```

```
      '([^/]*)(?=/[^/])','match');
   while(k <= n) & (x(q(k)) >= .005)
       %trim url
       Udomain = regexp(U(k), ...
            '([^/]*)(?=/[^/])','match');
       if(strcmp(Adomain(1),Udomain(1)))
           if(c(k)~=0)
               prank = x(pos)+x(k)/(c(k)+1);
           else
               prank = x(pos) + x(k);
           end
           disp(sprintf(' %3.0f %8.4f %4.0f %4.0f  %s  %8.4f', ...
               k,x(k),r(k),c(k),U{k},prank))
           k=n+1;
       else
           disp('URL with domain of A not found')
       end
       k=k+1;
   end
   ...
end
```

Αρχικα, κραταμε από το τυχαιο κόμβο A μονο το domain. Στην συνεχεια, κανουμε το ίδιο για κάθε ένα από τα 501 URLs που έχουμε. Αν βρω ένα URL με το ίδιο domain με αυτό του A, κάνω το k=n+1, για να σταματήσει το while. Όταν δεν βρίσκω URL για το domain αυτο εμφανίζεται μήνυμα `'URL with domain of A not found'`. Όταν βρίσκει ένα URL με το ίδιο domain, τότε υπολογίζεται το καινούριο PageRank, με τον ίδιο τρόπο που υπολογίζεται και στο 2ο ερώτημα, εξασφαλίζοντας ένα σύνδεσμο από το URL προς τον A. Τέλος, τυπώνεται το URL αυτο, το in-degree του, το out-degree του, το PageRank του και το πως αυξανεται το PageRank του A.