

Αλέξιος Νταβλούρος
Ορέστης Μαραζιώτης

1059653
1064028

Χρησιμοποιήθηκε η γλώσσα **Python**.

Χρησιμοποιήθηκαν οι βιβλιοθήκες ...

1. scikit-learn

Εγκατάσταση :

Install the 64bit version of Python 3, for instance from <https://www.python.org>.
Then run:

```
$ pip install -U scikit-learn
```

2. gensim

Εγκατάσταση :

To install gensim, simply run:

```
pip install --upgrade gensim
```

Για το μοντέλο word2vec :

import api module.

```
import gensim.downloader as api
```

“Κατέβασμα” text8 corpus and φόρτωση ως Python αντικείμενο ,το οποίο είναι προσβάσιμο ως stream.

```
corpus = api.load('text8')
```

Train word2vec model.

```
from gensim.models.word2vec import Word2Vec  
model = Word2Vec(corpus)
```

3. numpy

PIP

If you use `pip`, you can install NumPy with:

```
pip install numpy
```

3. pandas

Installing from PyPI

pandas can be installed via pip from PyPI.

Note

You must have `pip>=19.3` to install from PyPI.

```
pip install pandas
```

Διαδικασία υλοποίησης :

Ερώτημα 1

1. Άνοιγμα αρχείου BX-Books.csv
2. Bulk insert σε Elasticsearch , αφού πρώτα έχω δημιουργήσει σύνδεση.

Ερώτημα 2

1. Άνοιγμα αρχείου BX-Book-Ratings.csv και BX-Users.csv

2. Bulk insert σε Elasticsearch , αφού πρώτα έχω δημιουργήσει σύνδεση.
3. Αφού δώσω id , εύρεση βιβλίων ,τα οποία έχει βαθμολογήσει , μέσω Elasticsearch.
4. Συνυπολογισμός της μετρικής ομοιότητας της Elasticsearch, της βαθμολογίας που έχει βάλει ο χρήστης στο βιβλίο (αν είναι διαθέσιμη) και το μέσο όρο όλων των βαθμολογιών που έχουν βάλει οι υπόλοιποι χρήστες.
5. Ταξινόμηση κατά φθίνουσα σειρά και επιστροφή.

Ερώτημα 3

1. Άνοιγμα αρχείου BX-Book-Ratings.csv
2. Παίρνω ό,τι έχει βαθμολογήσει ο χρήστης με id ,που έχω δώσει ως όρισμα σε embed(uid).
3. Παίρνω περιλήψεις απο ό,τι έχει βαθμολογήσει ο χρήστης .
4. Μετατρέπω τις περιλήψεις σε διανύσματα με ίσες διαστάσεις. Κάποιες λέξεις δεν υπάρχουν στο λεξικό ,που πέρασα στο word2vec , ή οι περιλήψεις έχουν μικρότερο απο άλλες μήκος , οπότε αρχικά παίρνουν τιμή nan .Μετά θέτω κάθε nan ίσο με 0 , και τέλος το θέτω ίσο με τον μέσο όρο των στοιχείων της περίληψης.
5. Χρησιμοποιώ το μοντέλο πρόβλεψης KNN ή SVM και προβλέπω τις βαθμολογίες ,που θα έβαζε ο χρήστης.

Ερώτημα 4

Ο παραδοτέος κώδικας δεν είναι ο τελικός.

Υλοποίηση k-means για τα βιβλία ,που έχει βαθμολογήσει ο χρήστης ή είναι προς βαθμολόγηση.

1. Άνοιγμα αρχείου BX-Book-Ratings.csv
2. Παίρνω ό,τι έχει βαθμολογήσει ο χρήστης με id ,που έχω δώσει ως όρισμα σε embed(uid).
3. Παίρνω περιλήψεις απο ό,τι έχει βαθμολογήσει ο χρήστης .
4. Μετατρέπω τις περιλήψεις σε διανύσματα με ίσες διαστάσεις. Κάποιες λέξεις δεν υπάρχουν στο λεξικό ,που πέρασα στο word2vec , ή οι περιλήψεις έχουν μικρότερο απο άλλες μήκος , οπότε αρχικά παίρνουν τιμή nan .Μετά θέτω κάθε nan ίσο με 0 , και τέλος το θέτω ίσο με τον μέσο όρο των στοιχείων της περίληψης.
5. Τυχαία επιλέγω K απο τις περιλήψεις ως κεντροειδή.
6. Υπολογισμός νέων κεντροειδών μέχρι να έχω ίδια κεντροειδή μεταξύ 2 διαδοχικών επαναλήψεων ή να ξεπεράσω τον μέγιστο αριθμό επαναλήψεων.