Πρώτη Σειρά Ασκήσεων - Εξόρυξη Δεδομένων

Όνομα: Χρύσα Τεριζή
ΑΜ: 2553
Ημερομηνία: 02/4/2017
Free passes: Κανένα δεν χρησιμοποίησα

Ερώτηση 1(Reservoir Sampling)

1. Έχω ένα αρχείο. Το ανοίγω και διαβάζω το περιεχόμενο του ανά γραμμές. Έχω τον αριθμό των γραμμών που θέλω να εκτυπώσω, έστω κ. Έχω μία λίστα έστω reservoirList μεγέθους κ και εδώ θα κρατάω ουσιαστικά τις τελικές γραμμές που θέλω να εκτυπώσω, δηλαδή τα Κ αντικείμενα από τα Ν. Επειδή θέλω να διατρέξω το αρχείο μία φορά, κάθε φορά που θα διαβάζω μία γραμμή θα την αποθηκεύω κατευθείαν στην λίστα μου. Κάποια στιγμή η λίστα θα έχει γεμίσει με τις κ πρώτες γραμμές του αρχείου. Τώρα, θα πάρω έναν τυχαίο ομοιόμορφο αριθμό έστω r από το [1 , κ + 1]. Αν η τιμή του αριθμού αυτού είναι < κ τότε θα αντικαταστήσω στην θέση r της λίστας μου την γραμμή που μόλις διάβασα δηλαδή την κ + 1. Και συνεχίζω με αυτήν την λογική μέχρι να τελειώσει η ανάγνωση του αρχείου.

2. Οι πρώτες κ γραμμές μπαίνουν κατευθείαν μέσα στην λίστα οπότε η πιθανότητα τους θα μπουν στην λίστα είναι 1. Για την γραμμή κ + 1 η πιθανότητα της να μπει στην λίστα είναι (κ / κ + 1), της γραμμής κ + 2 είναι (κ + 1 / κ + 2) κτλ. Οπότε έχουμε ότι μία γραμμή από το αρχείο μας έχει πιθανότητα να εμφανιστεί στην λίστα (κ / κ + 1)*(κ + 1 / κ + 1)*(κ + 2 / κ + 3)* ... *(κ – Ν / Ν) = κ / Ν.

3. Το αρχείο έχει όνομα sample.py. Το αρχείο input.txt είναι ένα απλό αρχείο με το οποίο τέσταρα το πρόγραμμα μου, περιέχει μέσα πληροφορίες για την Ελλάδα τις οποίες βρήκα από το Greece Wikipedia(https://en.wikipedia.org/wiki/Greece).

Τρόπος για να τρέξει το πρόγραμμα:
> python sample.py 10 < input.txt

Αποτελέσματα: Στην αρχή εμφανίζει τους αριθμούς των γραμμών που έχουν επιλεγεί και μετά τυπώνει τις γραμμές αυτές.

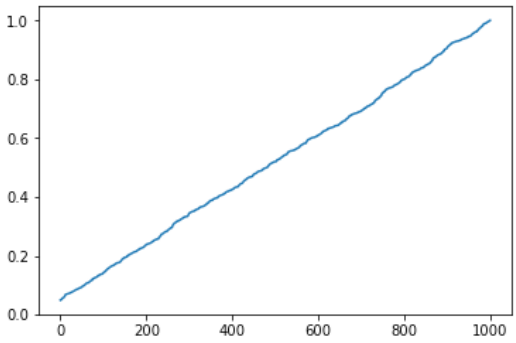
```
C:\Users\Chryssa\Documents\cse\8osemester\dataMining\assignments\assignment1\task1>python sample.py 10 < input.txt
[185, 152, 39, 297, 118, 368, 353, 141, 83, 64]
Greek called Demotic. Many of the educated elite saw this as a peasant dialect and were determined to restore the glories of Ancient Greek.

1815 onwards, to engage traditional strata of the Greek Orthodox world in their liberal nationalist cause.[79] The Filiki Eteria planned to

but collapsed violently around 1200 BC, during a time of regional upheaval known as the Bronze Age collapse.[32] This ushered in a period known

Representation through:[122]      embassy Γζô      embassy in another country

the United Kingdom in 1809 until their unification with Greece in 1864.[68]

dominated the country's political scene, and divided the country into two opposing groups. During parts of World War I, Greece had two

praised the fierceness of the Greek resistance. In an official notice released to coincide with the Greek national celebration of the Day of

class. These merchants came to dominate trade within the Ottoman Empire, establishing communities throughout the Mediterranean, the Balkans, and

religious practices were still in vogue in the late 4th century AD,[50] when they were outlawed by the Roman emperor Theodosius I in

cities in Asia and Africa.[40] Although the political unity of Alexander's empire could not be maintained, it resulted in the Hellenistic
```

Ερώτηση 2

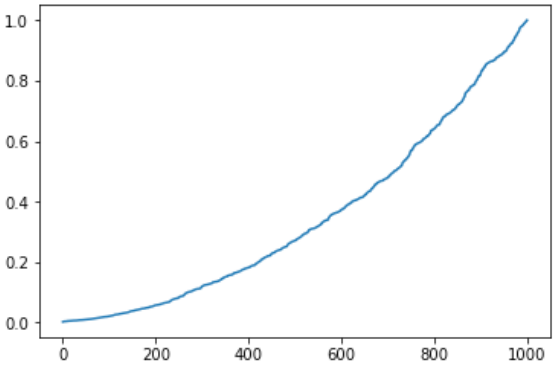
Αρχικά, προσπέλασα το αρχείο και δημιούργησα 3 διαφορετικές λίστες για την κάθε στήλη.

Σχέση ανάμεσα σε στήλες :

- A και B
  Έπειτα κανονικοποίησα τα δεδομένα των στηλών A και B. Έφτιαξα τις γραφικές τους παραστάσεις .

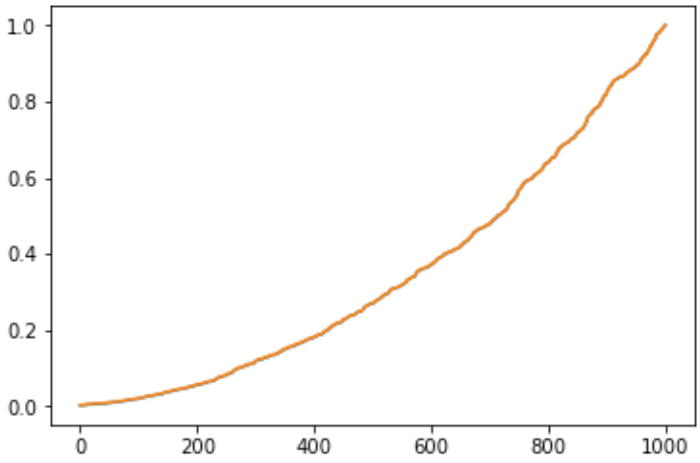


Στήλη A



Στήλη B

Παρατήρησα ότι η στήλη A αν την υψώσεις στην δύναμη 2 τότε οι γραφικές τους παραστάσεις θα ταυτίζονται και το MSE τους είναι 9.553926014086105e-07.

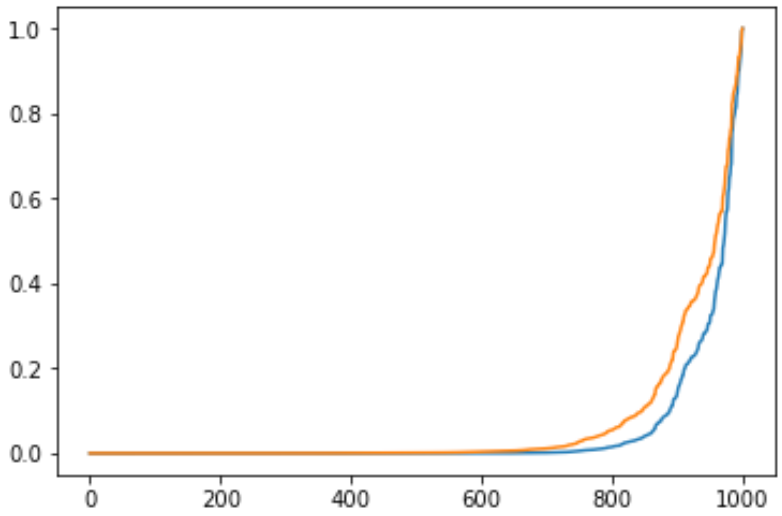


Στήλη A και B μαζί (ταυτίζονται)

Οπότε η τελική σχέση ανάμεσα στις στήλες A και B είναι B = A^2.

- A και C
  Από την γραφική παράσταση της στήλης C παρατήρησα ότι μοιάζει με την γραφική της e^(X). Οπότε προσπάθησα να φτιάξω τα δεδομένα της στήλης A με αυτόν τον τρόπο. Αρχικά η γραφική της e^(A) είχε κάποια απόκλιση στον άξονα των X οπότε δοκίμασα να προσθέσω από

το [0,100] και να βρω ποια έχει το ελάχιστο error. Βρέθηκε ότι ήτανε η τιμή 56 με απόκλιση 0.0019916702846461357.



Στήλες A και C

Οπότε η σχέση ανάμεσα στην στήλη A και C είναι C = e^(A) + 56.

Ερώτηση 3

Αρχικά ανοίγω μία φορά το αρχείο *twitter_dataset.txt* ώστε να βρω για κάθε tweet τα handles και τα hashtags που υπάρχουν σε αυτό. Ελέγχω για αρχή αν το tweet ξεκινάει με RT ώστε να το αγνοήσω και έπειτα καλώ την συνάρτηση getTagsHandles(x) η οποία μου επιστρέφει σε μία λίστα τα handles και τα hashtags τα οποία υπάρχουν μέσα στο tweet. Στο αρχείο *baskets.txt* αποθηκεύω τα “καλάθια” μου, κάθε γραμμή του αρχείου είναι και ένα “καλάθι”.

Έπειτα υπολογίζω κάποια στατιστικά σχετικά με τα καλάθια, τα handles και τα hashtags τα οποία είναι τα ακόλουθα:

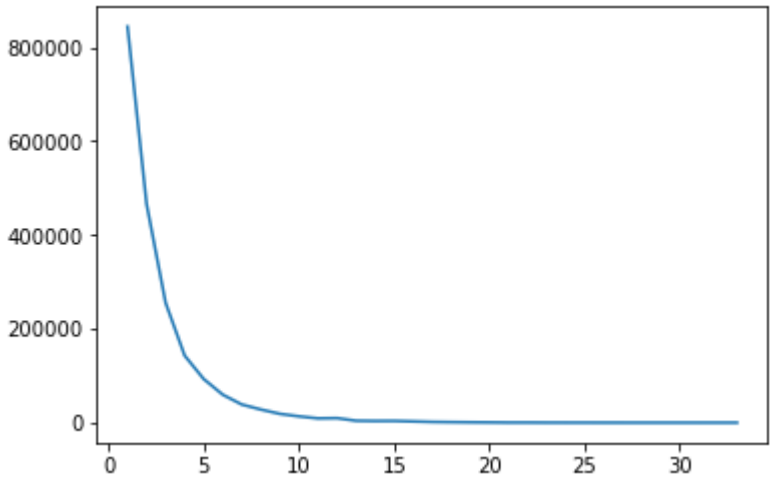
- Συνολικός αριθμός από καλάθια: 1997946
- Συνολικός αριθμός αντικειμένων που υπάρχουν μέσα σε όλα τα καλάθια: 648151
- Μέσο αριθμός από handles ή hashtags ανά καλάθι: 3.6284979674125326
- Συνολικός αριθμός από hashtags vs handles: 2548360 vs 2703237
- Συνολικός αριθμός από διακριτά hashtags vs handles: 163261 vs 484889

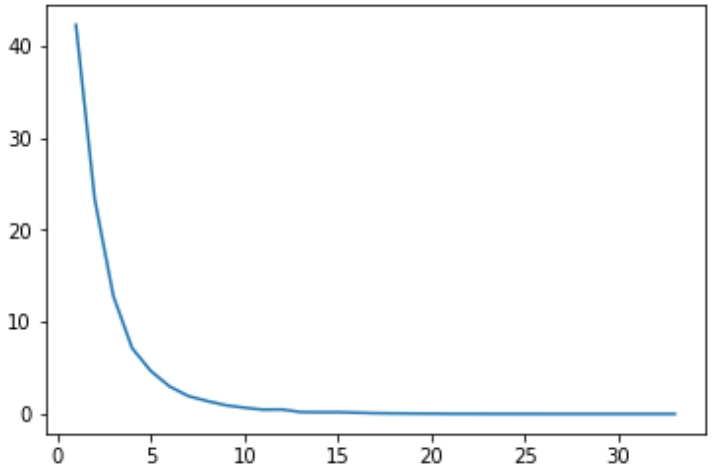
Πιο συγκεκριμένα για το μέγεθος των καλαθιών:

- Πλήθος καλαθιών με 1 αντικείμενο hashtag ή handle: 843818
- Πλήθος καλαθιών με 2 αντικείμενα hashtag ή handle: 466118
- Πλήθος καλαθιών με 3 αντικείμενα hashtag ή handle: 255370
- Πλήθος καλαθιών με 4 αντικείμενα hashtag ή handle: 142859
- Πλήθος καλαθιών με 5 αντικείμενα hashtag ή handle: 93032
- Πλήθος καλαθιών με 6 αντικείμενα hashtag ή handle: 59669
- Πλήθος καλαθιών με 7 αντικείμενα hashtag ή handle: 38701
- Πλήθος καλαθιών με 8 αντικείμενα hashtag ή handle: 28054

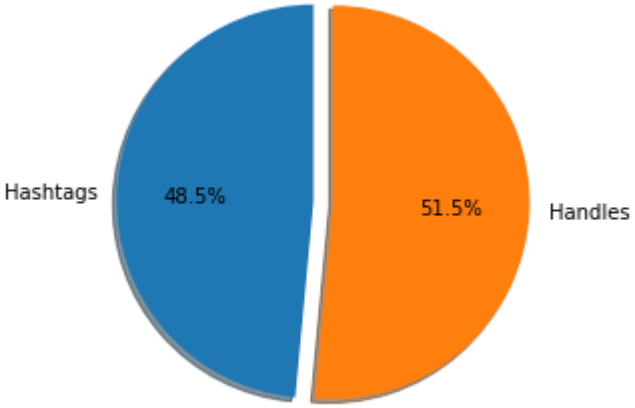
- Πλήθος καλαθιών με 9 αντικείμενα hashtag ή handle: 18828
- Πλήθος καλαθιών με 10 αντικείμενα hashtag ή handle: 13475
- Πλήθος καλαθιών με 11 αντικείμενα hashtag ή handle: 9155
- Πλήθος καλαθιών με 12 αντικείμενα hashtag ή handle: 9654
- Πλήθος καλαθιών με 13 αντικείμενα hashtag ή handle: 4093
- Πλήθος καλαθιών με 14 αντικείμενα hashtag ή handle: 3644
- Πλήθος καλαθιών με 15 αντικείμενα hashtag ή handle: 3743
- Πλήθος καλαθιών με 16 αντικείμενα hashtag ή handle: 2702
- Πλήθος καλαθιών με 17 αντικείμενα hashtag ή handle: 1620
- Πλήθος καλαθιών με 18 αντικείμενα hashtag ή handle: 1167
- Πλήθος καλαθιών με 19 αντικείμενα hashtag ή handle: 803
- Πλήθος καλαθιών με 20 αντικείμενα hashtag ή handle: 585
- Πλήθος καλαθιών με 21 αντικείμενα hashtag ή handle: 250
- Πλήθος καλαθιών με 22 αντικείμενα hashtag ή handle: 255
- Πλήθος καλαθιών με 23 αντικείμενα hashtag ή handle: 112
- Πλήθος καλαθιών με 24 αντικείμενα hashtag ή handle: 83
- Πλήθος καλαθιών με 25 αντικείμενα hashtag ή handle: 68
- Πλήθος καλαθιών με 26 αντικείμενα hashtag ή handle: 24
- Πλήθος καλαθιών με 27 αντικείμενα hashtag ή handle: 11
- Πλήθος καλαθιών με 28 αντικείμενα hashtag ή handle: 14
- Πλήθος καλαθιών με 29 αντικείμενα hashtag ή handle: 19
- Πλήθος καλαθιών με 30 αντικείμενα hashtag ή handle: 4
- Πλήθος καλαθιών με 31 αντικείμενα hashtag ή handle: 1
- Πλήθος καλαθιών με 32 αντικείμενα hashtag ή handle: 14
- Πλήθος καλαθιών με 33 αντικείμενα hashtag ή handle: 1

Γράφημα μέγεθος καλαθιών σε μορφή ποσοστού σε σχέση με το πλήθος των συνολικών καλαθιών.
Συμπέρασμα ότι το 40% των συνολικών καλαθιών έχουν 1 μόνο hashtag ή handle.

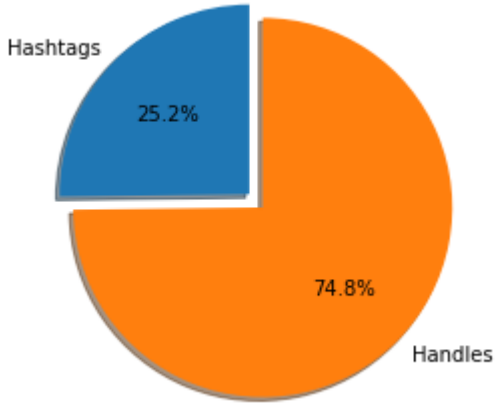




Γράφημα που δείχνει πόσα hashtags και πόσα handles έχουν χρησιμοποιηθεί συνολικά σε όλα τα tweets. Συμπέρασμα ότι περίπου ίσο πλήθος. Υπάρχουν διπλότυπα.

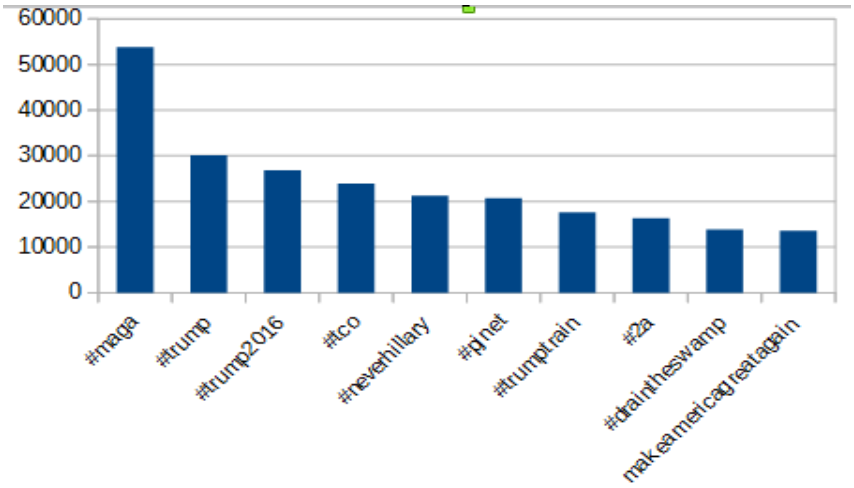


Γράφημα που δείχνει πόσα hashtags και πόσα handles έχουν χρησιμοποιηθεί συνολικά σε όλα τα tweets. Συμπέρασμα ότι έχουν χρησιμοποιηθεί περισσότερα διαφορετικά handles σε σχέση με τα hashtags όπου φαίνεται ότι από 48.5% έχουν γίνει 25.2%. Δεν υπάρχουν διπλότυπα.



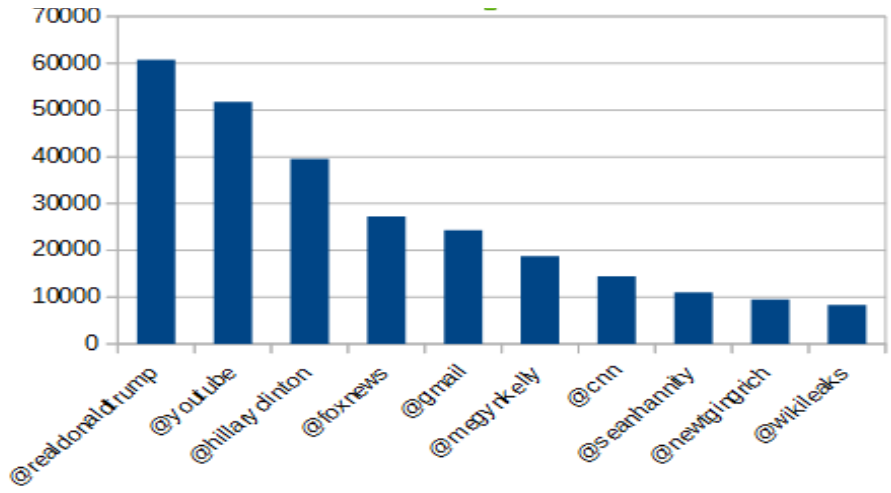
Τα 10 πιο συχνά hashtags είναι τα ακόλουθα σε φθίνουσα διάταξη:

1. ('#maga', 53777)
2. ('#trump', 30127)
3. ('#trump2016', 26846)
4. ('#tcot', 23892)
5. ('#neverhillary', 21264)
6. ('#pjnet', 20696)
7. ('#trumptrain', 17596)
8. ('#2a', 16314)
9. ('#draintheswamp', 13873)
10. ('#makeamericagreatagain', 13540)



Τα 10 πιο συχνά handles είναι τα ακόλουθα σε φθίνουσα διάταξη:

1. ('@realdonaldtrump', 60720)
2. ('@youtube', 51772)
3. ('@hillaryclinton', 39554)
4. ('@foxnews', 27241)
5. ('@gmail', 24293)
6. ('@megynkelly', 18764)
7. ('@cnn', 14429)
8. ('@seanhannity', 10997)
9. ('@newtgingrich', 9484)
10. ('@wikileaks', 8282)



<u>Maximal συχνά στοιχειοσύνολα</u>: Ο αλγόριθμος Apriori μας λέει ότι πρώτα θα βρω τα συχνά στοιχειοσύνολα μεγέθους 1, έπειτα από αυτά θα φτιάξω όλους τους συνδυασμούς μεγέθους 2 και θα βρω τα πιο συχνά, με τον ίδιο τρόπο θα συνεχίσω για να βρω για μέγεθος 3 κτλ. Maximal σημαίνει ότι κανένα υπερσύνολο δεν είναι συχνό. Συχνό εννοούμε ότι υπερβαίνει ένα κατώφλι το οποίο έχουμε θέσει εμείς.

Έχω υπολογίσει ταυτόχρονα τα συχνά στοιχειοσύνολα και από αυτά βρίσκω ποια είναι maximal. Αρχικά, ανοίγω το αρχείο με τα καλάθια *baskets.txt*. Έχω φτιάξει ένα λεξικό *d* όπου μέσα περιέχει σαν κλειδιά με 1 μέγεθος(ένα μόνο hashtag ή handle) και σαν τιμές σε αυτά τα κλειδιά είναι ένα σύνολο όπου περιέχει τον αριθμό της γραμμής όπου υπάρχει αυτό το κλειδί. Μέσα στην λίστα *remain* βάζω όλα τα συχνά στοιχειοσύνολα με μέγεθος 1. Φτιάχνω ένα ακόμα λεξικό *d_new* όπου εκεί μέσα κρατάω μόνο τα συχνά μονά στοιχειοσύνολα και σαν τιμή του κλειδιού έχω μετατρέψει σε ένα σύνολο τους αριθμούς των γραμμών στα οποία υπάρχουν αυτά τα κλειδιά.

Έχω μία συνάρτηση την *findTwo()* με την οποία υπολογίζω τα ζεύγη μεγέθους 2 που είναι συχνά και ταυτόχρονα βρίσκω ποια από τα μονά συχνά στοιχειοσύνολα είναι maximal.

def findTwo(): Έχει μία λίστα την *isMaximal* όπου είναι *Boolean*, στην αρχή την έχω αρχικοποιήσει με *True* και δείχνει ότι στην αρχή έστω ότι όλα τα μονά συχνά στοιχειοσύνολα είναι maximal. Φτιάχνω όλα τα ζεύγη μεγέθους 2. Παίρνω την ένωση των γραμμών στις οποίες βρίσκονται και αν το μέγεθος της ένωσης είναι μεγαλύτερο από το κατώφλι τότε σημαίνει ότι το ζεύγος αυτό είναι συχνό και ότι τα επιμέρους στοιχεία του σίγουρα δεν είναι maximal οπότε και αλλάζω στην λίστα *isMaximal* το *True* σε *False*. Ξέρω ποια από τα μονά συχνά στοιχειοσύνολα είναι maximal δεν χρειάζεται να ελέγξω όλα τα υπερσύνολα του για μέγεθος 3 και πάνω γιατί αν ένα υπερσύνολο του για μέγεθος 3 δεν γίνεται να έχει υποσύνολα τα οποία δεν είναι συχνά και οπότε δεν θα το έλεγχα καν αυτό.

Έχω φτιάξει την συνάρτηση *findFrequent(size)* η οποία μου υπολογίζει τα size ζεύγη και βρίσκει ποια είναι maximal. Και η οποία λειτουργεί με τον ίδιο τρόπο όπως και η findTwo() αλλά η διαφορά είναι

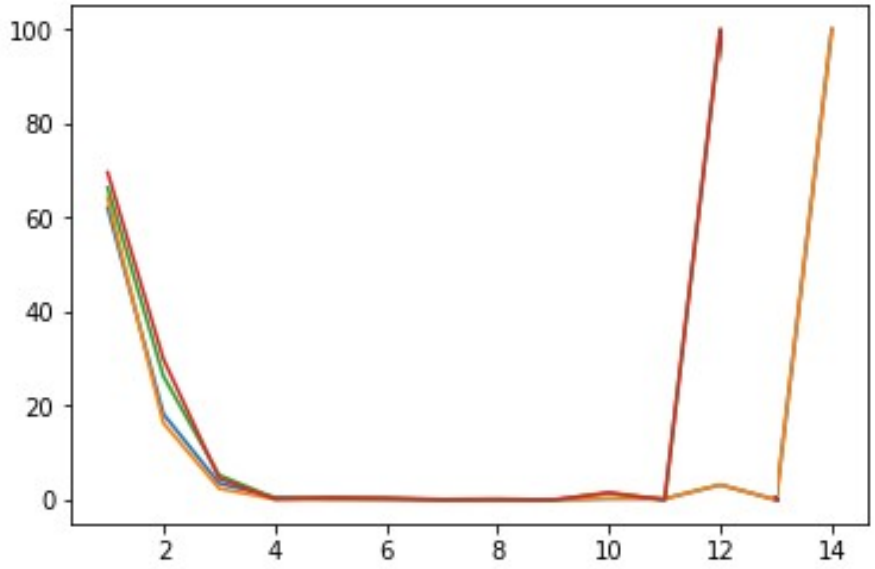
ότι για τα βρει τα ζευγάρια για παράδειγμα για μέγεθος 4 χρησιμοποιεί το σύνολο των συχνών στοιχειοσυνόλων μεγέθους 1 και το σύνολο των ζευγών μεγέθους 3.

Στατιστικά

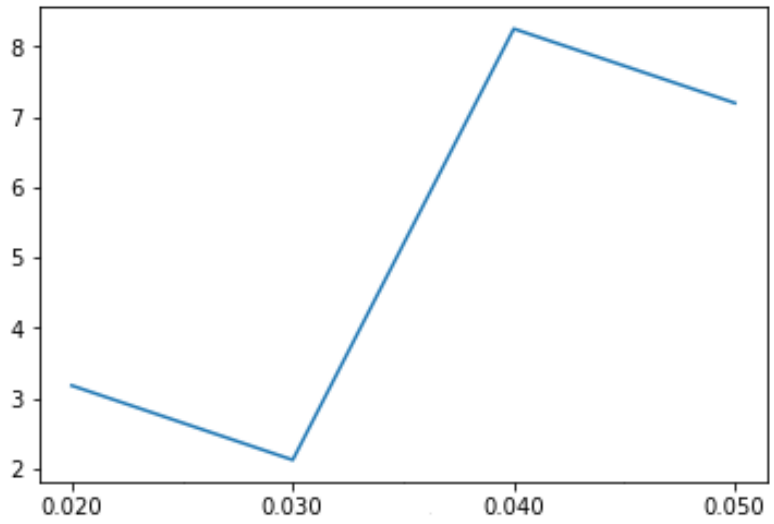
Γράφημα που δείχνει για όλα τα κατώφλια [0.02, 0.05] πόσο της % είναι maximal για όλα τα μεγέθη. Αντιστοιχούν σε:

- πορτοκαλί → 0.02%
- μπλε → 0.03%
- πράσινο → 0.04%
- κόκκινο → 0.05%

Συμπέρασμα, ότι για όλες τις τιμές για το κατώφλι κοντά στο 65% είναι τα maximal συχνά στοιχειοσύνολα για μέγεθος 1 και όσο μεγαλώνει το μέγεθος για τα ζεύγη τόσο μειώνονται τα συχνά maximal στοιχειοσύνολα. Δηλαδή ακόμα και που το κατώφλι αυξάνετε από 0.02% σε 0.05% και είναι λογικό να έχω λιγότερο συχνά στοιχειοσύνολα το πλήθος των maximal συχνών δεν αλλάζει.



Γράφημα που δείχνει για όλα τα κατώφλια [0.02, 0.05] πόσο της % είναι maximal συνολικά για όλα τα μεγέθη. Συμπέρασμα, ότι όταν το κατώφλι είναι 0.02% τα maximal συχνά στοιχειοσύνολα ανεξάρτητα το μέγεθος τους είναι κοντά στο 3% ενώ όταν αυξάνεται το κατώφλι σε 0.04% ενώ είναι λογικό να έχω βρει λιγότερα συχνά στοιχειοσύνολα τα maximal έχουν ανέβει και είναι κοντά στο 8%.



Συσχετίσεις στα συχνά στοιχειοσύνολα για κατώφλι [0.02, 0.03] :

- Ότι όταν υπάρχει κάποιο handle @realdonaldtrump θα συνοδεύεται συνήθως από τα hashtags #makeamericafreeagain #maga #2a #god #jesus #christian όταν τα ζεύγη είναι από 5 έως 15 στοιχεία.
- Τα hashtags #trump #israel #freedom #neverhillary πάνε μαζί δηλαδή όταν αναφέρετε κάποιος για το #israel θα έχει και κάποιο #freedom.
- Τα hashtags σχετικά με αγορές πάλι ταιριάζουν και χρησιμοποιούνται μαζί όσο αυξάνετε το μέγεθος για τα ζεύγη, #deals #ebay #online #toys #amazon #shopping #holidays
- Τα hashtags σχετικά με μουσική συνοδεύονται μαζί πάλι, #gmail #spotify #music #ceo #album #lgbt #yougma #download

Συσχετίσεις στα συχνά στοιχειοσύνολα για κατώφλι [0.04, 0.05] :

- Τα πιο συχνά ζεύγη περιέχουν περιεχόμενο για αγορές και λιγότερα hashtags ή handles για πολιτική, #amazon #bozanza #deals #gifts #online #toys και ότι συνδυάζονται πάλι μαζί