Corso di Big Data Analytics and Business Intelligence
Prof. Antonio Picariello
Anno Accademico 2018/2019

Expert Finding

Identificazione utenti con conoscenza su un dato tema

Davide Lauretano
M63/792

Michele Pommella
M63/790

Davide Trimaldi
M63/799

Problema

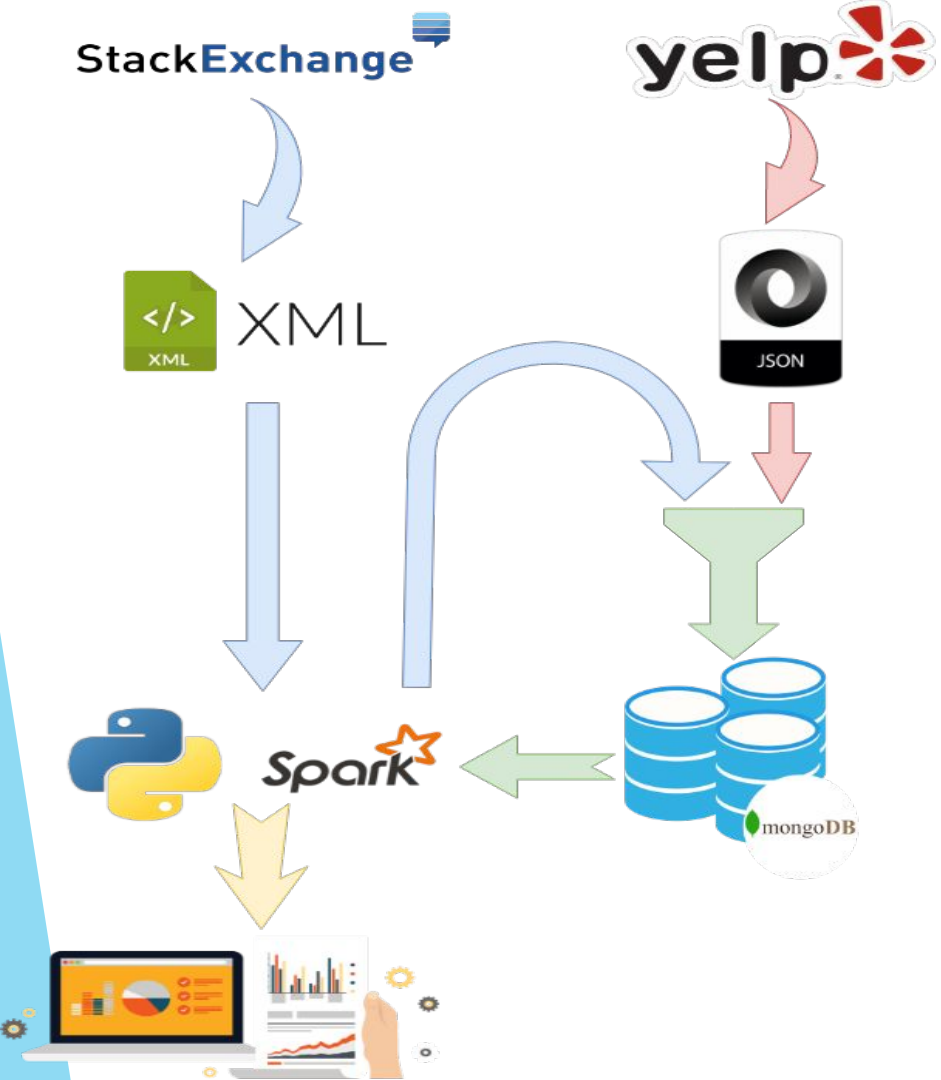
Ricerca di esperti per un dato argomento

Dataset

8 GB

60 GB

XML



Metodologia ed Architettura

MongoDB

- Database documentale per trattare aggregati strutturati
- I documenti BSON possono essere anche strutturalmente non identici.
- Gestione semplice ed efficace del dataset composto da file JSON e XML
- Ideale per applicazioni con grandi volumi di dati multi-strutturati e dall'elevato tasso di cambiamento
- Completa interoperabilità con il sistema Spark mediante il MongoDB Spark Connector

Spark

- Velocità sfruttando le ottimizzazioni in-memory;
- Framework unificato offrendo packages di librerie di alto livello (supporto a query SQL, Machine Learning, stream e graph processing);
- Semplicità includendo API facili da usare per operare su grandi dataset, come operatori per trasformare e manipolare dati semistrutturati.

Expert Finding

```
Input: topic
Result: experts
if topic composto da almeno una parola then
    Trasformazione topic in parole staccate, minuscole e con la prima
     lettera maiuscola per Yelp;
    Trasformazione topic in parole unite e minuscole per StackExchange;
    notYelp = False; notStack = False;
    Recupero da MongoDB le imprese di Yelp con categoria == topic;
    Recupero da MongoDB i post di StackExchange appartenenti alla
     collezione di topic;
    if numero imprese > 0 then
        Recupero da MongoDB le recensioni di Yelp e selezione degli
         attributi di interesse;
        Recupero delle recensioni relative alle imprese;
        Filtraggio recensioni con utilità > media(utilità);
        Riduzione ad almeno le 100 recensioni più utili;
        Recupero da MongoDB gli utenti di Yelp e selezione degli
         attributi di interesse;
        Recupero degli utenti autori delle recensioni più utili;
        Raggruppamento delle recensioni per utente e somma dell'utilità
         delle recensioni di uno stesso utente;
        Calcolo competenza come somma pesata degli attributi più
         importanti normalizzati;
        Ordinamento decrescente per competenza e selezione dei primi
         10 utenti;
    else
        notYelp = True
    if numero post > 0 then
        Filtraggio dei post di risposta e selezione degli attributi di
         interesse;
        Filtraggio dei post con score > media(score);
        Riduzione ad almeno i 100 post con score più alto;
        Recupero da MongoDB gli utenti di StackExchange appartenenti
         alla collezione di topic e selezione degli attributi di interesse;
        Recupero degli utenti autori dei migliori post;
        Raggruppamento dei post per utente e somma dello score dei
         post di uno stesso utente;
        Calcolo competenza come somma pesata degli attributi più
         importanti normalizzati;
        Ordinamento decrescente per competenza e selezione dei primi
         10 utenti;
    else
        notStack = True
    if notYelp AND notStack then
        print 'Topic inesistente';
else
    print 'Topic non selezionato';
```

Metrica

Somma pesata di attributi di interesse normalizzati.

Normalizzazione in [0,1] per rendere gli attributi omogenei

$$x_{normalizzato} = \frac{x - min(x)}{max(x) - min(x)}$$

Metrica

YELP

$$\sum_{1}^{7} w_i * x_{normalizato_i} \qquad w_i = \begin{cases} 0.25, & \text{se } x_i \in \{\text{topicuseful, elitedim}\} \\ 0.15, & \text{se } x_i \in \{\text{compliment hot}\} \\ 0.1, & \text{se } x_i \in \{\text{fans, compliment profile, useruseful}\} \\ 0.05, & \text{se } x_i \in \{\text{compliment list}\} \end{cases}$$

STACK EXCHANGE

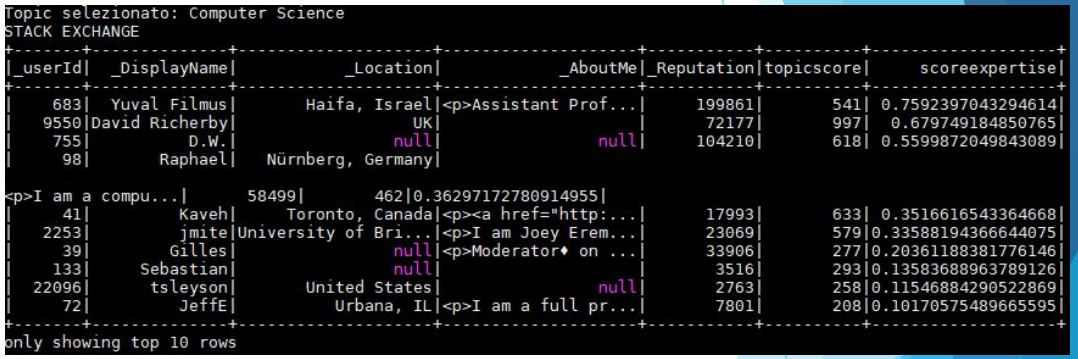
$$\sum_{1}^{2} 0.5 * x_{normalizato_i}$$

$$x_i \in \{\text{topicscore, reputation}\}$$

Risultati Sperimentali

Computer Science

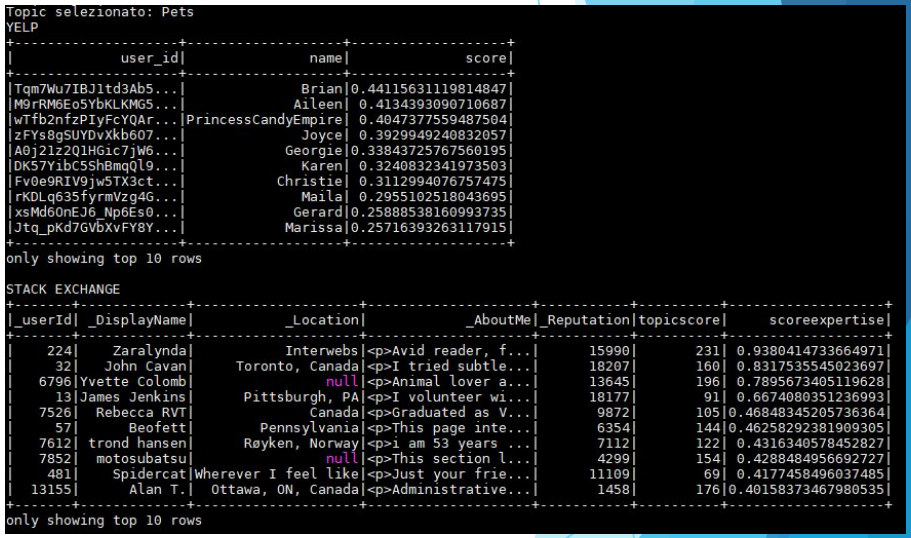
- Yuval Filmus: ricercatore del Technion
- Raphael: computer scientist esperto
- Kaveh: ingegnere del software di Google
- Jmite: dottorando della British Columbia
- JeffE: professore della Urbana-Champaign



Risultati Sperimentali

Pets

- Zaralynda: divide la casa con 3 gatti
- Yvette Colomb: amante degli animali e attivista per i loro diritti
- Trond Hansen: ha un gatto di nome Trinee ne ha sempre avuto uno
- Rebecca RVT: laureata come tecnico veterinario, lavora principalmente con animali (cani e gatti) e possiede 11 anni di esperienza con animali esotici
- James Jenkins : volontario presso la Rabbit Wranglers, istruisce e dà suggerimenti a chi vuole adottare un coniglio.

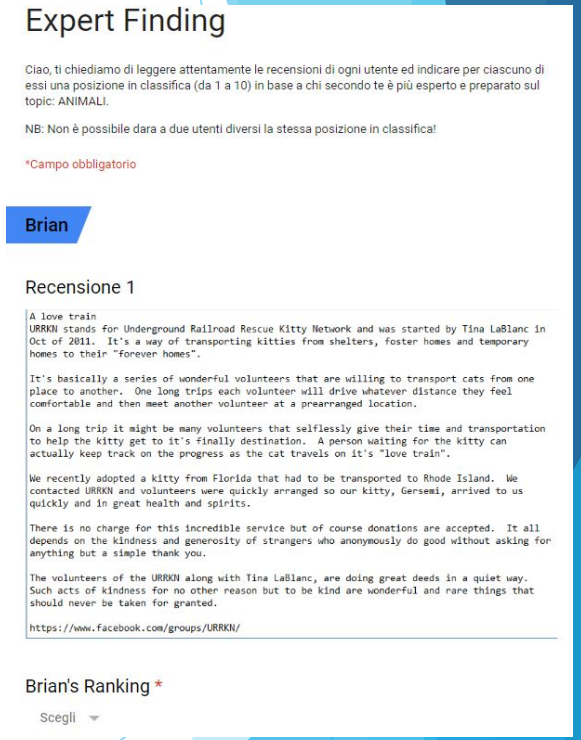


```
Topic selezionato: Pets
YELP
+--------------------+------------------+------------------+
|             user_id|              name|             score|
+--------------------+------------------+------------------+
|Tqm7Wu7IBJ1td3Ab5...|             Brian|0.44115631119814847|
|M9rRM6Eo5YbKLKMG5...|            Aileen| 0.4134393090710687|
|wTfb2nfzPIyFcYQAr...|PrincessCandyEmpire| 0.4047377559487504|
|zFYs8gSUYDvXkb6O7...|             Joyce| 0.3929949240832057|
|A0j21z2Q1HGic7jW6...|           Georgie|0.33843725767560195|
|DK57YibC5ShBmqQl9...|             Karen| 0.3240832341973503|
|Fv0e9RIV9jw5TX3ct...|           Christie| 0.3112994076757475|
|rKDLq635fyrmVzg4G...|             Maila| 0.2955102518043695|
|xsMd6OnEJ6_Np6Es0...|            Gerard|0.25888538160993735|
|Jtq_pKd7GVbXvFY8Y...|           Marissa|0.25716393263117915|
+--------------------+------------------+------------------+
only showing top 10 rows

STACK EXCHANGE
+-------+------------+------------------+------------------+----------+---------+------------------+
|_userId| _DisplayName|         _Location|          _AboutMe|_Reputation|topicscore|     scoreexpertise|
+-------+------------+------------------+------------------+----------+---------+------------------+
|    224|   Zaralynda|          Interwebs|<p>Avid reader, f...|     15990|      231| 0.9380414733664971|
|     32|  John Cavan|  Toronto, Canada|<p>I tried subtle...|     18207|      160| 0.8317535545023697|
|   6796|Yvette Colomb|              null|<p>Animal lover a...|     13645|      196| 0.7895673405119628|
|     13|James Jenkins|  Pittsburgh, PA|<p>I volunteer wi...|     18177|       91| 0.6674080351236993|
|   7526|  Rebecca RVT|            Canada|<p>Graduated as V...|      9872|      105|0.46848345205736364|
|     57|     Beofett|  Pennsylvania|<p>This page inte...|      6354|      144|0.46258292381909305|
|   7612| trond hansen|  Røyken, Norway|<p>i am 53 years ...|      7112|      122| 0.4316340578452827|
|   7852|  motosubatsu|              null|<p>This section l...|      4299|      154| 0.4288484956692727|
|    481|   Spidercat|Wherever I feel like|<p>Just your frie...|     11109|       69| 0.4177458496037485|
|  13155|     Alan T.|  Ottawa, ON, Canada|<p>Administrative...|      1458|      176|0.40158373467980535|
+-------+------------+------------------+------------------+----------+---------+------------------+
only showing top 10 rows
```

Risultati Sperimentali

- Sondaggio per ordinare gli utenti in base alla competenza sugli animali
- Partecipanti fidati possessori di animali
- Classifiche simili, tranne per Georgie
- Una parola può rappresentare differenti concetti



Expert Finding

Ciao, ti chiediamo di leggere attentamente le recensioni di ogni utente ed indicare per ciascuno di essi una posizione in classifica (da 1 a 10) in base a chi secondo te è più esperto e preparato sul topic: ANIMALI.

NB: Non è possibile dara a due utenti diversi la stessa posizione in classifica!

*Campo obbligatorio

Brian

Recensione 1

A love train
URRKN stands for Underground Railroad Rescue Kitty Network and was started by Tina LaBlanc in Oct of 2011. It's a way of transporting kitties from shelters, foster homes and temporary homes to their "forever homes".

It's basically a series of wonderful volunteers that are willing to transport cats from one place to another. One long trips each volunteer will drive whatever distance they feel comfortable and then meet another volunteer at a prearranged location.

On a long trip it might be many volunteers that selflessly give their time and transportation to help the kitty get to it's finally destination. A person waiting for the kitty can actually keep track on the progress as the cat travels on it's "love train".

We recently adopted a kitty from Florida that had to be transported to Rhode Island. We contacted URRKN and volunteers were quickly arranged so our kitty, Gersemi, arrived to us quickly and in great health and spirits.

There is no charge for this incredible service but of course donations are accepted. It all depends on the kindness and generosity of strangers who anonymously do good without asking for anything but a simple thank you.

The volunteers of the URRKN along with Tina LaBlanc, are doing great deeds in a quiet way. Such acts of kindness for no other reason but to be kind are wonderful and rare things that should never be taken for granted.

https://www.facebook.com/groups/URRKN/

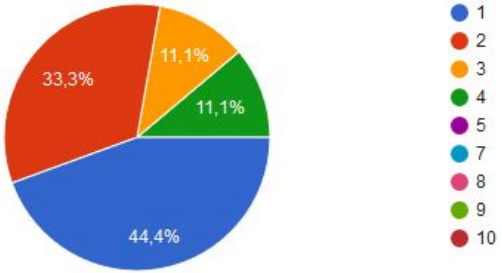
Brian's Ranking *

Scegli

Risultati sperimentali

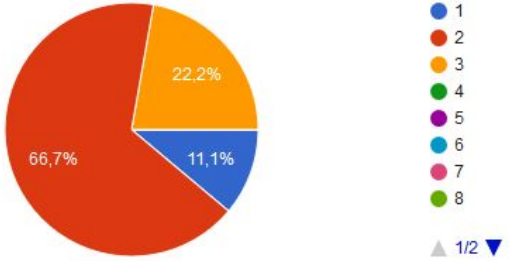
Brian's Ranking

9 risposte



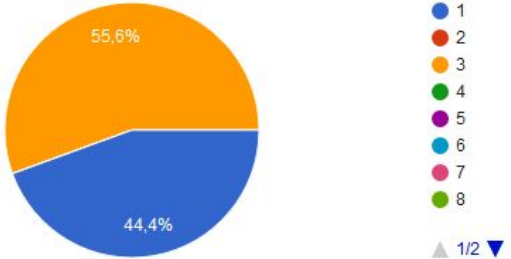
Aileen's Ranking

9 risposte



PrincessCandyEmpire's Ranking

9 risposte



Tempi e costi

- I tempi di esecuzione si attestano sui 12 minuti e sono dettati da Yelp
- Differenti configurazioni tecniche di Spark non apportano significativi vantaggi temporali
- Memoria centrale sufficiente per un'elaborazione efficiente
- Una fase di preprocessing onerosa potrebbe catalogare i dati di Yelp per topic nel database, riducendo i tempi di esecuzione ma provocando più ridondanza e maggior consumo di memoria
- Disco esiguo in capienza o volatile
- Trade off tempi-costi

Grazie per l’attenzione!