Descrizione Progetto per il corso di Modelli e Architetture Avanzate di DataBase (MAADB):

Obiettivo del progetto di laboratorio:
Si analizzi il dataset dei tweet proposto (lo trovate nella sotto-directory *Tweets* di Dataset-Laboratorio). Ad ogni messaggio è associato un sentimento che si suppone sia il sentimento espresso dal messaggio. Più avanti, nella descrizione degli input descriviamo il dataset.

Tale dataset consiste in un grande numero di messaggi reali di utenti, scambiati sulla popolare piattaforma di microblogging *Twitter* (detti tweets).
Si richiede di trovare quali parole sono le più frequenti per ogni sentimento. I messaggi sono in inglese.
Per ottenere questo risultato viene richiesta la realizzazione di due soluzioni software che rispettivamente utilizzeranno due diversi database, per mettere a confronto le rispettive soluzioni applicative:

1. con l'utilizzo di un database relazionale, come Oracle (Express Edition), Postgres o MySQL (o MariaDB)
2. con un database NOSQL: MongoDB.

Vogliamo valutare quale delle due soluzioni sia più adatta a questo trattamento dati e fare esperienza di programmazione nei due ambienti. In particolare, per MongoDB, occorre utilizzare soluzioni software di tipo parallelo/distribuito, con uso di primitive di programmazione funzionale del tipo Map-Reduce.

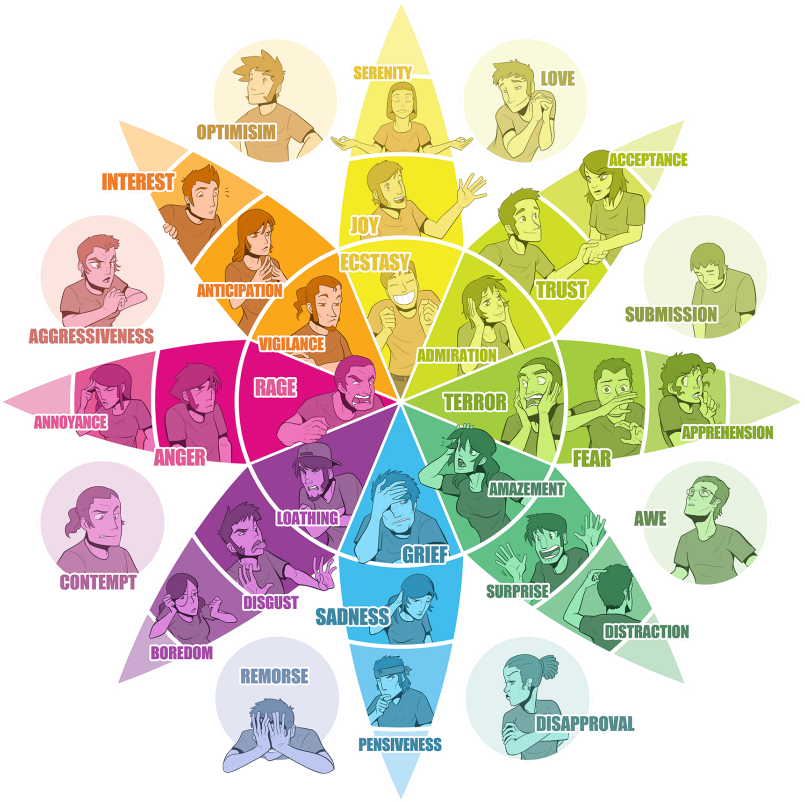

Figura 1 - Modello di Plutchik sulle emozioni

Input:

- *Risorse lessicali* (disponibili nella sotto-directory *Risorse-lessicali* di Dataset-Laboratorio)
per ogni sentimento vengono forniti dei file contenenti un elenco di parole (o insiemi di parole dette *n-grams*) associate a quel determinato sentimento dalla comunità degli studiosi di psicologia e del linguaggio.
- Il *modello delle emozioni* di tipo psicologico: la *Plutchik's Wheel*, illustrata nella figura soprastante. Le emozioni considerate sono le otto emozioni utilizzate nei petali del fiore, al secondo livello (ossia: *Anger, Anticipation, Joy, Trust, Fear, Surprise, Sadness, Disgust*).

- Dataset di tweets:
Il dataset consiste in un insieme di otto files contenenti ciascuno sessantamila messaggi reali di utenti, scambiati su Twitter (detti tweets). I file sono denominati con un sentimento associato ai messaggi.
I sentimenti sono quelli del secondo livello della Plutchik's Wheel.
- Librerie esterne per il Natural Language Processing (NLP): quella di Stanford (Risorse di Natural Language Processing). (Questa scelta è opzionale: le librerie di trattamento del linguaggio naturale possono essere sostituite con altre risorse da voi eventualmente preferite).

- Elenco emoji:
piccolo elenco di emoji ed altri insiemi di caratteri associati a ‘faccine’ (emoticons) per

facilitarne il riconoscimento nei tweet.

- Slang words:
elenco di alcune abbreviazioni con relativo significato per esteso.

Elenco delle operazioni richieste nell'elaborazione dei Tweets per estrarne le componenti (features) e contarne la frequenza nel corpus.

Come trattare le risorse lessicali
- Eliminare dalle risorse le parole composte (facilmente riconoscibili dalla presenza del carattere speciale '_') in quando da una prima analisi sommaria dei tweet non risultano sufficientemente presenti per essere rilevanti.
- Memorizzare la presenza delle parole nelle varie risorse (file) relative allo stesso sentimento.
- Caricare le risorse sui due database (relazionale e MongoDB).
- Calcolare la percentuale delle parole di ciascuna delle risorse (i diversi file relativi a uno dei sentimenti, rappresentati nella figura con il simbolo *X*) in cui le parole sono presenti anche nei messaggi tweet secondo lo schema mostrato qui sotto:

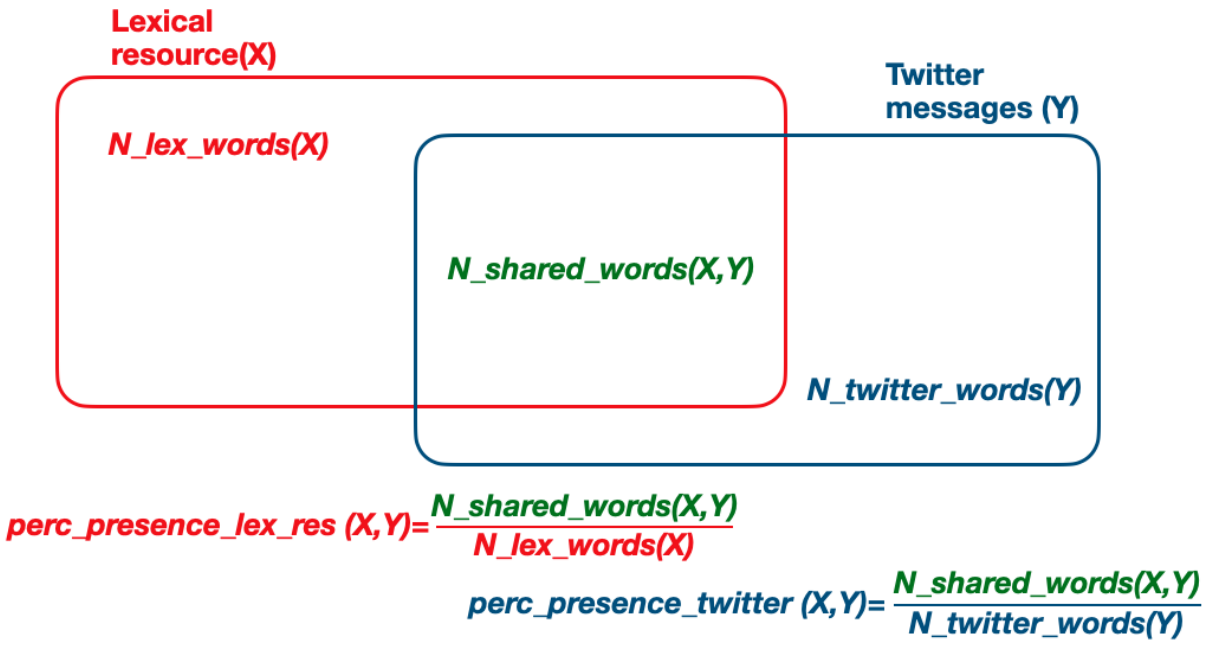

Figura 2 - Calcolo delle percentuali delle parole presenti nei tweets per la risorsa X e un determinato sentimento Y

In Figura 2 si mostrano due insiemi di parole che sono il risultato dell'elaborazione delle due sorgenti (risorse lessicali X per un determinato sentimento Y, in cui il numero totale delle parole è *N_lex_words(X)* e messaggi tweets etichettati con lo stesso sentimento Y, in cui il numero totale delle parole è *N_twitter_words(Y)*. L'intersezione tra i due insiemi contiene le parole comuni *N_shared_words(X,Y)* e le due formule mostrano come calcolare le percentuali.

Come trattare i tweet

- Riconoscere gli hashtag e memorizzarli in qualche modo (possibilmente anche contandone le occorrenze).
- Riconoscere gli emoji e rendicontarne la presenza nei vari sentimenti/tweet.
- Riconoscere le forme di slang
- Trattare la punteggiatura.
- Pulire i tweet dalla presenza delle parole tipiche anonimizzate USERNAME e URL.
- Eliminare le stop words.
- Lemmizzare le parole per permettere il match con le risorse lessicali
- Conteggiare la presenza nei vari tweet delle parole associate a un determinato sentimento.
- Memorizzare le 'nuove' parole trovate nei tweet ma assenti nelle risorse fornite (se alla fine del conteggio saranno altamente presenti avremo trovato nuova parole da aggiungere alle risorse o avremo creato una risorsa aggiuntiva!)

Quali risultati produrre dall'elaborazione dei tweets

- Visualizzare per ogni sentimento una *word cloud* con le parole maggiormente presenti nei tweet (la grandezza del carattere nella word cloud è proporzionale alla frequenza nei messaggi tweet). Si può creare una word cloud con le *x* parole più frequenti, con *x* parametro stabilito dall'utente.
- Creare una word cloud anche per le emoji e le emoticons (due cloud apposite, in quanto le frequenze tipiche sono diverse che nelle parole).
- Calcolare e mostrare un istogramma per ciascun sentimento con le percentuali delle parole delle risorse lessicali presenti nei tweets.
- Raccogliere le parole "nuove" presenti nella sorgente Tweet ma non nelle risorse lessicali: così abbiamo costruito una nuova risorsa lessicale, adatta a rappresentare i messaggi tweet.