AmazonRank

Giardino Di Lollo Chiara

Minervino Chiara

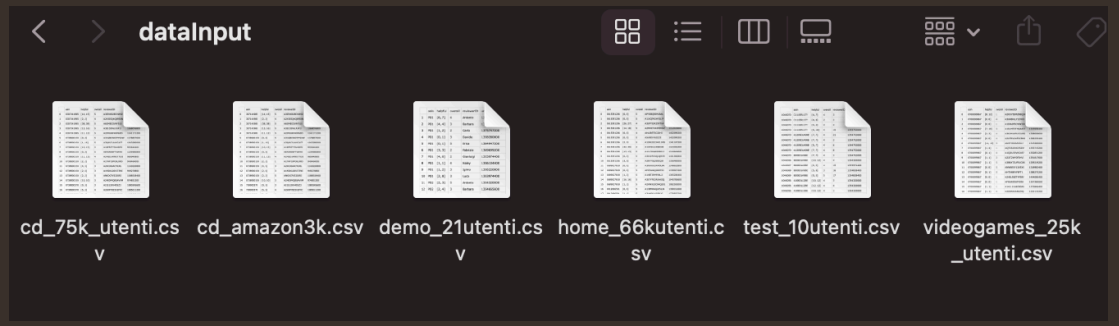
PageRank

- Il PageRank (PR) è un algoritmo utilizzato da Google per attribuire un punteggio alle pagine web , basandosi sulla quantità e sulla qualità dei link in ingresso.
- indica la probabilità che un utente casuale possa continuare a navigare in un sito web, o al di fuori di esso, cliccando sui link presenti nelle pagine
- Applicheremo questo algoritmo per analizzare l'utilità delle recensioni

Amazon

- Amazon è una piattaforma diffusa per gli acquisti di svariati prodotti
- Ogni giorno vengono ordinati milioni di prodotti
- per ognuno di esso vengono fatte svariate recensioni
- Amazon valuta queste recensioni e mette a disposizione tali dati

Dataset



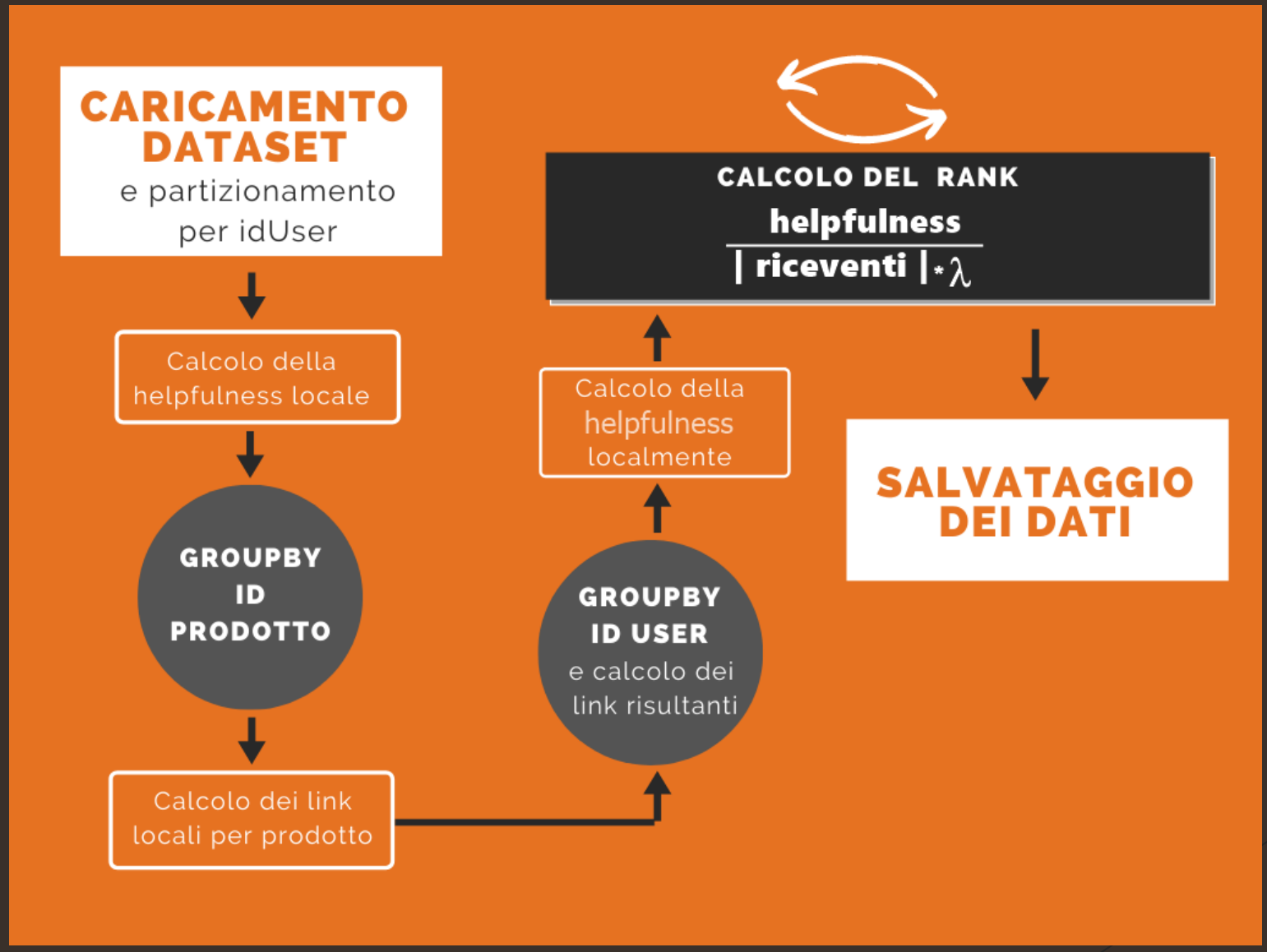
cd_amazon3k - Blocco note di Windows

File Modifica Formato Visualizza ?

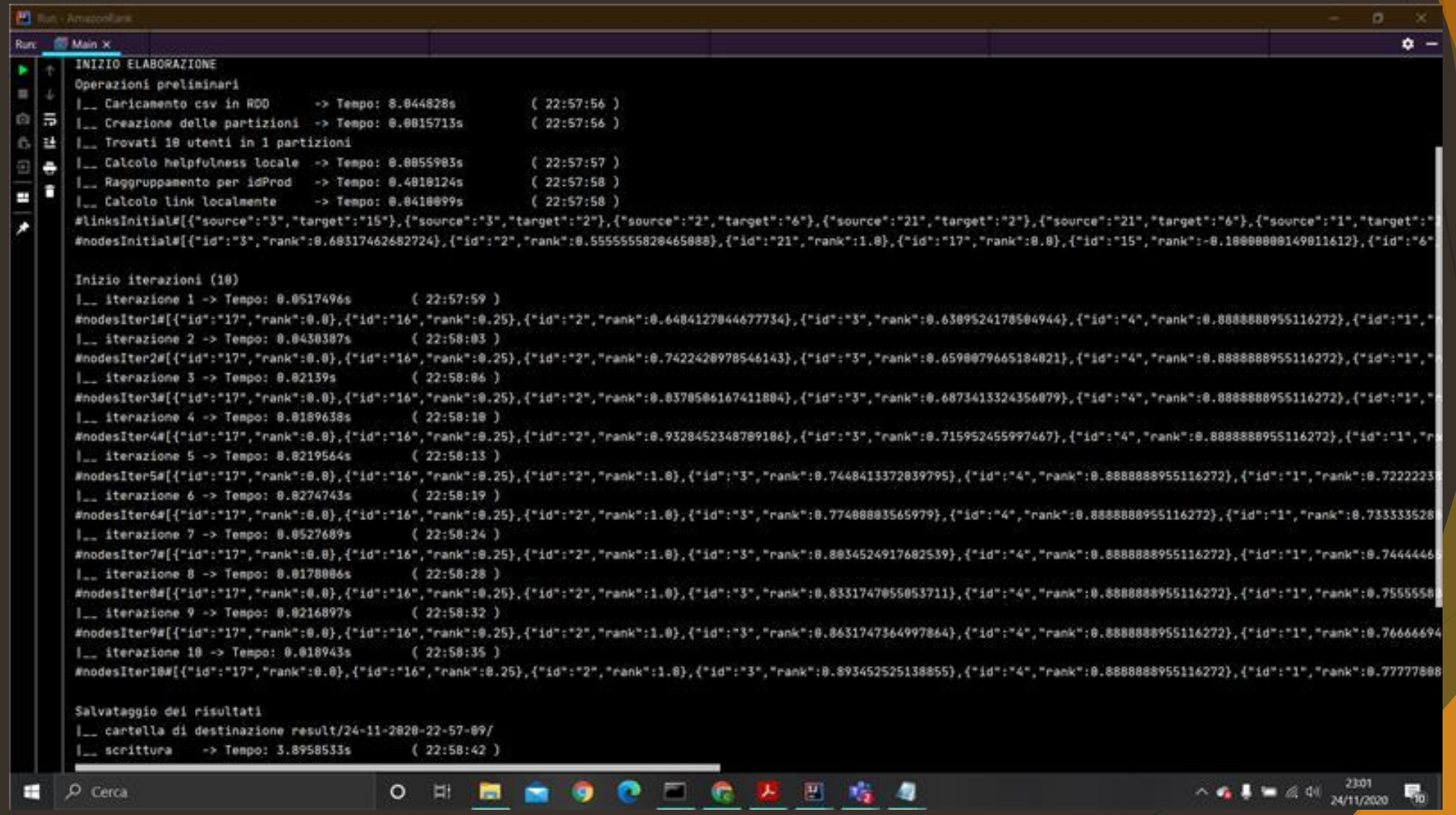
```
,asin,helpful,overall,reviewerID,unixReviewTime
0,307141985,"[14, 15]",5,A3IEV6R2B7VW5Z,1128556800
1,307141985,"[2, 2]",4,A2H3ISQ4QB95XN,1322006400
2,307141985,"[38, 38]",5,A6GMEO3VRY51S,1058140800
3,307141985,"[15, 16]",5,A3E102F6LPUF1J,1068076800
4,307141985,"[11, 12]",5,A2JP0URFHXP6DO,1141171200
```

- reviewerID - ID dell'utente che ha fatto la recensione
- asin - ID del prodotto
- helpful - utilità della recensione
- overall - voto dato al prodotto
- unixReviewTime - timestamp di inserimento della recensione

Procedimento applicato



Anteprima di quello che succede ai dati



Test a confronto

```
 -> Tempo: 6.426536s             ( 15:07:28 )
|__ Partizione per ID utente    -> Tempo: 0.001814172s          ( 15:07:28 )
|__ Calcolo helpfulness locale  -> Tempo: 0.008024963s          ( 15:07:28 )
|__ Raggruppamento per idProd   -> Tempo: 0.017848713s          ( 15:07:28 )
|__ Calcolo link localmente     -> Tempo: 0.031079348s          ( 15:07:29 )

Inizio iterazioni (10)
|__ iterazione 1 -> Tempo: 0.06494906s          ( 15:07:29 )
|__ iterazione 2 -> Tempo: 0.031749655s                 ( 15:07:29 )
|__ iterazione 3 -> Tempo: 0.06103472s          ( 15:07:29 )
|__ iterazione 4 -> Tempo: 0.03052626s          ( 15:07:29 )
|__ iterazione 5 -> Tempo: 0.025714029s                 ( 15:07:29 )
|__ iterazione 6 -> Tempo: 0.025735226s                 ( 15:07:29 )
|__ iterazione 7 -> Tempo: 0.028975567s                 ( 15:07:29 )
|__ iterazione 8 -> Tempo: 0.031882077s                 ( 15:07:29 )
|__ iterazione 9 -> Tempo: 0.024479622s                 ( 15:07:29 )
|__ iterazione 10 -> Tempo: 0.025820361s                ( 15:07:29 )
20/11/24 15:07:29 INFO ClientConfigurationFactory: Set initial getObject socket timeout
20/11/24 15:07:38 INFO MultipartUploadOutputStream: close closed:false s3://aws-logs-860

Salvataggio dei risultati
|__ cartella di destinazione s3n://aws-logs-860100786633-us-east-1/24-11-2020-15-07-11/
|__ scrittura    -> Tempo: 9.049339s            ( 15:07:38 )

FINE ELABORAZIONE

Riepilogo:
-> file: s3://aws-logs-860100786633-us-east-1/cd_amazon3k.csv
-> demo: false
-> ITER: 10
-> LAMBDA: 20
-> TIMEOUT: 3000)
-> Tempo totale: 26.573267s
[hadoop@ip-172-31-10-17 ~]$
```

```
INIZIO ELABORAZIONE
Operazioni preliminari
|__ Caricamento csv in RDD      -> Tempo: 1.8799341s          ( 23:13:39 )
|__ Creazione delle partizioni  -> Tempo: 0.0027865s          ( 23:13:39 )
|__ Calcolo helpfulness locale  -> Tempo: 0.0052696s          ( 23:13:39 )
|__ Raggruppamento per idProd   -> Tempo: 0.0327472s          ( 23:13:39 )
|__ Calcolo link localmente     -> Tempo: 0.0378044s          ( 23:13:39 )

Inizio iterazioni (10)
|__ iterazione 1 -> Tempo: 0.0604462s         ( 23:13:39 )
|__ iterazione 2 -> Tempo: 0.0389651s         ( 23:13:39 )
|__ iterazione 3 -> Tempo: 0.0601292s         ( 23:13:39 )
|__ iterazione 4 -> Tempo: 0.0391309s         ( 23:13:39 )
|__ iterazione 5 -> Tempo: 0.0483606s         ( 23:13:39 )
|__ iterazione 6 -> Tempo: 0.0483607s         ( 23:13:39 )
|__ iterazione 7 -> Tempo: 0.055771s          ( 23:13:39 )
|__ iterazione 8 -> Tempo: 0.0805377s         ( 23:13:39 )
|__ iterazione 9 -> Tempo: 0.040704s          ( 23:13:39 )
|__ iterazione 10 -> Tempo: 0.0299923s        ( 23:13:39 )

Salvataggio dei risultati
|__ cartella di destinazione result/23-11-2020 23-13-32/
|__ scrittura    -> Tempo: 7.9675555s         ( 23:13:47 )

FINE ELABORAZIONE

Riepilogo:
-> file: dataset\cd_amazon3k.csv
-> demo: false
-> ITER: 10
-> LAMBDA: 20
-> TIMEOUT: 3000
-> Tempo totale: 15.612831s

Process finished with exit code 0
```

Test a confronto

```
INIZIO ELABORAZIONE
Operazioni preliminari
|__ Caricamento csv in RDD     20/11/24 15:27:03 INFO GPLNativeCodeLoader: L
20/11/24 15:27:03 INFO LzoCodec: Successfully loaded & initialized native-li
20/11/24 15:27:04 INFO ClientConfigurationFactory: Set initial getObject soc
 -> Tempo: 6.237444s             ( 15:27:09 )
|__ Partizione per ID utente    -> Tempo: 0.002151843s          ( 15:27:09
|__ Calcolo helpfulness locale  -> Tempo: 0.016410515s          ( 15:27:09
|__ Raggruppamento per idProd   -> Tempo: 0.038205653s          ( 15:27:09
|__ Calcolo link localmente     -> Tempo: 0.08499782s           ( 15:27:09

Inizio iterazioni (10)
|__ iterazione 1 -> Tempo: 0.06774285s          ( 15:27:09 )
|__ iterazione 2 -> Tempo: 0.0583399s           ( 15:27:10 )
|__ iterazione 3 -> Tempo: 0.048529986s                ( 15:27:10 )
|__ iterazione 4 -> Tempo: 0.028474294s                ( 15:27:10 )
|__ iterazione 5 -> Tempo: 0.03673575s          ( 15:27:10 )
|__ iterazione 6 -> Tempo: 0.030329674s                ( 15:27:10 )
|__ iterazione 7 -> Tempo: 0.035407756s                ( 15:27:10 )
|__ iterazione 8 -> Tempo: 0.025828134s                ( 15:27:10 )
|__ iterazione 9 -> Tempo: 0.025445528s                ( 15:27:10 )
|__ iterazione 10 -> Tempo: 0.024913417s               ( 15:27:10 )
20/11/24 15:27:10 INFO ClientConfigurationFactory: Set initial getObject soc
20/11/24 15:35:15 INFO MultipartUploadOutputStream: close closed:false s3://

Salvataggio dei risultati
|__ cartella di destinazione s3n://aws-logs-860100786633-us-east-1/24-11-202
|__ scrittura    -> Tempo: 485.0373s            ( 15:35:15 )

FINE ELABORAZIONE

Riepilogo:
-> file: s3://aws-logs-860100786633-us-east-1/home_66kutenti.csv
-> demo: false
-> ITER: 10
-> LAMBDA: 20
-> TIMEOUT: 3000)
-> Tempo totale: 502.58936s
[hadoop@ip-172-31-10-17 ~]$
[hadoop@ip-172-31-10-17 ~]$
```

```
Inizio iterazioni (10)
|__ iterazione 1 -> Tempo: 0.0616341s          ( 15:21:05 )
|__ iterazione 2 -> Tempo: 0.0351551s          ( 15:21:05 )
|__ iterazione 3 -> Tempo: 0.0305329s          ( 15:21:05 )
|__ iterazione 4 -> Tempo: 0.0551064s          ( 15:21:05 )
|__ iterazione 5 -> Tempo: 0.0323586s          ( 15:21:05 )
|__ iterazione 6 -> Tempo: 0.0359986s          ( 15:21:05 )
|__ iterazione 7 -> Tempo: 0.0326666s          ( 15:21:05 )
|__ iterazione 8 -> Tempo: 0.0498047s          ( 15:21:05 )
|__ iterazione 9 -> Tempo: 0.0399586s          ( 15:21:05 )
|__ iterazione 10 -> Tempo: 0.0618601s         ( 15:21:05 )

Salvataggio dei risultati
|__ cartella di destinazione result/24-11-2020-15-20-49/
|__ scrittura    -> Tempo: 2001.3077s          ( 15:54:26 )

FINE ELABORAZIONE

Riepilogo:
-> file: dataset\home_66kutenti.csv
-> demo: false
-> ITER: 10
-> LAMBDA: 20
-> TIMEOUT: 3000
-> Tempo totale: 2017.4421s
```

Test a confronto

```
INIZIO ELABORAZIONE
Operazioni preliminari
|__ Caricamento csv in RDD     20/11/24 21:27:49 INFO GPLNativeCodeLoader: Loaded nati
20/11/24 21:27:49 INFO LzoCodec: Successfully loaded & initialized native-lzo library
20/11/24 21:27:50 INFO ClientConfigurationFactory: Set initial getObject socket timeou
 -> Tempo: 6.512505s            ( 21:27:55 )
|__ Partizione per ID utente    -> Tempo: 0.001753675s          ( 21:27:55 )
|__ Calcolo helpfulness locale  -> Tempo: 0.007744036s          ( 21:27:55 )
|__ Raggruppamento per idProd   -> Tempo: 0.018780205s          ( 21:27:55 )
|__ Calcolo link localmente     -> Tempo: 0.022339694s          ( 21:27:55 )

Inizio iterazioni (10)
|__ iterazione 1 -> Tempo: 0.058150504s                 ( 21:27:55 )
|__ iterazione 2 -> Tempo: 0.032193396s                 ( 21:27:55 )
|__ iterazione 3 -> Tempo: 0.030117359s                 ( 21:27:55 )
|__ iterazione 4 -> Tempo: 0.036709864s                 ( 21:27:55 )
|__ iterazione 5 -> Tempo: 0.030790962s                 ( 21:27:55 )
|__ iterazione 6 -> Tempo: 0.02760982s          ( 21:27:55 )
|__ iterazione 7 -> Tempo: 0.02917338s          ( 21:27:55 )
|__ iterazione 8 -> Tempo: 0.0766478s           ( 21:27:55 )
|__ iterazione 9 -> Tempo: 0.03532468s          ( 21:27:55 )
|__ iterazione 10 -> Tempo: 0.022038111s                ( 21:27:55 )

Salvataggio dei risultati
|__ cartella di destinazione s3n://aws-logs-743342462495-us-east-1/24-11-2020-21-27-36
20/11/24 21:27:55 INFO ClientConfigurationFactory: Set initial getObject socket timeou
20/11/24 21:29:31 INFO MultipartUploadOutputStream: close closed:false s3://aws-logs-7
|__ scrittura    -> Tempo: 96.09201s            ( 21:29:31 )

FINE ELABORAZIONE

Riepilogo:
-> file: s3://aws-logs-743342462495-us-east-1/videogames_25k_utenti.csv
-> demo: false
-> ITER: 10
-> LAMBDA: 20
-> TIMEOUT: 3000)
-> Tempo totale: 114.77661s
[hadoop@ip-172-31-21-59 ~]$
```

```
INIZIO ELABORAZIONE
Operazioni preliminari
|__ Caricamento csv in RDD     20/11/24 15:54:03 INFO GPLNativeCodeLoader: Loa
20/11/24 15:54:03 INFO LzoCodec: Successfully loaded & initialized native-lzo
20/11/24 15:54:04 INFO ClientConfigurationFactory: Set initial getObject socke
 -> Tempo: 5.967773s            ( 15:54:09 )
|__ Partizione per ID utente    -> Tempo: 0.001617593s          ( 15:54:09 )
|__ Calcolo helpfulness locale  -> Tempo: 0.007532478s          ( 15:54:09 )
|__ Raggruppamento per idProd   -> Tempo: 0.019353319s          ( 15:54:09 )
|__ Calcolo link localmente     -> Tempo: 0.023903059s          ( 15:54:09 )

Inizio iterazioni (10)
|__ iterazione 1 -> Tempo: 0.077365145s                 ( 15:54:09 )
|__ iterazione 2 -> Tempo: 0.03523353s          ( 15:54:09 )
|__ iterazione 3 -> Tempo: 0.03239091s          ( 15:54:09 )
|__ iterazione 4 -> Tempo: 0.039208688s                 ( 15:54:09 )
|__ iterazione 5 -> Tempo: 0.030067421s                 ( 15:54:09 )
|__ iterazione 6 -> Tempo: 0.025940828s                 ( 15:54:09 )
|__ iterazione 7 -> Tempo: 0.02793903s          ( 15:54:09 )
|__ iterazione 8 -> Tempo: 0.030373849s                 ( 15:54:09 )
|__ iterazione 9 -> Tempo: 0.026464554s                 ( 15:54:09 )
|__ iterazione 10 -> Tempo: 0.028987246s                ( 15:54:09 )
20/11/24 15:54:09 INFO ClientConfigurationFactory: Set initial getObject socke
20/11/24 15:57:35 INFO MultipartUploadOutputStream: close closed:false s3://aw

Salvataggio dei risultati
|__ cartella di destinazione s3n://aws-logs-860100786633-us-east-1/24-11-2020-
|__ scrittura    -> Tempo: 205.86342s           ( 15:57:35 )

FINE ELABORAZIONE

Riepilogo:
-> file: s3://aws-logs-860100786633-us-east-1/videogames_25k_utenti.csv
-> demo: false
-> ITER: 10
-> LAMBDA: 20
-> TIMEOUT: 3000)
-> Tempo totale: 222.7688s
[hadoop@ip-172-31-10-17 ~]$
[hadoop@ip-172-31-10-17 ~]$
```

```
INIZIO ELABORAZIONE
Operazioni preliminari
|__ Caricamento csv in RDD     20/11/24 21:40:21 INFO GPLNativeCodeLoader: Loa
20/11/24 21:40:21 INFO LzoCodec: Successfully loaded & initialized native-lzo
20/11/24 21:40:22 INFO ClientConfigurationFactory: Set initial getObject socke
 -> Tempo: 6.3536997s            ( 21:40:27 )
|__ Partizione per ID utente    -> Tempo: 0.001705136s          ( 21:40:27 )
|__ Calcolo helpfulness locale  -> Tempo: 0.008261979s          ( 21:40:27 )
|__ Raggruppamento per idProd   -> Tempo: 0.016989022s          ( 21:40:27 )
|__ Calcolo link localmente     -> Tempo: 0.040356666s          ( 21:40:27 )

Inizio iterazioni (10)
|__ iterazione 1 -> Tempo: 0.06319401s          ( 21:40:27 )
|__ iterazione 2 -> Tempo: 0.036432162s                 ( 21:40:27 )
|__ iterazione 3 -> Tempo: 0.039644383s                 ( 21:40:27 )
|__ iterazione 4 -> Tempo: 0.031932797s                 ( 21:40:27 )
|__ iterazione 5 -> Tempo: 0.02893037s          ( 21:40:27 )
|__ iterazione 6 -> Tempo: 0.024489593s                 ( 21:40:27 )
|__ iterazione 7 -> Tempo: 0.03300747s          ( 21:40:27 )
|__ iterazione 8 -> Tempo: 0.031594776s                 ( 21:40:27 )
|__ iterazione 9 -> Tempo: 0.031768195s                 ( 21:40:27 )
|__ iterazione 10 -> Tempo: 0.029035639s                ( 21:40:27 )

Salvataggio dei risultati
|__ cartella di destinazione s3n://aws-logs-743342462495-us-east-1/24-11-2020-
20/11/24 21:40:27 INFO ClientConfigurationFactory: Set initial getObject socke
20/11/24 21:42:22 INFO MultipartUploadOutputStream: close closed:false s3://aw
|__ scrittura    -> Tempo: 114.712906s          ( 21:42:22 )

FINE ELABORAZIONE

Riepilogo:
-> file: s3://aws-logs-743342462495-us-east-1/videogames_25k_utenti.csv
-> demo: false
-> ITER: 10
-> LAMBDA: 20
-> TIMEOUT: 3000)
-> Tempo totale: 133.97305s
[hadoop@ip-172-31-21-59 ~]$
```

Conclusioni

- Maggiore è la grandezza del file in input più è evidente la potenza di calcolo di aws e l'importanza delle risorse
- Se pur l'algoritmo non sia troppo complesso i risultati da noi ottenuti mostrano l'effettiva efficienza dell'algoritmo implementato.