HADOOP

Scaricare i file pubmed.txt e metadata.HGNC.

Il file pubmed contiene circa 4000 articoli (solo abstract, autori, titolo dell'articolo, e dati di identificazione dell'articolo). Il file metadata.HGNC contiene dati relativi ai nomi dei geni (gene symbol). Per ogni gene le informazioni sono suddivise in tre colonne.

Implementare in hadoop:

1. un inverted index (output1) dove per ogni articolo vengono estratti per ogni frase (presente nell'abstract (considerare la voce "AB -" nel file pubmex.txt) e solo i termini presenti nel file gene symbols (metadati, considerare solo la seconda colonna del file metadata.HGNC):
    - Chiave gene o elenco di geni presenti nella stessa frase,
    - Valore: elenco documenti che contengono il gene o l'elenco di geni (valore "TI -" del file pubmed.txt").
2. Implementare PCY per il calcolo delle coppie di articoli frequenti minsupp>2 nel file inverted index (output1) ottenuto al punto. Per ogni coppia di articoli frequenti restituire l'elenco di tutte le chiavi.