Lexicon-Based Sentiment Analysis

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

Centro Interdisciplinare di Ricerche per la Computerizzazione
dei Segni dell'Espressione (CIRCSE)

LEXICON-BASED SA

- La cartella lexicon-based-SA contiene 5 file:
  - doAnalysis-new.py: script in python
  - hr.csv: dataset su cui calcolare il sentiment, tweet in italiano pubblicati durante la messa in onda in tv di “Harry Potter e i doni della morte - parte 1” nel 2020 e scaricati con Twint (delimitatore: virgola)
  - W-MAL.tsv: lessico di sentiment per l’italiano (delimitatore: tab)
  - lexicon_easy.csv: lessico di sentiment per l’inglese (delimitatore: virgola)
  - README.md: file di descrizione dello script originale

N.B. Tutti i file si possono aprire con un editor di testo (Sublime Text)
Il file .csv e .tsv si possono aprire anche come fogli di calcolo

LEXICON-BASED SA

- APPROFONDIMENTI

- repository dello script originale: https://github.com/stepthom/lexicon-sentiment-analysis
- paper sul lessico W-MAL: http://ceur-ws.org/Vol-2769/paper_36.pdf
- paper sul lessico inglese (MPQA subjectivity lexicon): http://people.cs.pitt.edu/~wiebe/pubs/papers/emnlp05polarity.pdf

LEGGIAMO IL FILE PYTHON

- Apriamo il file doAnalysis-new.py con Sublime Text: lo script è stato scritto pensando ai tweet ma è applicabile a qualunque tipo di testo
  - Riga 55: lettura del file contenente il testo da processare, adesso hr.csv ma si può cambiare per applicare lo script ad altri dati
  - Riga 56: identificazione del tipo di delimitatore tra campi, da cambiare se il delimitatore non è la virgola

    delimiter=',' → virgola

    delimiter='\t' → tab
  - Righe 62-64: struttura del file di input, la seconda colonna deve essere un numero intero (int)
  - Righe 67-88: si riduce tutto in minuscolo, si eliminano i retweet, il carattere #, le vocali doppie (“beeeellooooooo” → “bello”)

    Le righe 83-88 sono da commentare se si processano dati in inglese: aggiungere il carattere # all’inizio di ogni riga

LEGGIAMO IL FILE PYTHON

- Riga 98: lettura del lexicon, adesso W-MAL.tsv ma si può cambiare per usare altri lexicon
- Riga 99: identificazione del tipo di delimitatore tra campi, da cambiare se il delimitatore non è la virgola
- Righe 100-101: se si lavora con un dataset inglese, decommentare queste righe (eliminare il carattere # ad inizio riga) e commentare le righe 98-99 (aggiungendo il carattere # ad inizio riga)
- Riga 104: struttura del file del lexicon, lo score di sentiment deve essere un numero decimale (float)

  Se si usa un lessico con numeri decimali cambiare float in int
- Righe 107-119: calcolo del sentiment in base al lexicon
- Righe 123-148: scrittura dell'output

USIAMO LO SCRIPT

1. Aprire il terminale
2. Digitare cd poi spazio poi trascinare la cartella lexicon-based-SA poi premere lo invio
3. Digitare il seguente comando e premere invio

   ```
   python ./doAnalysis-new.py
   ```

   Se non funziona provare una delle seguenti opzioni:
   - sostituire python con python3
   - su Windows: sostituire ./ con .\
   - sostituire python con py
4. Digitare il seguente comando poi premere il tasto Invio (SOLO per Windows): `set PYTHONIOENCODING=utf8`
5. Digitare il seguente comando poi premere il tasto Invio

   ```
   python doAnalysis-new.py > hr-out.txt
   ```

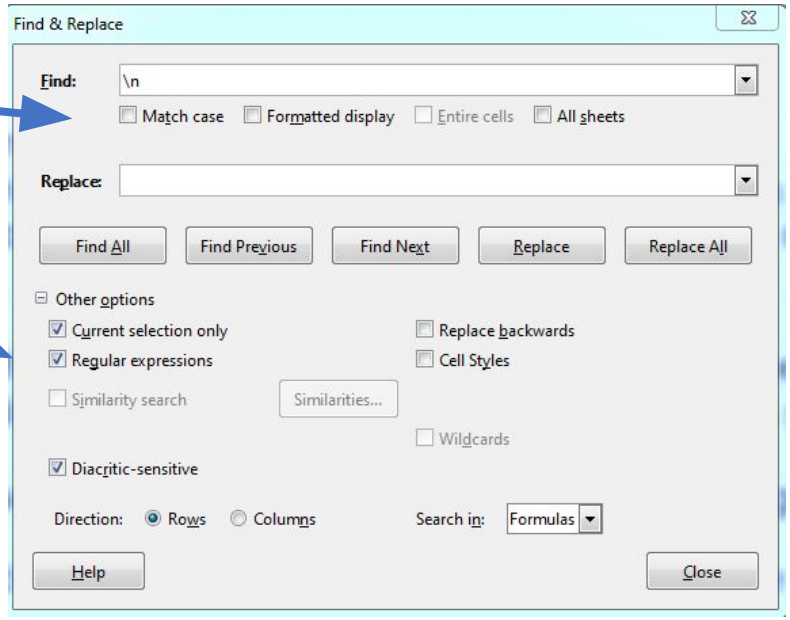
In questo modo l'output viene salvato in un file nella cartella lexicon-based-SA (hr-out.txt ma potete scegliere un nome a vostro piacimento)

PREPARIAMO UN DATASET

1. Aprire un file di output di Web scraper, copiare la colonna con il testo delle recensioni e incollarle su Sublime Text

N.B. Il testo di ogni recensione deve apparire **su una sola riga**. Se così non fosse: fare replace sul programma di foglio di calcolo

Su **LibreOffice Calc**: usare le espressioni regolari sostituendo \n (corrispondente all'a capo) con uno spazio



Su **Windows**: https://excelacademy.it/2145/excel-come-rimuovere-i-ritorni-a-capo-interruzioni-di-riga-dalle-celle/

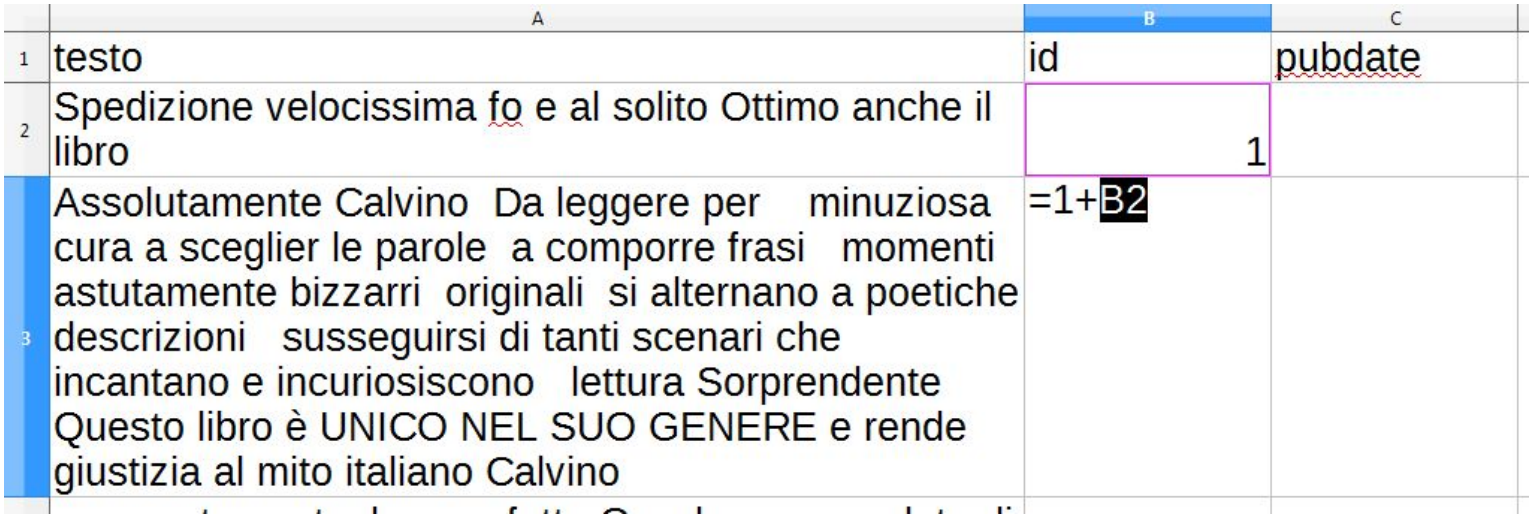
PREPARIAMO UN DATASET

2. Su Sublime Text, dopo aver incollato il testo, aprire la finestra di Replace (ctrl+h su Windows; alt+cmd+F su Mac), selezionare le opzioni come da figura che segue e sostituire la punteggiatura con uno spazio

3. Copiare il nuovo testo e incollarlo come prima colonna di un foglio di calcolo: aggiungere una riga per l'etichetta (ad es. testo)
4. Aggiungere due colonne (una con etichetta id e una con etichetta pubdate)

PREPARIAMO UN DATASET

5. Se si vuole aggiungere un numero intero progressivo nella colonna id: digitare 1 nella cella corrispondente alla prima recensione, digitare =1+B2 nella cella sotto e premere invio poi trascinare la cella B3 fino in fondo alla lista



	A	B	C
1	testo	id	pubdate
2	Spedizione velocissima fo e al solito Ottimo anche il libro	1	
3	Assolutamente Calvino Da leggere per minuziosa cura a sceglier le parole a comporre frasi momenti astutamente bizzarri originali si alternano a poetiche descrizioni susseguirsi di tanti scenari che incantano e incuriosiscono lettura Sorprendente Questo libro è UNICO NEL SUO GENERE e rende giustizia al mito italiano Calvino	=1+B2	

PREPARIAMO UN DATASET

6. Nelle celle di pubdate è possibile copiare la data di pubblicazione della recensione oppure un qualunque testo
7. Salvare il file come csv o tsv o usando un programma di fogli di calcolo oppure copiando e incollando il contenuto del file su Sublime Text e salvandolo da lì: verrà salvato con il tab (\t) come delimitatore
8. Aprire doAnalysis-new.py e modificare le righe 55 e 56 (nome del file e tipo di delimitatore)
9. Rilanciare lo script

GRAZIE!

Email: rachele.sprugnoli@unicatt.it
Twitter: @RSprugnoli