Business Intelligence per Big Data

Progetti di analisi di dati

Data Base and Data Mining Group of Politecnico di Torino

AA 2020-2021 - *Politecnico di Torino*

Datasets

- Campioni della collezione di tweet a tema COVID-19
  - mantenuta dal Panacea Lab di Georgia State
- La collezione originale copre quasi un anno di tweet raccolti
  - è la più grande collezione Open Data di tweet su questo tema.
- Ciascun gruppo dovrà analizzare solo un sotto-campione di questi dati

Datasets

- Caratteristiche di ogni dataset:
  - 4000 tweet circa
  - unico file CSV, in cui ad ogni riga corrisponde un tweet;
  - per ogni tweet sono presenti le seguenti informazioni:
    - “favorite_count”, “source”, “text”, “is_retweet”, “created_at”, “retweet_count”.
    - Il campo “source” identifica il dispositivo da cui è stato prodotto il tweet. I restanti campi sono auto-esplicativi.

Datasets

- Lingue presenti nel dataset
  - Inglese
  - Francese
  - Spagnolo.
- I tweet sono campionati in periodi temporali diversi
  - l'evoluzione della pandemia è stata significativamente diversa.
- I dati sono protetti da copyright e di proprietà di DBDMG, per cui non sono modificabili o ridistribuibili.

Obiettivo

- Utilizzo di una (o più di una) tecnica di data mining per analizzare un dataset reale
  - Analisi del dataset per caratterizzare la distribuzione dei dati
    - Analisi esplorativa dei dati
  - Tecniche di analisi di dati
    - *Regole di associazione*
      - *Effettuare più sessioni di analisi*
      - *Variare gli indici di qualità (e.g., supporto, confidenza, lift)*
    - *Clustering*
      - *Effettuare più sessioni di analisi con uno o più algoritmi(e.g., K-Means, DBSCAN)*
      - *Variare i parametri di input (analisi di sensitività)*
      - *Valutare i diversi espergli indici di qualità (e.g., SSE)*

Regole

- Gruppi di due persone
  - Registrarsi sul google sheet
  - https://docs.google.com/spreadsheets/d/1q6QAnzCEajo4CT85SKVwLmPqWpnDmB_jwQNvAzQQXGk/edit?usp=sharing
- Ogni gruppo deve
  - Caratterizzare il dataset
  - Effettuare diverse sessioni di analisi su un dataset utilizzando il tool RapidMiner e/o altri tools noti al gruppo di studenti
  - Analizzare i risultati e sintetizzarli in grafici
  - Discutere come sfruttare la conoscenza estratta in un'applicazione di business

Regole

- Preparare una breve ma completa presentazione sulle attività svolte
  - Caratterizzazione del dataset
  - Analisi effettuata (e.g., configurazione ottimale dell'algoritmo selezionato)
  - Risultati migliori ottenuti e comparativa (di performance e qualità della conoscenza) tra algoritmi diversi
- Presentare i risultati in 15 minuti
  - 5 minuti di presentazione a testa e 5 di domande

Consulenze per il progetto

- Consulenza
  - 3-4 slot da 1,5h
  - Da definire negli slot delle prossime lezioni e/o fuori orario se necessario

Materiale da consegnare

- Preparare il materiale
  - Processo di rapid miner e file memorizzati nel repository
  - Insieme di lucidi
  - Sorgenti dei grafici (e.g., file excel)

Date importanti

- Consegnare i lucidi via mail a Tania Cerquitelli
  - Entro il 6 giugno 2021
    - Discussione durante le lezioni del 8/6 e 10/6
  - entro il giorno precedente la prova scritta
    - Le presentazioni saranno svolte a partire dal giorno successivo alla prova scritta
    - Primo e secondo appello (Giugno-Luglio)
- *Consegnare tutto il materiale tramite link ad una cartella condivisa*

Valutazione

- Ogni studente del gruppo sarà valutato con un punteggio in trentesimi
  - In caso di lode, viene considerato 32 (per calcolare il voto finale)
- Il voto della tesina sarà mediato con il seguente punteggio
  - Voto conseguito all'esame scritto incrementato di
    - 1/30 se lo studente ha consegnato l'esercitazione sulle tecniche di classificazione
    - 1/30 se lo studente ha consegnato l'esercitazione su MapReduce/MongoDB
  - La lode viene riconosciuta se il voto finale è >=31