Τμήμα Εφαρμοσμένης Πληροφορικής
Ανάλυση Δεδομένων Μεγάλου Όγκου
Ακαδημαϊκό έτος 2023-2024

Δεύτερη ομαδική εργασία
Ημερομηνία παράδοσης: Τετάρτη, 14 Ιανουαρίου 2024

Σας δίνεται ένα σύνολο δεδομένων το οποίο περιλαμβάνει κείμενα με πρωτότυπο και παραφρασμένο κείμενο. Το σύνολο δεδομένων μπορείτε να το κατεβάσετε ως εξής:

wget -O Webis-CPC-11.zip https://zenodo.org/records/3251771/files/Webis-CPC-11.zip?download=1

Μας ενδιαφέρουν τα txt αρχεία τα οποία στο όνομά τους έχουν

1. original.txt
2. paraphrase.txt

Εντός της διαδρομής Webis-CPC-11. Τα αρχεία αυτά θα ανέβουν στο HDFS.

Προσοχή! Θα πρέπει να διαβάσετε τα αρχεία ώστε κάθε αρχείο να αποτελεί ξεχωριστή εγγραφή!

Θα πρέπει να συγκρίνετε τα έγγραφα *original.txt με τα έγγραφα *paraphrase.txt και να εντοπίσετε τα 10 ζεύγη εγγράφων με τη μεγαλύτερη ομοιότητα, χρησιμοποιώντας τη βιβλιοθήκη ML που αποτελεί μέρος του Apache Spark για MinHashing, βάζοντας στην παράμετρο number of hash Tables τις παρακάτω τιμές: 2, 5, 10 για κατώφλι απόστασης 0.8.

Πραγματοποιήστε 3 εκτελέσεις και καταγράψτε:

a. Τα έγγραφα αυτά, για κάθε περίπτωση σε μορφή csv
b. τον μέσο χρόνο που απαιτήθηκε για τον υπολογισμό.
c. Τα παραπάνω θα τα πραγματοποιήσετε:
    i. Για έναν worker
    ii. Για δύο workers

2. Σχολιάστε τα αποτελέσματα των παραπάνω μετρήσεων.

Τον χρόνο κάθε εκτέλεσης μπορείτε να τον δείτε στο πρόγραμμα περιήγησής σας στη διεύθυνση <ip-του-master>:8080

Οδηγίες

1. **Η εργασία είναι ομαδική αποκλειστικά σε ομάδες των 2 ατόμων.**
2. Θα παραδώσετε:
    2.1. Ένα αρχείο κώδικά και την εντολή που χρησιμοποιείτε για να εκτελεστεί (δείτε και στο παράδειγμα του μαθήματος).
    2.2. Μία αναφορά έως 3 σελίδες σχετικά με τα βήματα που ακολουθήσατε, τα προβλήματα που αντιμετωπίσετε και τους τρόπους επίλυσής τους.
    2.3. Τις εντολές και παραμέτρους που χρησιμοποιήσατε σε ένα .txt αρχείο.

2.3.1. Ένα αρχείο csv ανά τιμή hashTables με την έξοδο που θα παραχθεί.

2.4. Την έξοδο της εκτέλεσης του Spark.

3. Η εργασία θα πρέπει να παραδοθεί **ηλεκτρονικά** μέσω openeclass σε μορφή **ενός αρχείου** zip.

Ενδεικτικές Αναφορές:

1. Κεφάλαιο 3 από το βιβλίο «Εξόρυξη από Μεγάλα Σύνολα Δεδομένων», διαθέσιμο και online στο http://www.mmds.org/ (αγγλικά)
2. «Data-Intensive Text Processing with MapReduce». Διαθέσιμο online στο https://lintool.github.io/MapReduceAlgorithms/
3. PySpark Recipes https://link.springer.com/book/10.1007/978-1-4842-3141-8
4. PySpark SQL Recipes https://link.springer.com/book/10.1007/978-1-4842-4335-0
5. Learn PySpark https://link.springer.com/book/10.1007/978-1-4842-4961-1
6. Applied Data Science Using PySpark https://link.springer.com/book/10.1007/978-1-4842-6500-0
7. Machine Learning with PySpark https://link.springer.com/book/10.1007/978-1-4842-4131-8
8. Beginning Apache Spark 3 https://link.springer.com/book/10.1007/978-1-4842-7383-8
9. Big Data Analytics with Spark https://link.springer.com/book/10.1007/978-1-4842-0964-6
10. https://spark.apache.org/docs/latest/ml-features.html
11. Σημειώσεις και παραδείγματα του μαθήματος

Χρήσιμα:

Παράμετροι ορίζονται ως εξής κατά την εκτέλεση του pyspark ή spark-submit:

--driver-memory 2000M : Μνήμη στον driver
--executor-memory 6000M : Μνήμη στον executor
--num-executors 4 : Πλήθος executors
--executor-cores 4 : Πλήθος πυρήνων στον executor

Στον κώδικα:

```
from  pyspark.sql.functions import input_file_name

df = df.withColumn("filename", input_file_name())


from pyspark.sql.functions import regexp_replace
df = df.withColumn('id', regexp_replace('id', 'a', 'b'))
```