Linee guida per la parte del laboratorio che usa MongoDB

Rosa Meo

Linee guida

- La seconda parte del laboratorio del corso di Laboratorio avanzato di basi di dati richiede di adottare un database NOSQL

- Sviluppiamo i dizionari delle parole associate ai vari sentimenti come mappe da memorizzare in un database NOSQL

- Abbiamo scelto MongoDB

Ipotesi di lavoro

- Ipotizziamo di avere un certo numero di nodi-server *N* membri di un cluster di computer per effettuare la computazione parallela e distribuita

- Se non abbiamo a disposizione più di una macchina server, allochiamo tutti i nodi alla stessa macchina (localhost)

- Ogni nodo avrà in locale la propria istanza di database MongoDB e avrà caricato lo stesso dizionario delle parole derivanti dal preprocessing dei file delle risorse lessicali associate ai sentimenti

- Quest'ultimo dizionario, così come l'elenco delle altre risorse (elenco delle emoticons, emojis, slang words, stop words, etc), saranno caricati uguali nei database di ogni nodo (usate lo stesso programma di preprocessing che avete usato nel laboratorio con Oracle)

Proposta di mappa

- La mappa generale per il modello dei sentimenti è :

chiave: nome sentimento	valore: mappa annidata con il dizionario e le frequenze delle parole trovate nei Tweets

Mappa annidata

- La mappa annidata ha il seguente formato:

chiave: parola	valore: frequenza (numero dei Tweets) della parola in chiave

Requisiti proposti

1. Dividere l'insieme dei Tweets in porzioni di M Tweets e distribuire M_l Tweets di uno stesso sentimento (in *sentiment*) a ciascun nodo *l* del cluster

2. Implementare una funzione map(sentiment,M_l) che processa i Tweets nell'insieme di Tweets in M_l associati al sentimento *sentiment* e costruisce una mappa che ha la forma della mappa annidata vista nella slide precedente; ogni Tweet in M_l verrà processato secondo le stesse linee guida (pipeline sequenziale di elaborazione dei Tweets) viste già per il Laboratorio con Oracle e verificherà la presenza delle parole nella mappa (se non presenti, le aggiungerà alla mappa) come chiave e come valore avrà il numero dei Tweets in M_l con quella parola. Le emoji o emoticons trovate nei Tweets verranno trattate come se fossero parole.

Requisiti (cont.)

3.Implementare una funzione
Reduce(sentiment,M_l)
che prenderà dai vari nodi l le mappe M_l e farà il merge delle mappe che avranno lo stesso valore di sentiment. Il merge farà la somma delle frequenze delle stesse chiavi (parole) nelle varie mappe. produrrà una mappa M globale per ogni sentimento

4.Implementare una funzione
Filter(K)
che filtrerà le mappe M con le K chiavi che hanno il valore maggiore. Produrrà una mappa *MK* per ogni sentimento.
Le mappe MK saranno l'input al software di Word Cloud (che produrrà una nuvola di parole di dimensione variabile a seconda della frequenza).