Analisi del corpus epistolare di Italo Svevo

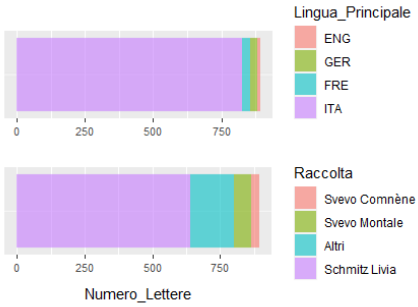
Progetto di Viol Alessandro – IN2000163

Descrizione del problema

A partire da una raccolta epistolare di Italo Svevo si vogliono estrarre informazioni inedite ai più sulla vita dell'autore. Per ottenere questa visione più personale di Italo Svevo, vogliamo applicare delle tecniche di text mining per portare alla luce le tematiche di cui lo scrittore è solito parlare ed evidenziare in che modo queste siano in relazione con lo stato d'animo dell'autore, con i suoi corrispondenti e con l'anno storico.

Descrizione dei dati

Il corpus è composto da 894 lettere che riportano informazioni su data, nome e città del mittente e del destinatario, la lingua principale e le lingue con cui è scritta la lettera.



Le lettere, infatti, fanno spesso uso concomitante di più lingue, principalmente italiano, francese, inglese e tedesco, con la frequente presenza di frasi in dialetto triestino.

Per ogni lettera è a nostra disposizione anche il nome della raccolta epistolare di provenienza, solitamente intitolata al corrispondente di Svevo. Ciò ci permette di evidenziare che le lettere non sono equamente distribuite tra queste raccolte, e quindi tra i vari interlocutori. Anche la distribuzione delle lettere negli anni non è equa. Si conclude quindi che il dataset non è bilanciato.

Soluzione proposta e metodi di verifica

Per l'estrazione delle tematiche, si è deciso di stimare un modello a 4 topic di Latent Dirichlet Allocation mediante l'algoritmo VEM[1] incluso nel pacchetto topicmodels di R. Il numero di topic del modello è stato scelto tramite un confronto tra le misure di log-likelihood e perplexity ottenute sui modelli costruiti da 2 a 20 topic. Il seed per l'algoritmo VEM è stato scelto confrontando i modelli prodotti da diversi valori del parametro, decidendo infine per il valore 10211 che a nostro giudizio portava al modello migliore.

Per quanto riguarda la parte di sentiment analysis, si è voluto utilizzare il NRC Emotion lexicon[2] incluso nel pacchetto syuzhet che ci permette di rilevare, per ogni documento, un punteggio su 10 indici emotivi. Viste le dimensioni variabili delle lettere, i punteggi attribuiti a ciascun documento necessitano di essere normalizzati.

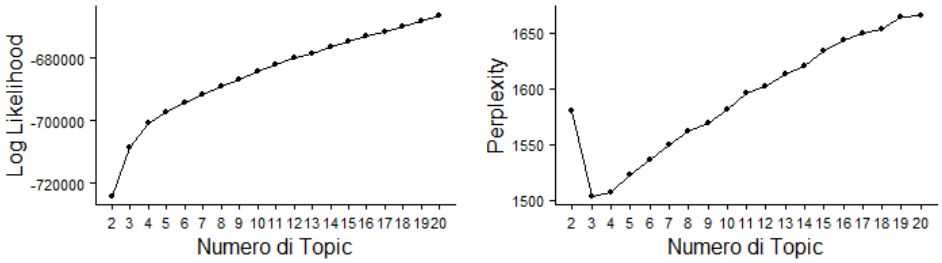
Per verificare la qualità della sentiment analysis, in mancanza di misure di qualità, ci si è limitati ad esplorare alcuni casi limite che si sono riscontrati, come le lettere con punteggi nulli su tutti gli indici o con dei valori molto alti.

Procedura sperimentale

Per poter superare il problema posto dalla compresenza di più lingue nelle lettere, si è deciso di considerare le sole 826 lettere la cui lingua principale è l'italiano. Questo poiché il rapporto tra il numero di parole in lingua straniera e quello delle parole italiane è così basso dal renderle simili al rumore dal punto di vista dell'algoritmo VEM.

Il testo del corpus subisce poi una fase di preprocessing in cui vengono rimosse le maiuscole, la punteggiatura e le stopwords della lingua italiana. Per il topic modeling vengono eseguite delle operazioni ulteriori che non vengono invece eseguite per la sentiment analysis. In particolare, viene eseguito lo stemming del testo e vengono eliminate delle stopwords aggiuntive, determinate osservando i topic modellati e tenendo conto della struttura caratteristica della lettera (*schmitz-ettore-signor-signore-signora-mano-mani-lettera-lettere-parola-fare-cosa-cose-così-tanto-tanti-poi-quando-prima-molto-caro-cara-carissimo-carissima-carissimi-sua-suo-suoi*).

Viene a questo punto eseguito l'algoritmo VEM per la stima del modello LDA. Prima, tuttavia, è stato necessario determinare quale sia il numero ottimale di topic da ricercare.



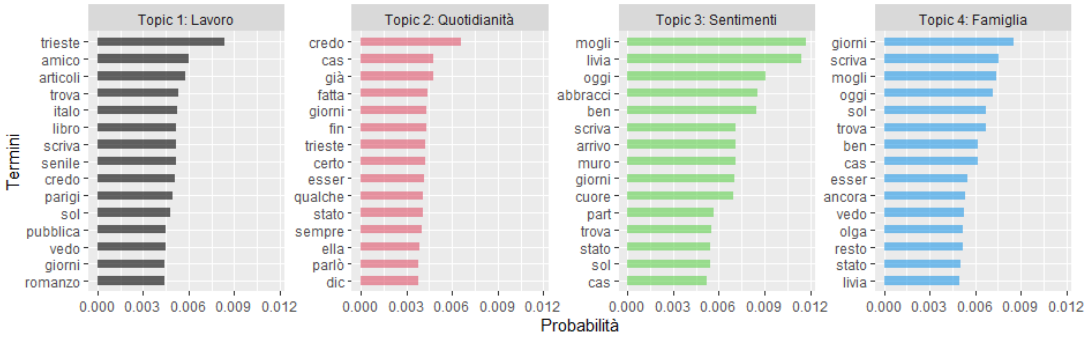
Per fare ciò, sono stati determinati i modelli LDA da 2 a 20 topic e sono stati confrontati i valori delle log-likelihood (più è alto l'indice, migliori le prestazioni) determinate dall'algoritmo VEM. Inoltre, tramite 5-fold cross validation, sono stati calcolati i valori di perplexity (più è basso l'indice, migliori le prestazioni) sullo stesso range di topic. Da questa analisi si è deciso di estrarre 4 topic.

Una volta ottenuta la distribuzione di parole di ciascun topic, si tenterà di invertire l'operazione di stemming mediante il metodo stemCompletion applicato usando come dizionario il corpus di partenza. Ciò completerà le parole troncate con le parole compatibili più corte del dizionario. Viene fatto questo per avere dei dati più facilmente analizzabili, di fatto usando lo stemming e stemCompletion per approssimare un'operazione di lemmatization, che non si è potuta applicare per mancanza di un adatto lessico in italiano.

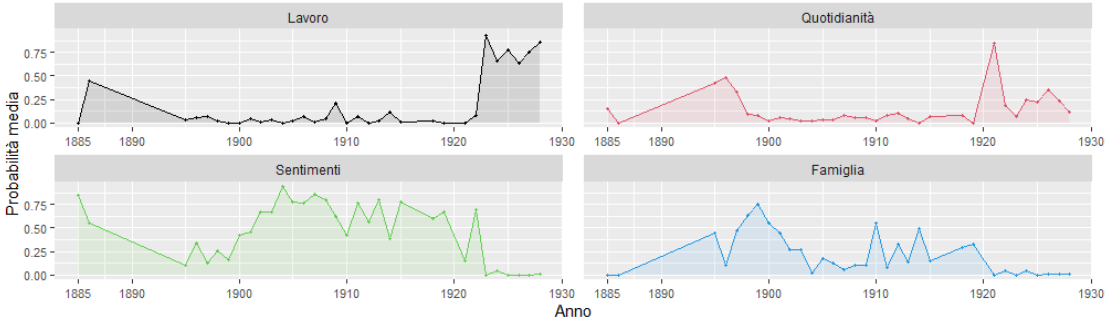
Ottenuto il modello, abbiamo a disposizione le probabilità che ad ogni documento sia associato un topic. Poiché è noto l'anno in cui è stata spedita ciascuna lettera, per mettere in evidenza come variano i topic di interesse di Svevo è bastato considerare la media delle probabilità dei topic su tutte le lettere di ogni anno. In modo analogo, una volta stilata una lista di tutti i corrispondenti, possiamo visualizzare quali sono le persone con cui l'autore discute di più per ogni tematica, con l'accortezza di considerare le sole persone con cui Svevo abbia scambiato più di una lettera.

Per la sentiment analysis, una volta finito il preprocessing, viene utilizzato il NCR Emotional lexicon per calcolare per ogni lettera i punteggi nelle emozioni di rabbia, anticipazione, disgusto, paura, gioia, tristezza, sorpresa, fiducia, negatività e positività. Poiché questo lessico è disponibile nelle 4 lingue principali usate dall'autore, vengono sommati i punteggi di tutte e quattro per ognuna delle 826 lettere. I risultati ottenuti vengono poi normalizzati dividendoli per il numero di parole del testo preprocessato dal quale sono stati estratti. A questo punto basta ricavare la media dei punteggi raggruppando le lettere per anno, topic e interlocutori per mettere in evidenza le relazioni tra questi e le emozioni dello scrittore.

Risultati ottenuti



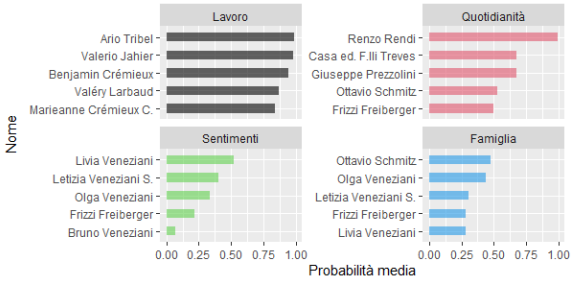
Nonostante gli sforzi profusi, è stato difficile ottenere dei topic chiaramente distinguibili e facilmente interpretabili. Nei 4 topic individuati, la principale difficoltà si è incontrata nell'interpretare le tematiche 3 e 4. Utilizzando il POS Tagging nel preprocessing del topic modeling, si sarebbero potuti mantenere solo i nomi e i verbi all'interno del testo, riducendo il rumore e permettendo di ottenere dei risultati migliori. Anche l'uso di lemmatization avrebbe contribuito a una miglior leggibilità dei risultati. Infine, poiché è stato provato che le misure di perplexity e log-likelihood spesso non rappresentano una misura di qualità dei modelli in linea con il giudizio umano, si ritiene che usare una misura alternativa come la coherence per individuare il numero di topic possa risultare beneficiario.



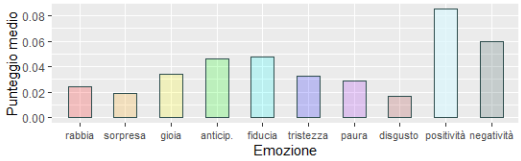
Si osserva che, benché i topic Famiglia e Sentimenti si mantengono presenti pressoché per tutta la vita dell'autore, poco dopo il 1920 vengono soppiantati dai topic Lavoro e in

minor misura Quotidianità, probabilmente a causa del successo del romanzo della Coscienza di Zeno.

I legami tra argomenti e persone rispettano le aspettative, con Sentimenti e Famiglia associati ai famigliari e Lavoro associato ai colleghi. Più inaspettate sono le persone legate al topic di Quotidianità, che appartengono anch'esse alla sfera lavorativa di Svevo. Si ritiene quindi di aver male interpretato quest'ultimo tema.

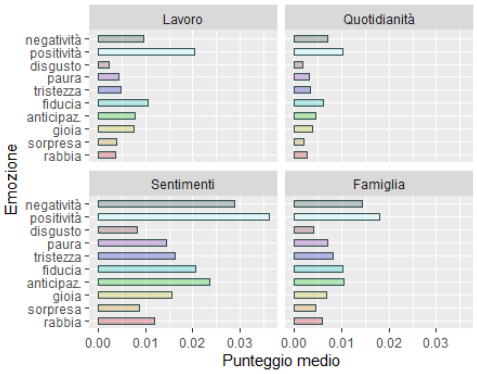


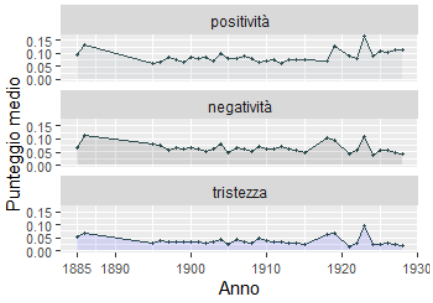
La sentiment analysis ci permette in primo luogo di evidenziare la prevalenza di positività, che potrebbe essere causata dalla presenza di termini con connotazioni positive che solitamente vengono poste in capo e in coda alle lettere.

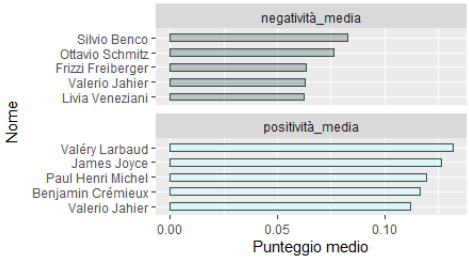


Come ci aspettiamo, i temi di Famiglia e Sentimenti sono positivi per l'autore. In particolare, quest'ultimo topic è quello caratterizzato dalle emozioni più forti. Si osserva anche che il profilo emotivo è simile tra le tematiche, differendo principalmente per intensità.

Si osservano picchi di negatività e positività molto simili, soprattutto negli anni di maggior successo dell'autore. La negatività di quel periodo sembra corrispondere a lutti personali, come confermerebbero anche i picchi nel grafico di tristezza. Le emozioni di positività sembrano interessare principalmente colleghi di lavoro, mentre la negatività principalmente i famigliari.







Riferimenti

[1] *Blei D.M., Ng A.Y., Jordan M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.2020. Lecture Notes in Computer Science, vol 12101. Springer, Cham. https://doi.org/10.1007/978-3-030-44094-7_10*

[2] *Saif Mohammad and Peter Turney. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California. See: http://saifmohammad.com/WebPages/lexicons.html)*