Ανάκτηση Πληροφορίας

Εργαστηριακή Άσκηση 2021

Όνομα: Μηναδάκης Εμμανουήλ **ΑΜ:** 1041815 **email:** minadakis@ceid.upatras.gr

Η άσκηση υλοποιήθηκε στη γλώσσα προγραμματισμού python. Η εγκατάσταση της python στα Windows 10 γίνεται μέσω του Microsoft Store. Η έκδοση που χρησιμοποιήθηκε είναι η 3.10. Για το πρώτο και δεύτερο ερώτημα χρησιμοποιήθηκαν οι βιβλιοθήκες elasticsearch, csv και warnings. Για το τρίτο ερώτημα χρησιμοποιήθηκαν επίσης οι βιβλιοθήκες genism.models και nltk. Το τέταρτο ερώτημα δεν έχει υλοποιηθεί.

Μια σύντομη περιγραφή των βιβλιοθηκών που αναφέρονται παραπάνω είναι:

Elasticsearch: Παρέχει τη δυνατότητα σύνδεσης με την elasticsearch. Εγκαθίσταται με την εντολή
```
python -m pip install elasticsearch
```

csv: Βοηθάει στο import των dataset στην elasticsearch
```
python -m pip install csv
```

warnings: Μορφοποιεί την έξοδο του προγράμματος ώστε να μην φαίνονται μερικές άχρηστες πληροφορίες. Δεν χρειάζεται εγκατάσταση.

gensim.models: Κάνουμε import το Word2Vec από αυτήν τη βιβλιοθήκη για να χρησιμοποιήσουμε την τεχνική των word embeddings.
```
python -m pip install gensim
python -m pip install word2vec
```

nltk: Βοηθάει στο tokenize των περιλήψεων των βιβλίων από όπου και κατεβάζουμε τα αρχεία ‘stopwords’.
```
python -m pip install nltk
```

Σύντομη περιγραφή της διαδικασίας:

Αρχικά, εγκατέστησα την Elasticsearch και το γραφικό περιβάλλον Kibana. Δημιούργησα ένα .bat file για να ενεργοποιώ και τις δύο υπηρεσίες ταυτόχρονα με τον παρακάτω κώδικα:

```
START cmd /C E:\DOWNLOADS\elasticsearch-7.16.1-windows-x86_64\elasticsearch-
7.16.1\bin\elasticsearch.bat
START cmd /C E:\DOWNLOADS\kibana-7.16.1-windows-x86_64\kibana-7.16.1-windows-
x86_64\bin\kibana.bat
```

Αρχικά δημιουργήσαμε το αρχείο parser.py το οποίο κάνει import τα BX-Books.csv και BX-Book-Ratings.csv στην elasticsearch. Η Η έξοδος του προγράμματος (αφού περιμένουμε μερικά λεπτά) είναι όπως παρακάτω:

```
Microsoft Windows [Version 10.0.19043.1466]
(c) Microsoft Corporation. All rights reserved.

E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>python parser.py
Connection established

E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>
```

Βεβαιωνόμαστε ότι τα αρχεία φορτώθηκαν εκτελώντας την παρακάτω εντολή στο console του Kibana:

History Settings Help

200 - OK 24 ms

```
1  GET _cat/indices
2  DELETE books
3  DELETE ratings
```

```
1  #! Elasticsearch built-in security features are not enabled. Without authentication, your cluster could be accessible to anyone. See https://www.elastic.co/guide/en/elasticsearch/reference/7.16/security-minimal-setup.html to enable security.
2  green  open .geoip_databases xegLNKYoR7O2lF9wTi7G5Q 1 0     42   4  40.4mb  40.4mb
3  yellow open books Dtzfa5v9TO2mg5DlM5L3sg 1 1 134692   0   97mb    97mb
4  green  open .apm-custom-link 05NOHHxCQCGbZUeAYhg55A 1 0      0   0   226b    226b
5  green  open .kibana_7.16.1_001 Y-YUXP4ORteZpFlxP3vqfg 1 0    442  24  4.8mb   4.8mb
6  yellow open ratings jOW0SdxGT6GyBk8Jcxpb7g 1 1 958500   0   47mb    47mb
7  green  open .apm-agent-configuration 11aIkkImQpGOt9Q_P_6Pcg 1 0      0   0   226b    226b
8  green  open .kibana_task_manager_7.16.1_001 sCALzZLfTsyNhJvHNmowrw 1 0     17 321 217.2kb 217.2kb
9  green  open .async-search 4gVtHhRFTPqqM5EokbJR4Q 1 0      0   0   252b    252b
10 green  open .tasks jnfNsUrsQkiE42yqTV4--Q 1 0     48   0  67.4kb  67.4kb
11
```

Βλέπουμε ότι τα indexes των books και ratings υπάρχουν.

Για το δεύτερο ερώτημα υλοποίησα τη main.py η οποία εκτελεί την αναζήτηση φαίνεται παρακάτω:

```
E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>python main.py
Connection established

User ID: 8
Keyword: clara
[ISBN] 2005018 [Title] Clara Callan [Rating] 7.52
[ISBN] 815410255 [Title] Clara Bow: Runnin' Wild [Rating] 6.25
[ISBN] 679823115 [Title] Sweet Clara and the Freedom Quilt [Rating] 5.78
[ISBN] 899972845 [Title] South Bay Trails: Outdoor Adventures in & Around Santa Clara
 Valley : From the Diablo Range to the Pacific Ocean [Rating] 4.75
[ISBN] 312135084 [Title] Henry and Clara [Rating] 4.71
[ISBN] 451178734 [Title] Dancing With Clara (Signet Regency Romance) [Rating] 4.45
[ISBN] 64441342 [Title] Clara and the Bookwagon (I Can Read Book 3) [Rating] 4.27
[ISBN] 1931561168 [Title] Clara Mondschein's Melancholia [Rating] 3.87
[ISBN] 684844494 [Title] Clara : A Novel [Rating] 3.87
[ISBN] 088240069X [Title] This Old House: The Story of Clara Rust Alaska Pioneer [Rat
ing] 3.81
[ISBN] 20418205 [Title] Clara Barton : Founder Of The American Red Cross (Childhood O
f Famous Americans) [Rating] 3.74
[ISBN] 819310573 [Title] Clara Joins the Circus [Rating] 3.58
[ISBN] 1555466419 [Title] Clara Barton (Women of Achievement) [Rating] 3.33
[ISBN] 1932270116 [Title] Driven to Kill: The Clara Harris Story [Rating] 2.93
[ISBN] 743257502 [Title] The Glory Cloak : A Novel of Louisa May Alcott and Clara Bar
ton [Rating] 2.24

E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>
```

Έχουμε επιλέξει τον χρήση 8 με το keyword “clara”. Γνωρίζουμε ότι ο συγκεκριμένος χρήστης έχει βαθμολογήσει το βιβλίο Clara Callan με 5. Παρακάτω φαίνεται η αναζήτηση του ίδιου keyword από έναν άλλο χρήστη ο οποίος δεν έχει βαθμολογήσει το βιβλίο. Παρατηρήστε τη διαφορά:

```
E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>python main.py
Connection established

User ID: 2
Keyword: clara
[ISBN] 815410255 [Title] Clara Bow: Runnin' Wild [Rating] 6.25
[ISBN] 2005018 [Title] Clara Callan [Rating] 5.86
[ISBN] 679823115 [Title] Sweet Clara and the Freedom Quilt [Rating] 5.78
[ISBN] 899972845 [Title] South Bay Trails: Outdoor Adventures in & Around Santa Clara
 Valley : From the Diablo Range to the Pacific Ocean [Rating] 4.75
[ISBN] 312135084 [Title] Henry and Clara [Rating] 4.71
[ISBN] 451178734 [Title] Dancing With Clara (Signet Regency Romance) [Rating] 4.45
[ISBN] 64441342 [Title] Clara and the Bookwagon (I Can Read Book 3) [Rating] 4.27
[ISBN] 1931561168 [Title] Clara Mondschein's Melancholia [Rating] 3.87
[ISBN] 684844494 [Title] Clara : A Novel [Rating] 3.87
[ISBN] 088240069X [Title] This Old House: The Story of Clara Rust Alaska Pioneer [Rat
ing] 3.81
[ISBN] 20418205 [Title] Clara Barton : Founder Of The American Red Cross (Childhood O
f Famous Americans) [Rating] 3.74
[ISBN] 819310573 [Title] Clara Joins the Circus [Rating] 3.58
[ISBN] 1555466419 [Title] Clara Barton (Women of Achievement) [Rating] 3.33
[ISBN] 1932270116 [Title] Driven to Kill: The Clara Harris Story [Rating] 2.93
[ISBN] 743257502 [Title] The Glory Cloak : A Novel of Louisa May Alcott and Clara Bar
ton [Rating] 2.24

E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>
```

Βλέπουμε ότι το βιβλίο Clara Callan δεν βρίσκεται τόσο ψηλά στα αποτελέσματα διότι έχει χαμηλότερο rating. Αυτό σημαίνει ότι η αναζήτηση λειτουργεί με βάση το τι βιβλία έχει βαθμολογήσει ο χρήστης. Ένα βιβλίο το οποίο έχει βαθμολογηθεί με καλό βαθμό να φαίνεται πιο ψηλά.

Για το τρίτο ερώτημα υλοποίησα το Word2Vec.py το οποίο μαντεύει τις βαθμολογίες που θα έβαζε ο χρήστης σε παρόμοια βιβλία με αυτά που έχει βαθμολογήσει. Αυτό γίνεται αφού παίρνουμε όλα τα summaries από τα βιβλία που έχει βαθμολογήσει ένας χρήστης, τα ενοποιούμε σε ένα, κάνουμε tokenize και δημιουργούμε ένα model το οποίο με βάσει τις λέξεις που υπάρχουν μέσα σε αυτό βρίσκει ομοιότητες με τα summaries που προκύπτουν από τις αναζητήσεις. Υπάρχουν αναλυτικά σχόλια μέσα στο αρχείο του κώδικα. Μία έξοδος του προγράμματος είναι όπως φαίνεται παρακάτω:

```
E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>python word2vec.py
Connection established

User ID: 8
Keyword: october
[ISBN] 034532448X [Title] The October Country [Guessed User Rating] 9.17
[ISBN] 1888996366 [Title] Wedding in October [Guessed User Rating] 9.14
[ISBN] 1565120035 [Title] The Queen of October [Guessed User Rating] 9.05
[ISBN] 671024205 [Title] The October Horse : A Novel of Caesar and Cleopatra [Guessed
 User Rating] 8.87
[ISBN] 684853310 [Title] The October Horse : A Novel of Caesar and Cleopatra [Guessed
 User Rating] 8.87
------
[ISBN] 380973871 [Title] The October Country [Rating] 6.58
[ISBN] 1564580911 [Title] Scorpio: October 24-November 22 [Rating] 6.09
[ISBN] 945575211 [Title] Queen of October [Rating] 5.25
[ISBN] 394499123 [Title] October Light [Rating] 4.79
[ISBN] 026274015X [Title] Looking Awry: An Introduction to Jacques Lacan through Popu
lar Culture (October Books) [Rating] 4.74
[ISBN] 440235502 [Title] October Sky: A Memoir [Rating] 4.71
[ISBN] 034532448X [Title] The October Country [Rating] 4.66
[ISBN] 862419425 [Title] Silence in October [Rating] 4.58
[ISBN] 870212850 [Title] The Hunt for Red October [Rating] 4.45
[ISBN] 140346147 [Title] Mystery on October Road (A Puffin Book) [Rating] 4.37
[ISBN] 684853310 [Title] The October Horse : A Novel of Caesar and Cleopatra [Rating]
 4.30
[ISBN] 038533320X [Title] Rocket Boys (aka October Sky) [Rating] 4.25
[ISBN] 1565120035 [Title] The Queen of October [Rating] 4.25
[ISBN] 425083837 [Title] The Hunt for Red October [Rating] 3.99
[ISBN] 671024205 [Title] The October Horse : A Novel of Caesar and Cleopatra [Rating]
 3.96
[ISBN] 425120279 [Title] The Hunt for Red October [Rating] 3.90
[ISBN] 812516818 [Title] October Wind [Rating] 3.90
[ISBN] 471415340 [Title] October Fury [Rating] 3.90
[ISBN] 451458958 [Title] October Dreams: A Celebration of Halloween [Rating] 3.72
[ISBN] 192713841 [Title] The October Child [Rating] 3.58
[ISBN] 1888996366 [Title] Wedding in October [Rating] 3.58
[ISBN] 156012979 [Title] Silence in October [Rating] 3.58
[ISBN] 553254324 [Title] The October Circle [Rating] 3.58
[ISBN] 590460110 [Title] Mystery on October Road [Rating] 3.32
[ISBN] 1854592920 [Title] Pentecost: The Rsc/Allied Domecq Young Vic Season : First P
erformed at the Other Place, Stratford-Upon-Avon, 12 October 1994 [Rating] 3.18
[ISBN] 075730012X [Title] We Are Not Afraid: Strength and Courage from the Town That
Inspired the #1 Bestseller and Award-Winning Movie \October Sky\"" [Rating] 3.13
[ISBN] 425172902 [Title] The Hunt for Red October (Special 15th Anniversary Edition)
[Rating] 2.41
[ISBN] 345407857 [Title] The October Country: By Ray Bradbury ; Illustrated by Joemug
naini ; All-New Introduction by the Author [Rating] 1.82
[ISBN] 805423249 [Title] October Dawn: A Novel Based on the Cuban Missile Crisis (Wal
ker, James, Mysteries in Time Series.) [Rating] 1.75
[ISBN] 671768069 [Title] The Missiles of October: The Declassified Story of John F. K
ennedy and the Cuban Missile Crisis [Rating] 1.75

E:\DESKTOP\IR_project_2021\IR_PROJECT_MINADAKIS_EMMANOUIL_1041815>
```

Αυτό που βλέπουμε είναι ότι στο μέρος πάνω από τις παύλες, έχει βρει ποια βιβλία από αυτά που υπάρχουν στα αποτελέσματα, παρουσιάζουν ομοιότητα με αυτά που έχει βαθμολογήσει και τους βάζει ένα rating το οποίο προκύπτει από τα summaries που αυτά έχουν.