ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΦΟΡΑ ΥΛΟΠΟΙΗΤΙΚΟΥ PROJECT
ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΒΑΚΑΛΟΠΟΥΛΟΣ ΔΗΜΗΤΡΙΟΣ
ΑΜ: 1059564
up1059564@upnet.gr

ΚΑΡΑΜΗΤΣΟΠΟΥΛΟΣ ΔΗΜΗΤΡΙΟΣ
ΑΜ: 1056277
up1056277@upnet.gr

ΠΑΤΡΑ, ΙΑΝΟΥΑΡΙΟΣ 2022

Γλώσσα - Περιβάλλον Υλοποίησης

Η υλοποιητική εργασία, υλοποιήθηκε χρησιμοποιώντας γλώσσα Python (3.8.5). Το εργαλείο το οποίο χρησιμοποιήθηκε για την συγγραφή του κώδικα είναι το : Anaconda Spyder.

Σημαντικές Βιβλιοθήκες που χρησιμοποιήθηκαν

Για την διευκόλυνση της υλοποίησης, εγκαταστάθηκαν και χρησιμοποιήθηκαν οι παρακάτω βιβλιοθήκες:

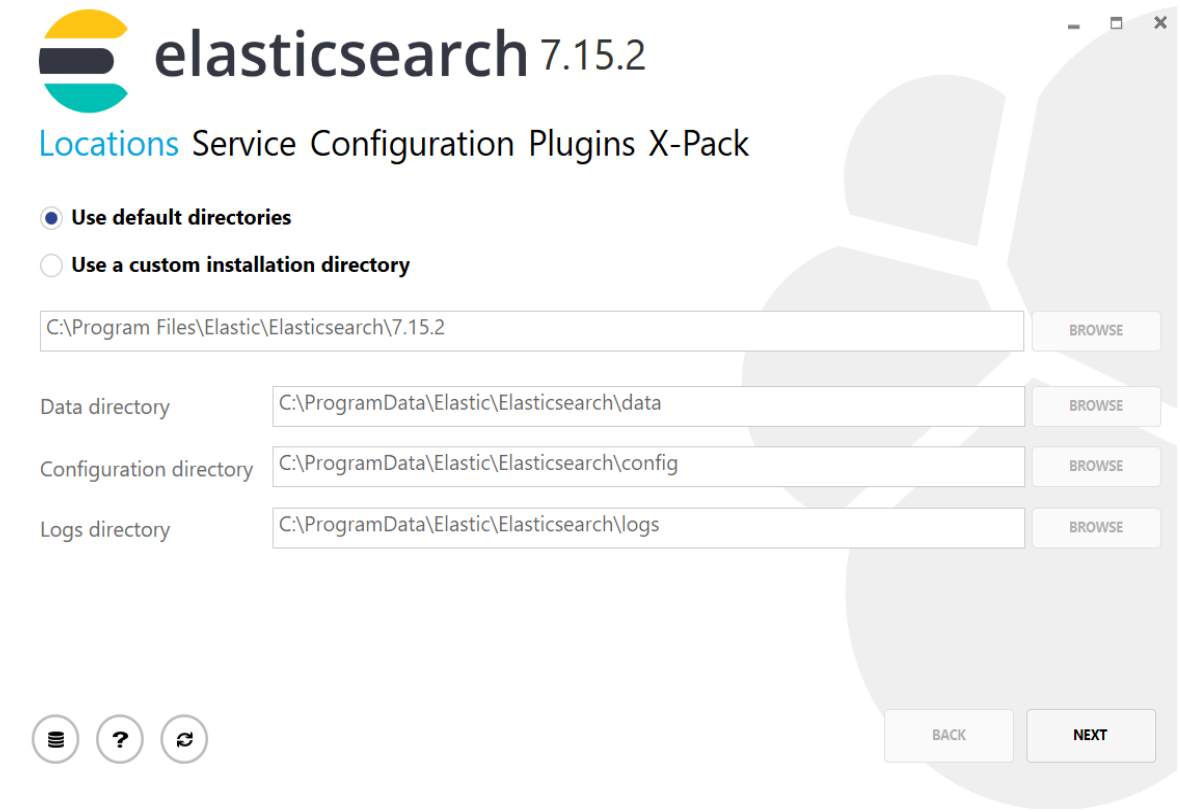
- elasticsearch
- csv
- pandas
- numpy
- gensim
- keras
- sklearn
- math

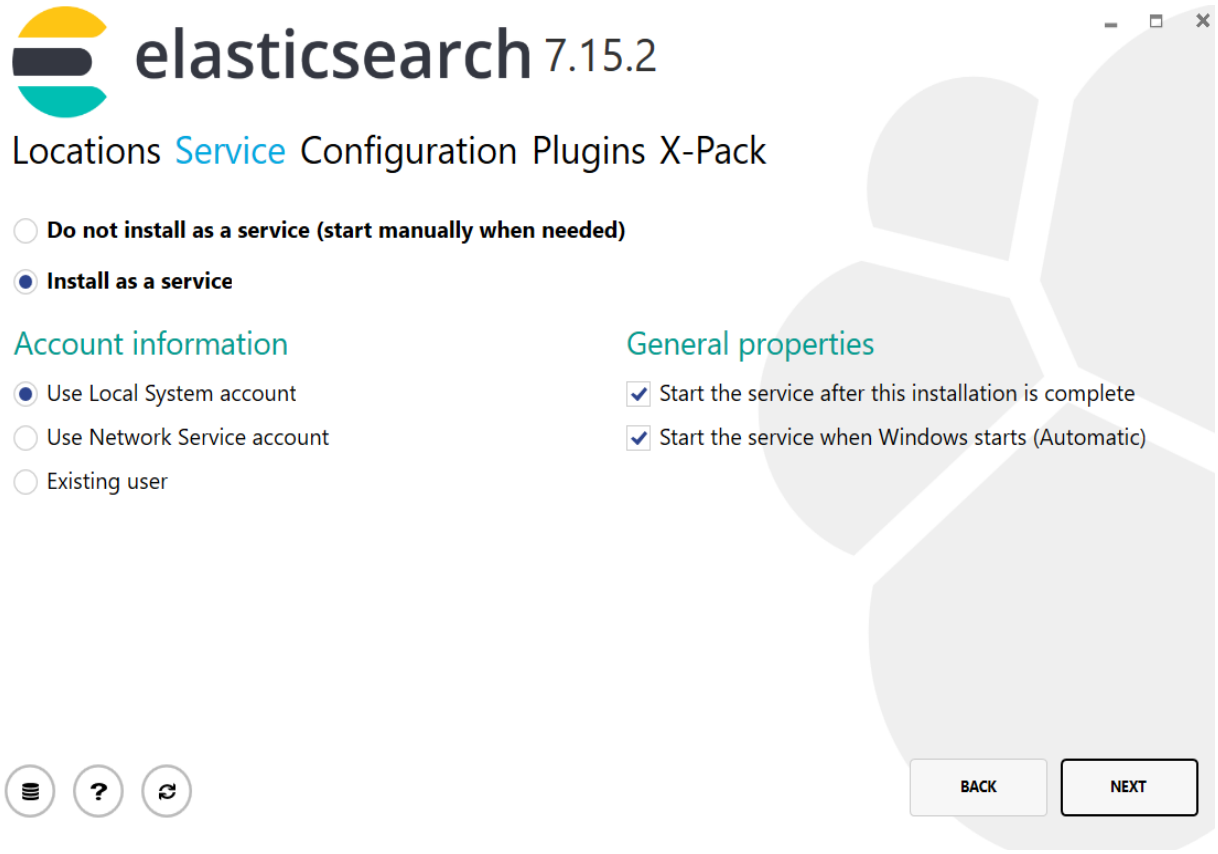
Όλες οι παραπάνω βιβλιοθήκες εγκαταστάθηκαν με την παρακάτω εντολή στο command prompt:

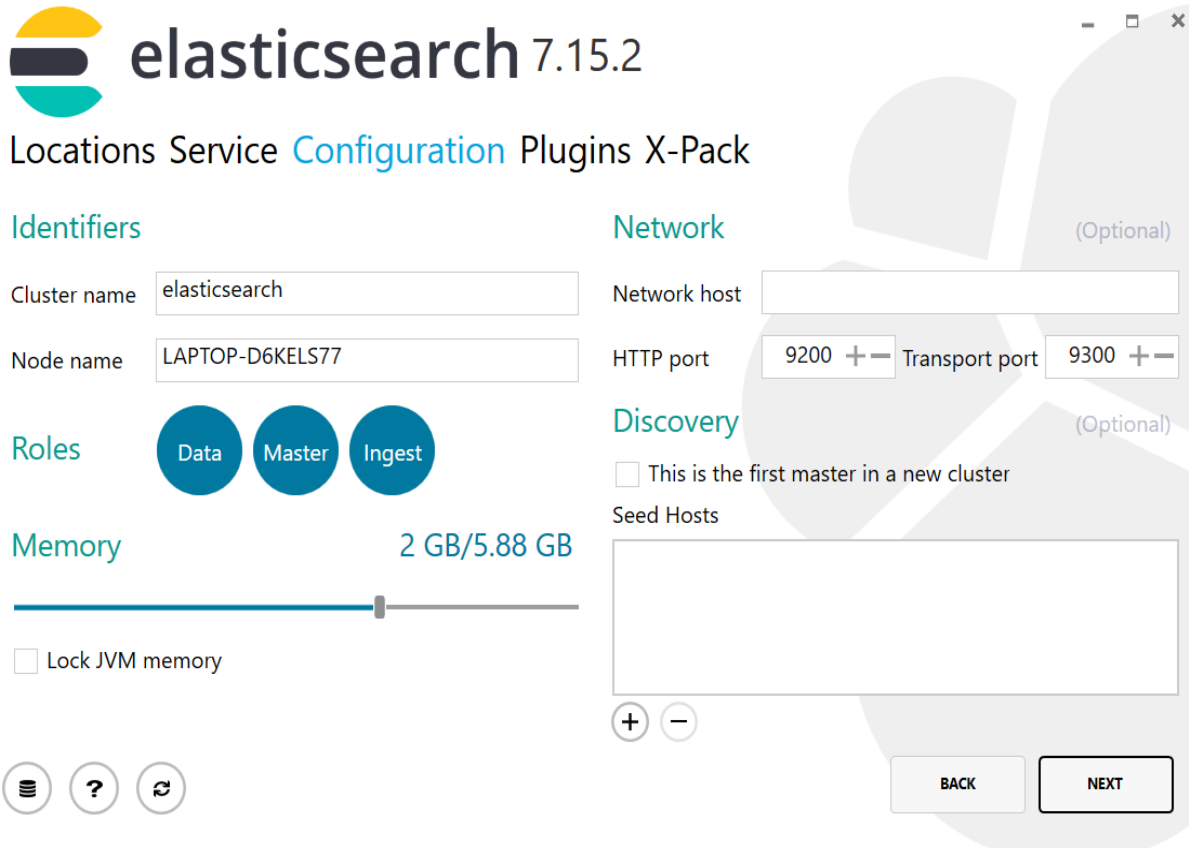
pip install <όνομα_βιβλιοθήκης)

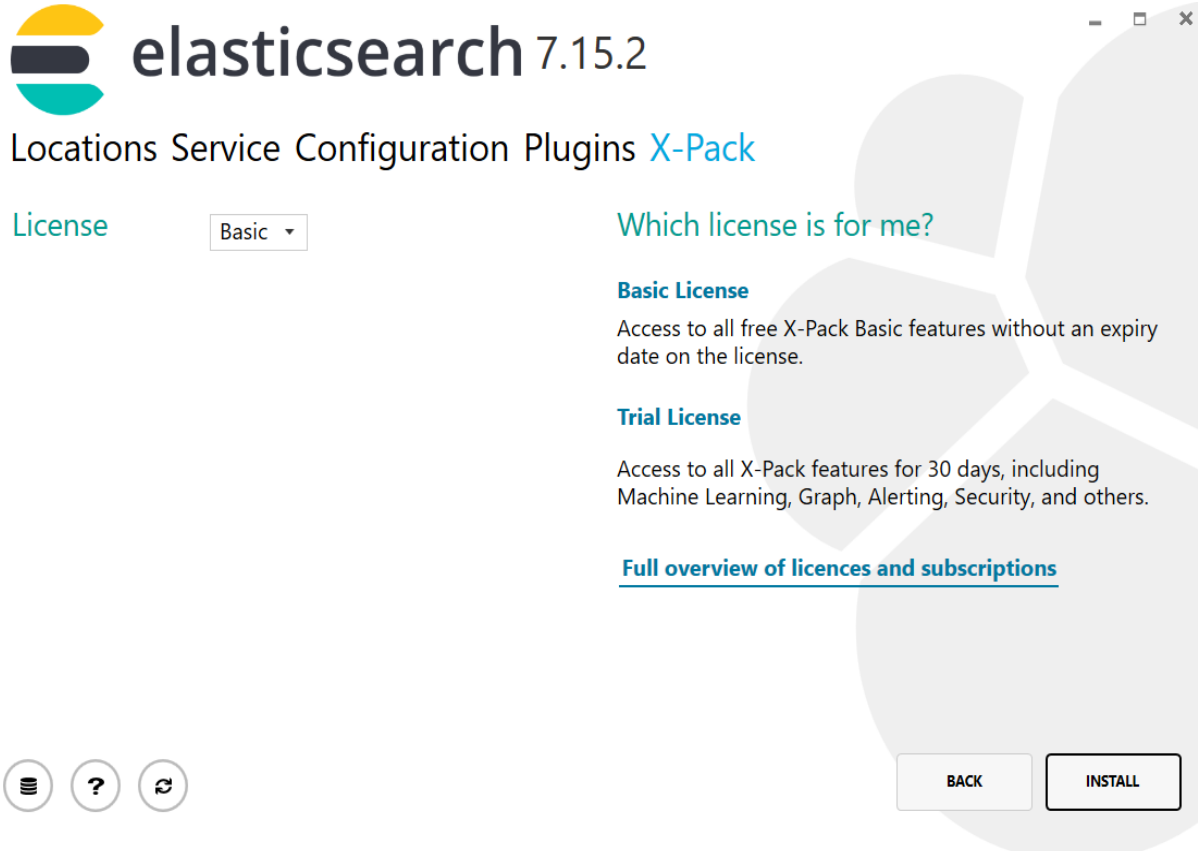
Βήματα Εγκατάστασης Elasticsearch

Στην συνέχεια παραθέτουμε τα βήματα, τα οποία ακολουθήθηκαν για την εγκατάσταση της Elasticsearch (αφού κατεβάσαμε πρώτα το αντίστοιχο .msi αρχείο):


elasticsearch 7.15.2
Locations Service Configuration Plugins X-Pack
Use default directories
Use a custom installation directory
C:\Program Files\Elastic\Elasticsearch\7.15.2
BROWSE
Data directory
C:\ProgramData\Elastic\Elasticsearch\data
BROWSE
Configuration directory
C:\ProgramData\Elastic\Elasticsearch\config
BROWSE
Logs directory
C:\ProgramData\Elastic\Elasticsearch\logs
BROWSE
BACK
NEXT


elasticsearch 7.15.2
Locations Service Configuration Plugins X-Pack
Do not install as a service (start manually when needed)
Install as a service
Account information
Use Local System account
Use Network Service account
Existing user
General properties
Start the service after this installation is complete
Start the service when Windows starts (Automatic)
BACK
NEXT


elasticsearch 7.15.2
Locations Service Configuration Plugins X-Pack
Identifiers
Cluster name
elasticsearch
Node name
LAPTOP-D6KELS77
Roles
Data
Master
Ingest
Memory
2 GB/5.88 GB
Lock JVM memory
Network
(Optional)
Network host
HTTP port
9200
Transport port
9300
Discovery
(Optional)
This is the first master in a new cluster
Seed Hosts
BACK
NEXT


elasticsearch 7.15.2
Locations Service Configuration Plugins X-Pack
License
Basic
Which license is for me?
Basic License
Access to all free X-Pack Basic features without an expiry date on the license.
Trial License
Access to all X-Pack features for 30 days, including Machine Learning, Graph, Alerting, Security, and others.
Full overview of licences and subscriptions
BACK
INSTALL


```
{
  "name" : "LAPTOP-D6KELS77",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "P0y4gjctScCJxfEMOvDRcw",
  "version" : {
    "number" : "7.12.0",
    "build_flavor" : "default",
    "build_type" : "zip",
    "build_hash" : "78722783c38caa25a70982b5b042074cde5d3b3a",
    "build_date" : "2021-03-18T06:17:15.410153305Z",
    "build_snapshot" : false,
    "lucene_version" : "8.8.0",
    "minimum_wire_compatibility_version" : "6.8.0",
    "minimum_index_compatibility_version" : "6.0.0-beta1"
  },
  "tagline" : "You Know, for Search"
}
```

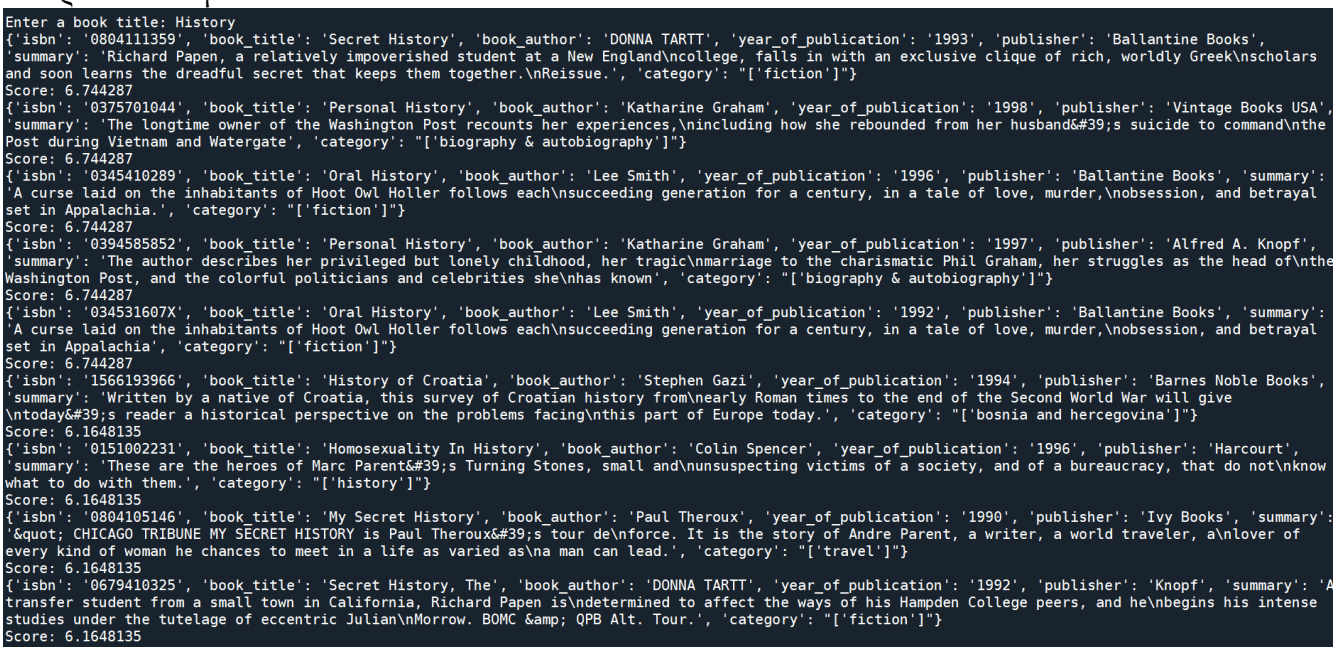

Όπως γίνεται αντιληπτό από το παραπάνω screenshot, η Elasticsearch εγκαταστάθηκε με επιτυχία.

Ερώτημα 1

Για την απάντηση στο ερώτημα αυτό, υλοποιήθηκαν τα εξής αρχεία:

- **import_csv.py** : για την εισαγωγή των εγγραφών στην Elasticsearch
- **search_by_title.py:** στο οποίο δίνεται η δυνατότητα στον χρήστη να εισάγει τίτλο βιβλίου, και η Elasticsearch επιστρέφει τα αντίστοιχα αποτελέσματα, έχοντας υπολογίσει την αντίστοιχη τιμή ομοιότητας για κάθε αποτέλεσμα

Παραθέτουμε ενδεικτικό screenshot:



Ερώτημα 2

Για το συγκεκριμένο ερώτημα υλοποιήθηκε το αρχείο:

- **question2.py**

Βασικά σημεία στο αρχείο αυτό είναι:

- Η εύρεση των βαθμολογιών του χρήστη για τα βιβλία τα οποία έχουν επιστραφεί από την Elasticsearch
- Ο υπολογισμός των μέσων όρων βαθμολογίας που έχουν δώσει οι χρήστες στα βιβλία που έχουν επιστραφεί από την Elasticsearch
- Ο υπολογισμός των νέων score των βιβλίων

Η μετρική η οποία προτάθηκε για τον υπολογισμό των νέων scores είναι η εξής:

Νέο score = a * βαθμολογία_χρήστη + b * βαθμολογία_elastic + c*μέσος_βαθμός_χρηστών

Σημείωση: Οι τιμές των a,b,c δίνονται σαν είσοδος από τον χρήστη από το πληκτρολόγιο. Έτσι λοιπόν, ανάλογα με τις τιμές του κάθε φορά δίνεται και διαφορετική βαρύτητα (είτε στην βαθμολογία του χρήστη, είτε στην μέση βαθμολογία των άλλων χρηστών, είτε στην βαθμολογία της Elasticsearch). Με τον τρόπο αυτό μπορούν να εξαχθούν χρήσιμα συμπεράσματα.

Παράδειγμα εκτέλεσης:

```
Enter a book title: History

Enter a user id:8

Give weight of user ratings :0.7

Give weight of elastic_search ratings :0.2

Give weight of average ratings :0.1
{'isbn': '0804111359', 'book_title': 'Secret History', 'book_author': 'DONNA TARTT', 'year_of_publication': '1993', 'publisher': 'Ballantine Books',
'summary': 'Richard Papen, a relatively impoverished student at a New England\ncollege, falls in with an exclusive clique of rich, worldly Greek\nscholars
and soon learns the dreadful secret that keeps them together.\nReissue.', 'category': "['fiction']"}
Updated Score: 1.6482585976047903
{'isbn': '0375701044', 'book_title': 'Personal History', 'book_author': 'Katharine Graham', 'year_of_publication': '1998', 'publisher': 'Vintage Books USA',
'summary': 'The longtime owner of the Washington Post recounts her experiences,\nincluding how she rebounded from her husband&#39;s suicide to command\nthe
Post during Vietnam and Watergate', 'category': "['biography & autobiography']"}
Updated Score: 1.5693702205128206
{'isbn': '0345410289', 'book_title': 'Oral History', 'book_author': 'Lee Smith', 'year_of_publication': '1996', 'publisher': 'Ballantine Books', 'summary':
'A curse laid on the inhabitants of Hoot Owl Holler follows each\nsucceeding generation for a century, in a tale of love, murder,\nobsession, and betrayal
set in Appalachia.', 'category': "['fiction']"}
Updated Score: 1.7988574
{'isbn': '0394585852', 'book_title': 'Personal History', 'book_author': 'Katharine Graham', 'year_of_publication': '1997', 'publisher': 'Alfred A. Knopf',
'summary': 'The author describes her privileged but lonely childhood, her tragic\nmarriage to the charismatic Phil Graham, her struggles as the head of\nthe
Washington Post, and the colorful politicians and celebrities she\nhas known', 'category': "['biography & autobiography']"}
Updated Score: 1.7695470551724137
{'isbn': '034531607X', 'book_title': 'Oral History', 'book_author': 'Lee Smith', 'year_of_publication': '1992', 'publisher': 'Ballantine Books', 'summary':
'A curse laid on the inhabitants of Hoot Owl Holler follows each\nsucceeding generation for a century, in a tale of love, murder,\nobsession, and betrayal
set in Appalachia', 'category': "['fiction']"}
Updated Score: 1.525780476923077
{'isbn': '1566193966', 'book_title': 'History of Croatia', 'book_author': 'Stephen Gazi', 'year_of_publication': '1994', 'publisher': 'Barnes Noble Books',
'summary': 'Written by a native of Croatia, this survey of Croatian history from\nearly Roman times to the end of the Second World War will give
\ntoday&#39;s reader a historical perspective on the problems facing\nthis part of Europe today.', 'category': "['bosnia and hercegovina']"}
Updated Score: 1.6329627000000002
{'isbn': '0151002231', 'book_title': 'Homosexuality In History', 'book_author': 'Colin Spencer', 'year_of_publication': '1996', 'publisher': 'Harcourt',
'summary': 'These are the heroes of Marc Parent&#39;s Turning Stones, small and\nunsuspecting victims of a society, and of a bureaucracy, that do not\nknow
what to do with them.', 'category': "['history']"}
Updated Score: 1.5329627000000001
```

Ερώτημα 3

Για το ερώτημα αυτό υλοποιήθηκε το αρχείο:

- **question3.py**

Βασικά σημεία στο ερώτημα αυτό είναι:

- η διανυσματοποίηση των περιλήψεων των βιβλίων
- η χρήση νευρωνικού δικτύου και η εκπαίδευσή του έτσι ώστε να μπορεί να προβλέπει τις βαθμολογίες οι οποίες λείπουν

Σημείωση: το νευρωνικό δίκτυο έχει εκπαιδευτεί έτσι ώστε να μπορεί να προβλέπει και να συμπληρώνει τις βαθμολογίες του χρήστη, ο οποίος μας ενδιαφέρει. Σε μελλοντική επέκταση θα μπορούσαμε να προβλέπουμε και τις βαθμολογίες όλων των χρηστών για όλα τα βιβλία, των οποίων οι βαθμολογίες λείπουν.

Παρατήρηση: Διαπιστώνουμε ότι μετά από την ολοκλήρωση της εκτέλεσης (η οποία απαιτεί αρκετό χρόνο εξαιτίας των διαθέσιμων υπολογιστικών μας πόρων, αλλά του όγκου του συνόλου δεδομένων), συμπληρώνονται οι βαθμολογίες του χρήστη οι οποίες λείπουν, που με την σειρά τους επηρεάζουν την τελική κατάταξη των αποτελεσμάτων. Αυτό οφείλεται στο γεγονός ότι πλέον οι βαθμολογίες έχουν αποκτήσει τιμές μη μηδενικές, επηρεάζοντας την τελική βαθμολογία.

Ερώτημα 4

Για την απάντηση στο ερώτημα αυτό υλοποιήθηκε το αρχείο:

- **question4.py**

Βασικά σημεία της υλοποίησης στο αρχείο αυτό:

- Clustering των βιβλίων με βάση τις περιλήψεις τους (εφαρμόστηκε διανυσματοποίηση στα κείμενα των περιλήψεων)
- Clustering των χρηστών με βάση τα δημογραφικά τους χαρακτηριστικά(location και ηλικία). Στην περίπτωση του location εφαρμόστηκε επίσης διαδικασία διανυσματοποίησης.

Και στις 2 περιπτώσεις χρησιμοποιήθηκε ο αλγόριθμος K-Means και πιο συγκεκριμένα :

- 10 clusters για τα βιβλία
- 5 clusters για τους χρήστες

Αφού σχηματίστηκαν τα clusters, για να μπορέσουμε να υπολογίσουμε την συσχέτιση μεταξύ των clusters των χρηστών και των clusters των βιβλίων, αποφασίσαμε να υπολογίσουμε την μέση βαθμολογία που έχουν δώσει οι χρήστες του κάθε cluster στα βιβλία του κάθε cluster.

Με τον τρόπο αυτό υπολογίστηκαν 10 χ 5 = 50 βαθμολογίες, μέσω των οποίων μπορούν να εξαχθούν χρήσιμα συμπεράσματα για τον τρόπο με τον οποίο βαθμολογούν οι χρήστες.
Σαν μελλοντική επέκταση, θα μπορούσαν να χρησιμοποιηθούν και άλλοι αλγόριθμοι ομαδοποίησης, έτσι ώστε να μπορούμε να κάνουμε σύγκριση των αποτελεσμάτων.