# sentiment analysis software

Autore: Andrea Cattaneo  
Matricola: 909227  
Data: 06/04/2018  
Versione del documento: 1.0

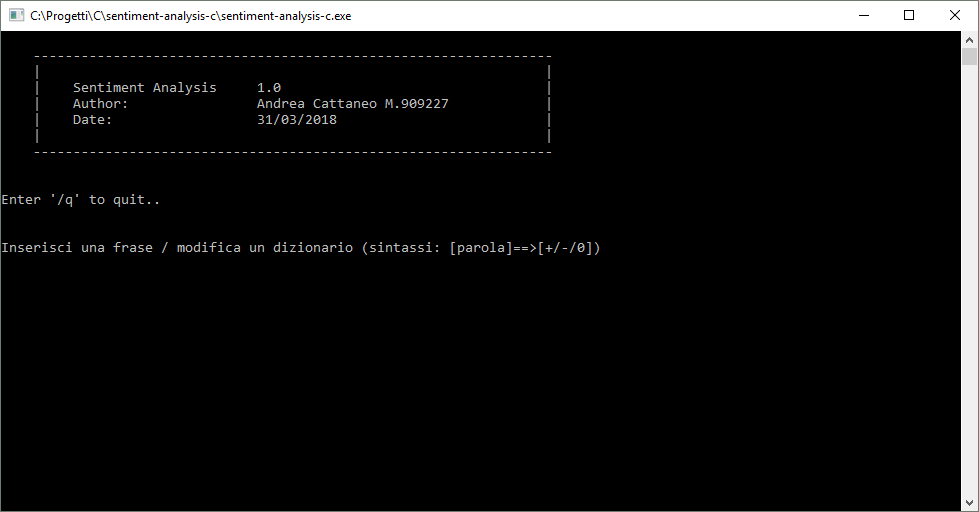
## OBIETTIVO DEL SOFTWARE

L’obiettivo del software realizzato è quello di effettuare la sentiment analysis di una frase inputata dall’utente. A seguito dell’analisi viene restituito un valore che indica se la frase inserita ha un’accezione positiva o negativa.

## UTILIZZO

L’utente avvia il programma ed il programma propone 3 possibili azioni:

1. Inserire una frase da valutare
2. Modificare un dizionario tramite l’apposita sintassi (vedi modifica dei dizionari)
3. Terminare l’esecuzione inserendo la stringa “/q”

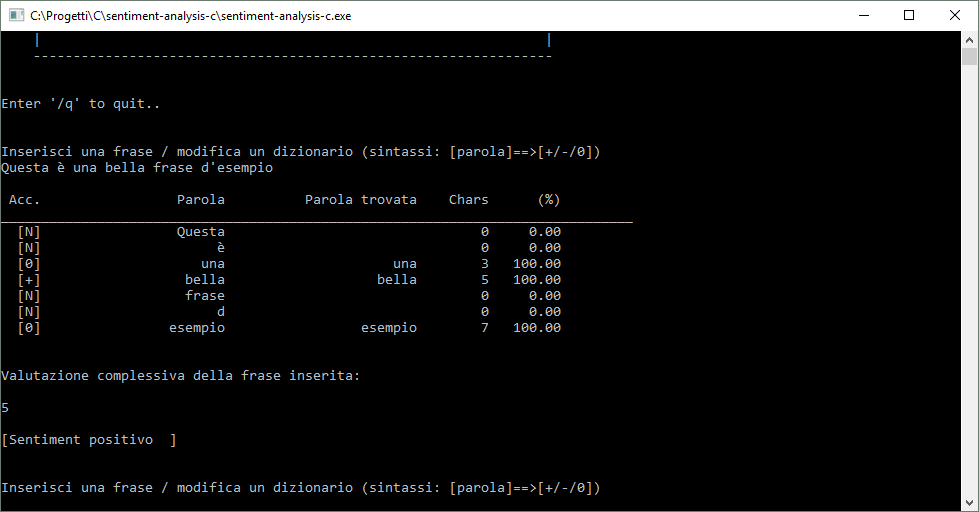


Vengono di seguito illustrate le operazioni di valutazione e codifica

### valutazione di una frase

L’operatore scrive una frase a suo piacimento e ne conferma l’inserimento con la pressione del tasto “Enter”.

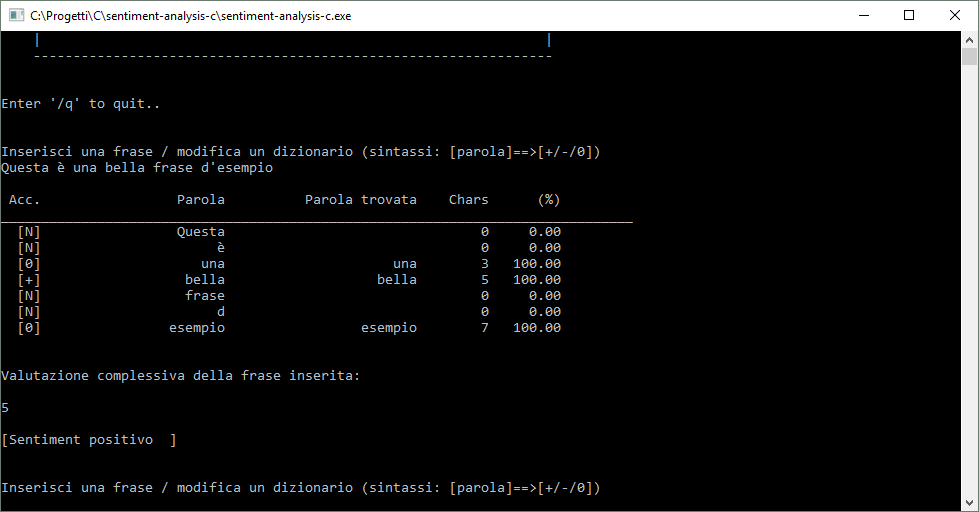
Come prima cosa viene restituita una tabella che riporta le parole valutate ed alcuni dati utilizzati durante il processo.



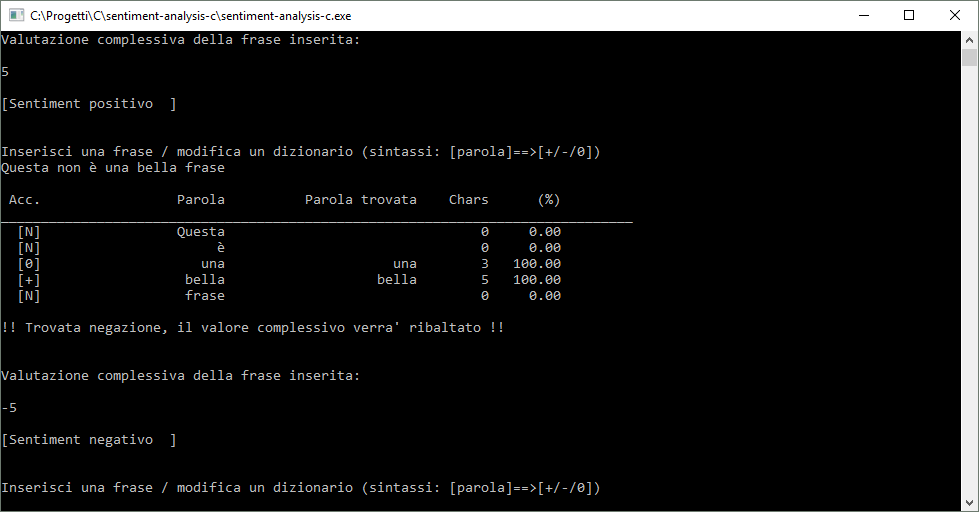
Le colonne hanno il seguente significato:

“**Acc**.”: Accezione della parola, viene indicato se la parola è risultata di accezione positiva (+), negativa (-) o neutra (N/0). Una parola viene valutata neutra se non viene trovata o se viene trovata nella lista delle parole “ignoreList.txt”. Questo perché i vocabolari inseriti nel software si concentrano sulle parole positive e negative. Le neutre si ottengono per esclusione o filtraggio di quelle che si vuole ignorare.   
“**Parola**”: La parola analizzata  
“**Parola trovata**”: La parola trovata nei dizionari. E’ la parola che ha la maggior somiglianza con la parola inputata.  
“**Chars**”: E’ la quantità di caratteri corrispondenti tra la parola di input e la parola trovata.  
“**%**”: E’ la percentuale di corrispondenza tra la parola di input e la parola trovata.

Dopo la tabella, il software riporta la valutazione dell’intera frase inserita. Un valore maggiore di zero indica una frase con accezione positiva, un valore minore di zero una frase con accezione negativa ed infine il valore zero una valutazione neutra. Questo risultato si ottiene sommando tutti i punteggi (caratteri corrispondenti) delle parole trovate.



Se viene trovata una keyword di negazione all’interno della frase il risultato viene ribaltato:



Il ribaltamento avviene solo una volta, le doppie negazioni vengono valutate come le singole perché spesso nella lingua italiana non vanno ad invertire nuovamente il significato (non sempre per lo meno). L’unica negazione inserita nel dizionario delle negazioni è la parola “non”.

### Modifica dei dizionari

A seguito di una valutazione potrebbe accadere che una parola venga trovata nel dizionario sbagliato, o che ci siano parole molto simili che necessitano di essere specializzate in modo da ottenere una valutazione migliore.  
Per farlo è sufficiente inserire la parola della quale modificare il dizionario seguita dalla keyword “==>”e quindi dal carattere “+”, “-“ oppure “0”. La parola sarà scritta rispettivamente nel dizionario positivo, negativo o neutro.  
Quando inseriamo una parola in un dizionario essa sarà anche rimossa dagli altri due.

Esempi di modifica dei dizionari:  
*splendido==>+  
bruttino==>-  
legno==>0*

## Valutazione delle parole

Per determinare se una parola trovata nel dizionario corrisponde alla parola inserita, si valuta la quantità di caratteri corrispondenti tra l’inizio di una parola e l’inizio dell’altra. Esse devono iniziare nello stesso modo. Si valuta il valore di caratteri coincidenti come punteggio della parola. La quantità di caratteri corrispondenti deve essere almeno del 70%. La parola più lunga tra la cercata e la trovata rappresenta il 100%.

Il motivo per cui vengono valutati solo i caratteri identici nella parte iniziale della parola è che con buona probabilità due parole che hanno inizio uguale hanno la stessa radice e quindi un significato simile.

Se la parola trovata coincide con i caratteri della stringa cercata per l’intera lunghezza della parola cercata, ma ha una lunghezza maggiore (la cercata è inferiore del 70% della trovata), la parola viene scartata. Il motivo è che essendo la parola trovata molto più lunga della cercata, potrebbe avere significato completamente diverso.

Ogni parola inserita viene confrontata con tutte le parole dei dizionari. La parola con corrispondenza percentuale maggiore è quella che viene considerata come “parola trovata” (a patto che superi la soglia minima del 70%).

Una parola trovata nel dizionario delle parole neutre o una parola non trovata assume valore zero. Al contrario una parola trovata nei dizionari positivo o negativo, assume appunto un segno positivo o negativo rispettivamente. In questo caso il modulo del punteggio è il numero di caratteri corrispondenti tra la parola cercata e la parola trovata.

Una parola corrispondente molto lunga vale più punti di una parola molto corta. Il motivo di questa scelta è che se una parola è più lunga tendenzialmente ha una peso maggiore, ad esempio quando stiamo analizzando dei superlativi.

*Bello < Bellissimo  
Brutto < Bruttissimo*

Oltre ai superlativi, anche normali aggettivi tendono ad essere più lunghi se hanno un significato di maggior peso:

*Bello < Fantastico  
Brutto < Schifoso < Orripilante*

## Valutazione della frase

La valutazione complessiva della frase viene ottenuta sommando la valutazione di ogni singola parola ed ottenendo così un punteggio finale positivo o negativo a seconda del risultato.  
In questa valutazione hanno molto peso le parole espressamente positive o negative anche se come quantità sono decisamente inferiori. Questo perché alle parole neutre che vanno per la maggiore è stato attribuito un punteggio zero.

## DIZIONARI

Per ottenere performance migliori i dizionari vengono caricati in memoria in modo da essere confrontabili velocemente con le parole inserite. Il caricamento avviene all’avvio del software oppure a seguito della modifica dei dizionari. Il processo consiste nell’allocare in memoria un array di stringhe tramite la funzione “*malloc*”. Una volta effettuata l’allocazione si leggono i file txt dei dizionari e si popolano gli array con le stringhe di testo lette nei file. Alla modifica dei dizionari, prima di effettuare un nuovo caricamento in memoria, viene eseguita la “*free*” della memoria liberando così spazio dallo “*heap*” prima di occuparla nuovamente con un nuovo array di valori.

## PROGRAMMA

Il programma si avvia nel **main.c** che è il punto d’ingresso. Esso richiama il caricamento dei dizionari all’avvio e gestisce gli inserimenti dell’utente andando poi a richiamare le funzioni per la valutazione del testo, le funzioni per la modifica dei dizionari o la chiusura del programma.

## MODULI

Per rendere più leggibile il software è stato organizzato in moduli suddivisi per macro-funzionalità. Ogni modulo ha un file di interfaccia \*.h che espone solo le funzioni necessarie. Gli altri metodi e variabili non necessarie dall’esterno del modulo, sono incapsulate non esponendole nei file di interfaccia e semplificando così l’utilizzo del modulo stesso. Di seguito vengono elencati ed illustrati i principali moduli:

**common** Contiene funzionalità comuni che possono essere utilizzate in qualsiasi parte del programma.

**fileAccess** Modulo di accesso ai file. L’implementazione contiene funzioni di lettura / scrittura da file di testo esterni.

**sentiment** Modulo per la valutazione del sentiment di una frase. Espone la funzione di valutazione del sentiment della frase chiamata “**LaunchSentimentAnalysis**”. In questo modulo viene anche definita la struttura “*WordEvaluation”* che viene utilizzata per salvare una parola trovata, numero di caratteri corrispondenti e percentuale di corrispondenza.

**vocabulary** Modulo contenente la gestione dei vocabolari e la loro esposizione alle altre parti del programma.   
Le funzionalità disponibili sono:  
-modifica del contenuto dei vocabolari tramite la funzione “*LaunchVocabularyEditing*”. Esso riceve in ingresso la frase (*statement*) e la stringa per lo splittaggio tra parola e attributo positivo/neutro/negativo (*vocabEditSplitter*).  
-verifica se una parola è contenuta nel dizionario delle negazioni tramite “*NegationVocabularyContains”*   
-caricamento in memoria dei dizionari tramite funzione “*LoadAllVocabularies*”  
-liberazione della memoria dedicata ai dizionari tramite funzione “*FreeAllVocabularies*”

## FONTI ESTERNE

I dizionari positivi e negativi sono stati reperiti da questo sito:

<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>

Una volta scaricati sono stati tradotti con l’ausilio di Google Translate ed inseriti nella cartella Data del software.