

VERIFICA TEXT MINING

Master Data Science A.A. 2018/19

Ceritello Graziano - De Giorgi Stefano Raffaele - La Mantia Roberto
Stentella Eleonora - Tomarchio Alessandro - Tomeo Valeria

Sommario

Introduzione e analisi dei file di output.....	3
Pre-processing	4
Corpus e metodo.....	5
Strategia 1 – Analisi dei primi 5 risultati più recenti per tutti i ristoranti.....	6
Barplot.....	6
Wordcloud cumulata strategia 1	7
Istogramma dei rating	7
Wordcloud per rating	8
Comparaison Cloud per promoter e detractor	9
Commonality cloud per promoter e detractor	10
Cluster Analysis	10
Strategia 2 - analisi degli ultimi 50 commenti dei 20 ristoranti con più recensioni del quartiere Testaccio. .	14
Barplot.....	14
Wordcloud cumulata strategia 2	15
Istogramma dei rating	15
Word cloud per rating	16
Comparison cloud per promoter e detractor	16
Commonality cloud per promoter e detractor	17
Cluster	17
Analysis in Iramuteq	18
Strategia 1	18
La struttura del corpus	18
Setting	18
Statistical summary	20
Wordcloud.....	21
Analisi delle specificità per ristoranti.....	22
Analisi delle similarità o Textometrical analysis	23
Cluster analysis.....	26
The Hierarchical Descending Classification.....	26
Subcorpus per metadata	31
Strategia 2	33
Wordcloud.....	34
Similarity analysis	36
Similarity analysis in Gephi	37

Analysis in Voyant strategia 1.....	39
Pre- processing su file di strategia 1	39
Wordcloud e Summary statistics Voyant.....	39
Analisi dei link all'interno del corpus	39
Il grafo delle collocazioni mostra un grafo a rete delle parole che con maggiore frequenza appaiono in prossimità di una parola data.....	39
Analisi dei sintagmi per le principali parole presenti nel testo	40
Analisi degli andamenti delle principali parole nel corpus	41
Analysis in Voyant strategia 2.....	41
Appendice 1.....	43
References.....	45

Introduzione e analisi dei file di output

L'obiettivo dell'analisi è quella di analizzare le recensioni dei ristoranti del quartiere Testaccio di Roma provenienti dal sito di TripAdvisor.

La raccolta dei dati è avvenuta mediante tecniche di Web Scraping utilizzando degli script in Python. Per l'attività di raccolta dati sono state utilizzate le librerie BeautifulSoup e Selenium, vedi "Appendice 1" per maggiori dettagli.

Per ogni ristorante sono state collezionate le recensioni con le relative stelle, il nickname dell'autore del commento e la data di pubblicazione.

Una volta ottenuti i dati sono state eseguite le seguenti strategie di analisi:

- **Strategia 1:** analisi degli ultimi 5 commenti postati per ciascun ristorante del quartiere Testaccio
- **Strategia 2:** analisi degli ultimi 50 commenti dei primi 20 ristoranti con il maggior numero di recensioni del quartiere Testaccio.
- **Strategia complementare alla 1 e 2:** analisi comparativa delle recensioni in base alle stelle assegnate (rating).

Infine, è stata eseguita un'analisi esplorativa dei due corpora prodotti dalle singole strategie.

Sono stati, poi, effettuati due tipi di approfondimenti con strumenti differenti. Una prima analisi è stata svolta utilizzando il software R e diversi pacchetti a disposizione per l'analisi testuale, facendo riferimento ai testi delle singole recensioni. Nella seconda analisi sono state accorpate le recensioni per ciascun ristorante e, nei paragrafi successivi, verranno descritti i risultati ottenuti utilizzando i software Iramuteq, Gephi e Voyant.

Di seguito si presentano i risultati ottenuti con il software R.

Per ogni strategia sono stati creati i seguenti output:

- **Barplot** con le frequenze delle 20 principali parole
- **"Wordcloud"**

Questa modalità consente di visualizzare le keyword presenti nelle recensioni con maggior rilevanza. Ogni parola è rappresentata con un carattere più o meno grande in relazione alla sua frequenza di occorrenza all'interno delle varie recensioni. Per ogni strategia è presente sia una **wordcloud complessiva di tutto il corpus**, costituito dall'insieme di tutte le recensioni, sia una **wordcloud per rating**, che riportano esclusivamente le parole chiave presenti per ogni singolo gruppo. I gruppi sono suddivisi considerando il rating, che assume i seguenti valori: 10 (1 stella), 20 (2 stelle), 30 (3 stelle), 40 (4 stelle), 50 (5 stelle), in base al gradimento assegnato dagli utenti nell'ambito di ciascuna recensione.

Insieme alle wordcloud sopra citate sono state elaborate anche:

- **comparison.cloud** compara due wordcloud di parole tra due subset stabiliti mediante la divisione in rating o in recensioni di *promoter* (con rating uguale a 50) e *detractor* (con rating da 10 a 30)
- **commonality.cloud** evidenzia le parole in comune tra due subset testuali suddivisi per rating o in recensioni di *promoter* (con rating uguale a 50) e *detractor* (con rating da 10 a 30)
- **cluster analysis** applicata direttamente alla matrice termini per documenti, genera dei gruppi di parole in base alla loro presenza all'interno delle recensioni.

Pre-processing

Per procedere con la classificazione automatica delle recensioni si utilizzerà la matrice Bag of Words.

La **Bag of words** consente di analizzare ed estrarre le *features* da documenti testuali e di inserirle in una matrice a 2 dimensioni.

La creazione della Bag of words per N recensioni prevede 2 fasi:

- Conversione della prima recensione in un vettore binario in cui si contrassegna l'occorrenza di una parola con 1 o 0 se non è presente.
- Trasformazione vettoriale reiterata per tutte le recensioni con l'obiettivo di ottenere una matrice "Termini x Frequenze" dove i termini sono le parole uniche che appaiono nell'intera raccolta di recensioni e le frequenze sono rappresentate dalle occorrenze di ogni termine per ogni recensione.

Questo processo da un lato produce la matrice Termini x Frequenze, dall'altro ci fa perdere ogni informazione relativa all'ordine, struttura e relazione semantica delle parole. Il focus del Modello BoW infatti è il conteggio del numero di occorrenze di ogni parola e non la posizione all'interno della frase.

Per ottimizzare la gestione delle risorse computazionali durante il processo si cerca di perseguire una logica di diminuzione delle dimensioni del vocabolario totale delle parole adottando tecniche come:

- Rimozione della punteggiatura e dei caratteri numerici.
- Rimozione delle stopwords (nel nostro caso abbiamo utilizzato una lista precompilata presente online e adatta per la lingua italiana).
- Processo di stemming.

Rimuovere la punteggiatura e i caratteri numerici aiuta a evitare una eccessiva frammentazione in token del testo. Questo consente, in una fase di pre-processing, di rimuovere una serie di dati che in genere non forniscono informazioni utili ai fini della ricerca.

Con lo stesso scopo vengono gestite le stopwords, ossia quelle parole che sono prive di contenuto informativo in relazione al modello di studio.

Infine, viene eseguito il processo di stemming, ossia la riduzione delle forme flesse e delle parole derivate ad una forma base chiamata "radice" di parola. Quest'attività è utile per ridurre le dimensioni del vocabolario totale e per standardizzare e ridurre le varianti delle parole da analizzare. Nelle operazioni di stemming si può scegliere il grado di riduzione a radice, prestando attenzione che all'aumentare del grado di stem si rischia di incappare nell'over-stemming, ossia raggruppare parole con significato diverso con la conseguente perdita di specificità dell'analisi. Al contrario, diminuendo il grado di stem si rischia di trovarsi nell'under-stemming con la conseguente perdita di generalità del testo.

In genere, le stopwords vengono eliminate prima di procedere con lo stemming in modo da minimizzare il numero di termini da ridurre alla radice.

Per la fase di pre-processing nella nostra analisi abbiamo utilizzato il pacchetto di R "tm" che al suo interno contiene un vocabolario in Italiano e consente di gestire diverse fasi del pre-processing.

Le operazioni sopra descritte non sempre, però, hanno avuto un esito ottimale. In particolare, nella fase di rimozione dei caratteri anomali e simboli e nella fase di stemming si sono evidenziate alcune casistiche che il software non è riuscito a trattare opportunamente. Ad esempio, applicando in "tm_map" la funzione che rimuove i caratteri anomali in alcune wordcloud si evidenzia la presenza di parole troncate delle lettere accentate come "à" o "é", o ancora se si utilizza la funzione di stemming ("stemDocument") vengono tagliate tutte le parole per ricondurle alla radice comune per cui si uniformano i casi in cui una stessa parola è declinata al maschile/femminile o al singolare/plurale (ad es. "ottimo", "ottima", "ottimi") ma le wordcloud appaiono meno leggibili. In alternativa, si poteva utilizzare il pacchetto "udpipe" che restituisce risultati migliori. Si è deciso, però, di utilizzare il software Iramuteq per descrivere i risultati ottenuti con un pre-processing ottimale.

Corpus e metodo

Utilizzando R sono stati creati due corpora delle recensioni. Il primo passo è stato quello di verificare la componente lessicografica dei messaggi delle due strategie.

Di seguito si presentano i risultati ottenuti utilizzando, nella fase di pre-processing, le operazioni di pulizia del testo sopra descritte ma prima senza applicare la funzione di stemming (che restituisce delle wordcloud più chiare) e poi applicandola. Abbiamo optato per questa soluzione anche per confrontare approcci differenti. L'analisi con il pre-processing migliore è stata fatta con Iramuteq.

Sono state raccolte in totale 1.724 recensioni così suddivise: 724 per la strategia 1 e 1.000 per la strategia 2.

Si è proceduto in prima istanza con l'analisi lessicale ricavando informazioni utili alla successiva analisi testuale volta a localizzare unità di testo di rilievo per gli obiettivi del presente studio.

Il primo corpus analizzato, senza utilizzare lo stemming, è composto da 724 testi, 20009 occorrenze (tokens), 6177 forme (types), 3797 hapax, ossia il numero di forme grafiche che si presentano una sola volta, che rappresentano il 61.47% di tutte le forme presenti. La ricchezza lessicale del corpus è discreta ($V/N*100 = 30.87\%$) per cui, a fronte di un testo non molto ampio, si riscontra un vocabolario non proprio ridotto. Confrontando con i risultati ottenuti applicando la funzione di stemming, appare evidente come si riducano sensibilmente il numero delle forme grafiche diverse (4129 types) e, quindi, della complessità lessicale del testo (20.64%).

Lexicometric measures	Values (no stem)	Values (stem)
N. of messages	724.00	724.00
Total word count (tokens, N)	20009.00	20009.00
N. of different graphic forms (types, V)	6177.00	4129.00
Complexity Factor (lexical density, $V/N*100$)	30.87	20.64
N. of Hapax (V_1)	3797.00	2342.00
Hapax Percentage ($H= V_1/V*100$)	61.47	56.72
General Average Frequency (N/V)	3.24	4.85
Guiraud Index (V/\sqrt{N})	43.67	29.19

Il secondo corpus analizzato è composto da 1000 testi, 27775 occorrenze, 7018 forme, 4349 hapax pari al 61.97% delle forme totali. La ricchezza lessicale del corpus è leggermente inferiore rispetto al primo corpus ($V/N*100 = 25.27\%$). Anche in questo caso, applicando la funzione di stemming il numero di forme differenti e la complessità lessicale si riducono drasticamente (4546 e 16.37% rispettivamente).

Lexicometric measures	Values (no stem)	Values (stem)
N. of messages	1000.00	1000.00
Total word count (tokens, N)	27775.00	27775.00
N. of different graphic forms (types, V)	7018.00	4546.00
Complexity Factor (lexical density, $V/N*100$)	25.27	16.37
N. of Hapax (V_1)	4349.00	2608.00
Hapax Percentage ($H= V_1/V*100$)	61.97	57.37
General Average Frequency (N/V)	3.96	6.11
Guiraud Index (V/\sqrt{N})	42.11	27.28

Strategia 1 – Analisi dei primi 5 risultati più recenti per tutti i ristoranti
 Barplot

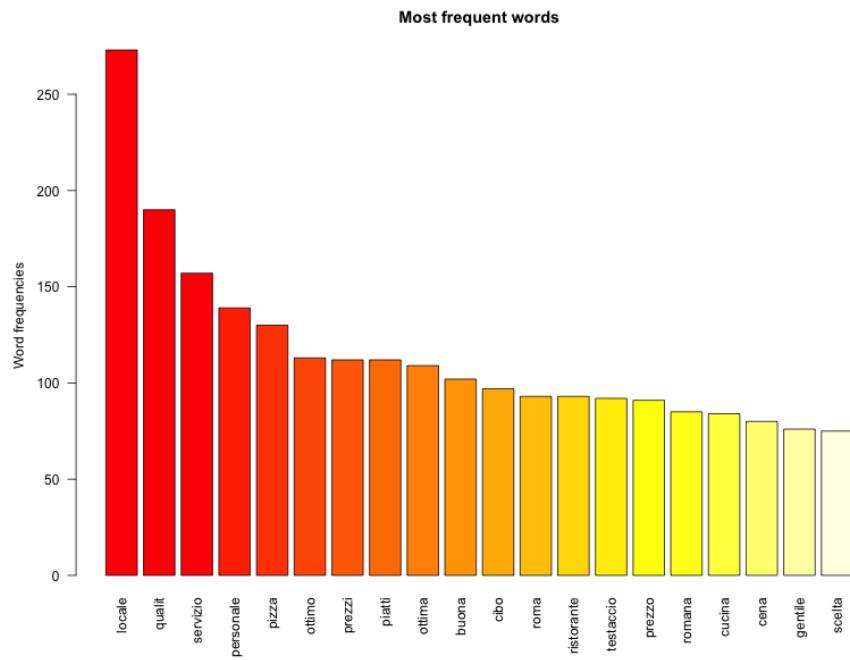


Figura: Parole più frequenti presenti all'interno del corpus della strategia 1 - senza stemming

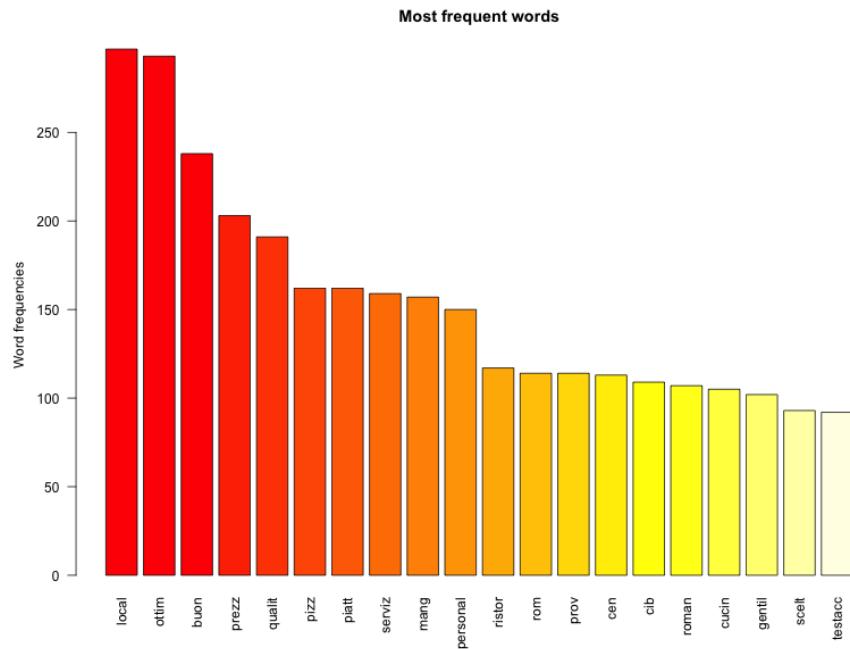


Figura: Parole più frequenti presenti all'interno del corpus della strategia 1 - con stemming

Wordcloud cumulata strategia 1

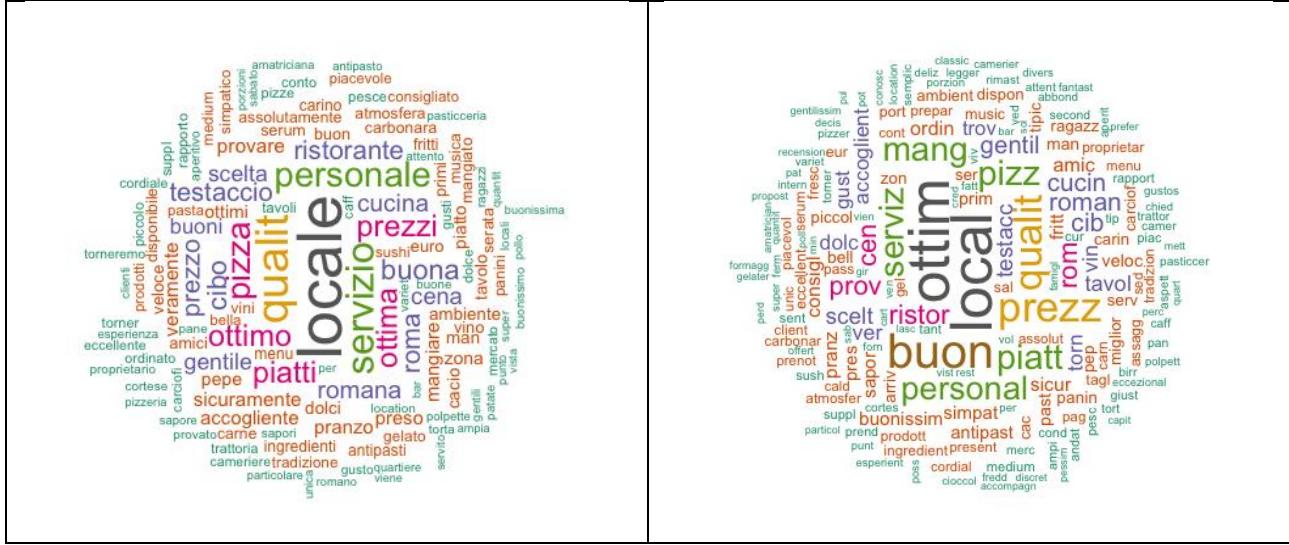


Figura: Parole più frequenti all'interno del corpus costituito da tutte le recensioni - senza stemming e con stemming

L'analisi della distribuzione di frequenze delle parole che compaiono nelle recensioni, per la prima strategia, evidenzia come sia la parola "locale" quella maggiormente utilizzata, essendo, insieme al sinonimo "ristorante" che compare poco più in fondo, il termine che contestualizza l'oggetto principale della recensione stessa. Seguono, poi, parole che identificano le dimensioni prese in esame nella critica, quali, ad esempio, "qualità", "servizio", "personale", "pizza", e aggettivi che definiscono il giudizio di merito sui singoli temi, come "ottimo", "buono", "gentile". Appaiono, infine, anche nomi che definiscono il luogo in cui si trovano i ristoranti, come "Roma" o "testaccio".

È interessante vedere come, confrontando il risultato ottenuto applicando lo stemming (con tutti i limiti già esposti) gli aggettivi “ottimo” e “buono” (che appaiono troncati) aumentino notevolmente in termini di numero di occorrenze, così come la parola “prezzo”.

Iistogramma dei rating

Attraverso l'istogramma dei rating è possibile evidenziare come si distribuiscono le recensioni in base al punteggio attribuito dagli utenti. Si nota che nel quartiere romano preso in esame i ristoranti tendono ad avere recensioni con rating molto alto. I punteggi 5 e 4 (che nel codice R sono espressi in 50 e 40) sono i più frequenti e questo evidenzia un alto grado di soddisfazione da parte degli utenti sia per quanto concerne la prima analisi, che prende in esame i primi 5 commenti per ogni ristorante della zona, sia per la seconda, che considera solo i primi 20 ristoranti con il maggior numero di recensioni.

La valutazione emersa è coerente con l'area in quanto la zona del quartiere Testaccio è notoriamente riconosciuta per avere alcuni tra i migliori ristoranti e trattorie di Roma.

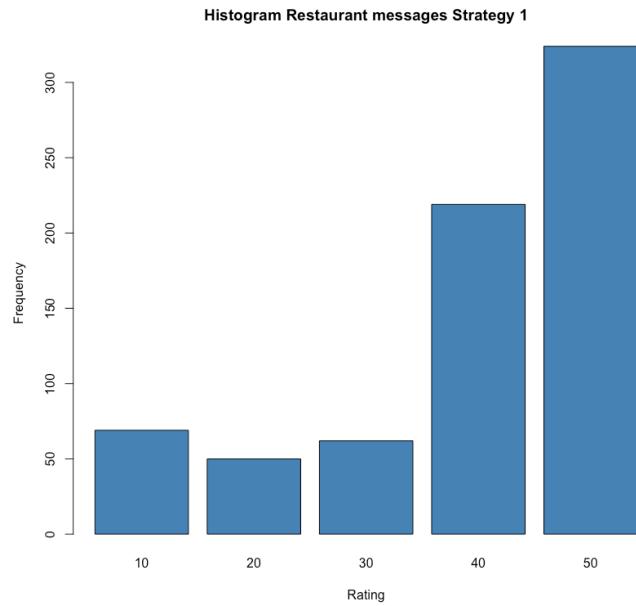


Figura: Distribuzione di frequenze delle recensioni in base al rating

Wordcloud per rating

Le wordcloud per rating sono le nuvole di parole chiave che evidenziano le parole con maggiori occorrenze per ogni singolo subset creato.



Comparaison Cloud per promoter e detractor

La presente analisi fonda le sue radici sull'approccio metodologico dell'NPS Scoring inventato dalla Bustom Consulting Group. I messaggi presenti all'interno del corpus sono stati suddivisi tra quelli con rating basso e quelli con rating alto come segue:

```
##### Genera i dataframe per i rating più bassi e per quello più alto
df_s1_low <- subset.data.frame(msg_s1, no_rating %in% c(10, 20, 30) )
df_s1_high <- subset.data.frame(msg_s1, no_rating %in% c(50) )
```

E' stata creata una funzione per richiamare quanto sopra evidenziato che confronta due liste di messaggi suddivisi tra quelli con un rating basso e quelli

```
sentiment_comparison<-function(msg_low, msg_high, lab1, lab2){
  # Prepara i due dataset su due vettori collassando i documenti
  low = paste(msg_low, collapse=" ")
  high = paste(msg_high, collapse=" ")
  low <- removeWords(low, stop_vec)
  high <- removeWords(high, stop_vec)
  ...
}
```

s1_low



s1_high

Il confronto tra i risultati ottenuti per i promoter (High = 50) e i detractor (10 -30) evidenzia come i temi che maggiormente contraddistinguono la prima tipologia di utenti, per altro più conspicua della seconda, sono legati: alla qualità dei prodotti e degli ingredienti; all'accoglienza e alla gentilezza del personale; menzionato anche l'ambiente. Si noti l'uso di aggettivi come "ottimo" ed eccellente". I detractor usano di più l'aggettivo "buona", presumibilmente legato alla pizza, "scadente", "poco" e "quantità" che, vista la presenza anche della parola "costo" lascia presumere una scontentezza legata al rapporto qualità/prezzo.

Commonality cloud per promoter e detractor

L'analisi che segue mette in evidenza le parole condivise tra le recensioni che hanno un rating che va da 10 a 30 (da 1 stella a 3 stelle) verso quelle con valore pari a 50 (5 stelle)



Tra le parole comuni torna la "qualità", il "servizio", il "cibo" ovviamente con diversa valenza rispetto al giudizio espresso come per altro già visto dalla wordcloud precedente.

Cluster Analysis

Di seguito si presentano i risultati della **cluster analysis** effettuata provando diversi metodi di formazione dei gruppi (**gerarchico agglomerativo con i criteri del legame completo, singolo, ward e il k-means**), e diverse misure di prossimità (**distanza euclidea e coseno**). Per ciascun dendogramma ottenuto con i metodi gerarchici è stato calcolato il **coefficiente di correlazione cofenetico**.

Si è scelto, anche in questo caso, di presentare i risultati ottenuti senza lo stemming perché di più facile lettura e non molto diversi da quelli ricavati effettuando lo stemming nel pre-processing.

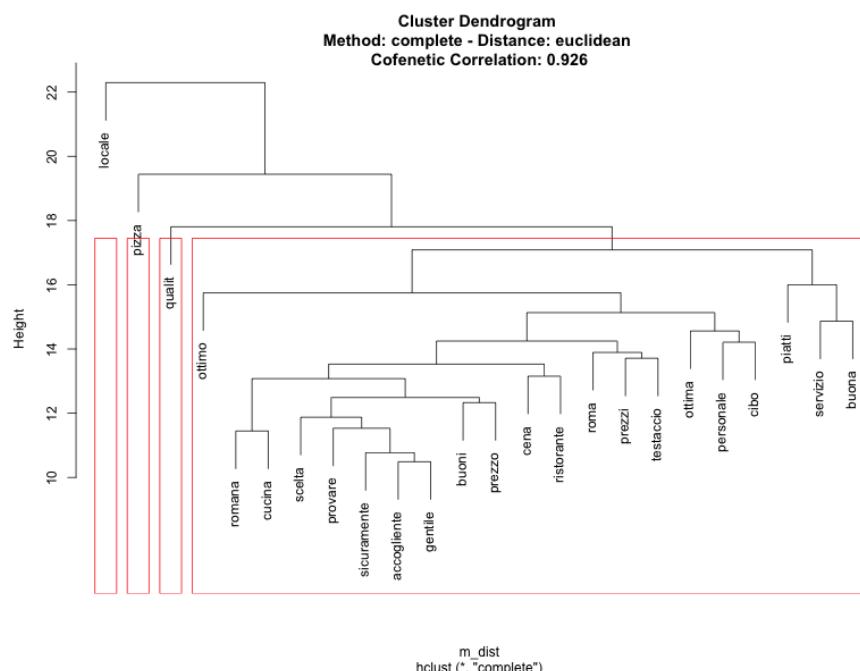


Figura 1: Cluster dendrogram – Method: complete – Distance: euclidean

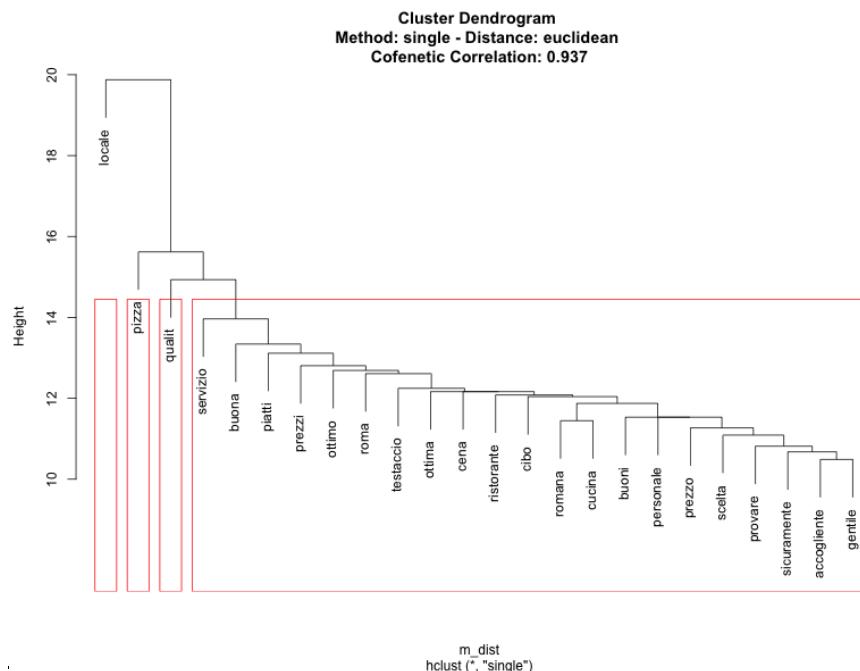


Figura 2: Cluster dendrogram – Method: single – Distance: euclidean

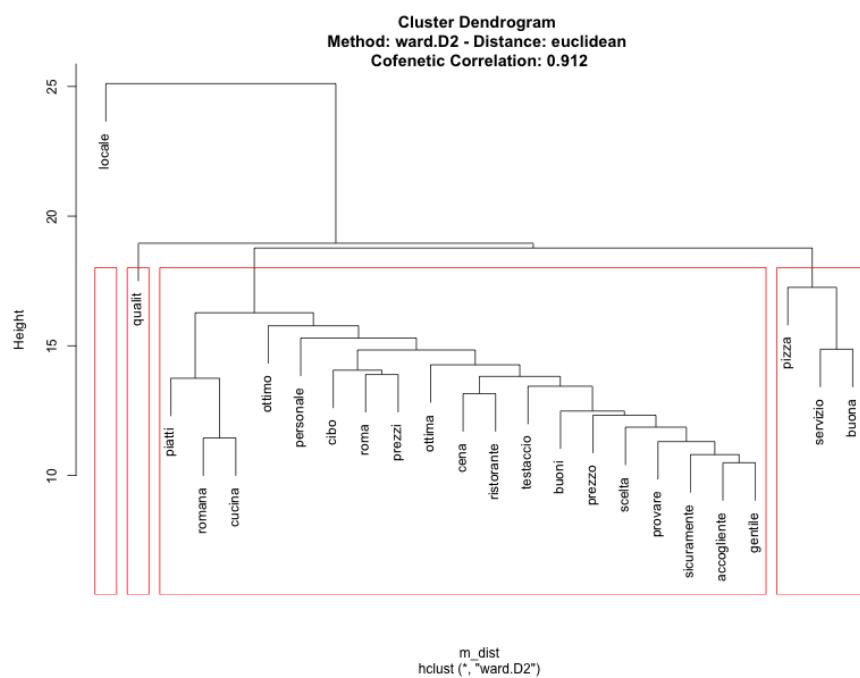


Figura 3: Cluster dendrogram – Method: ward.D2 – Distance: euclidean

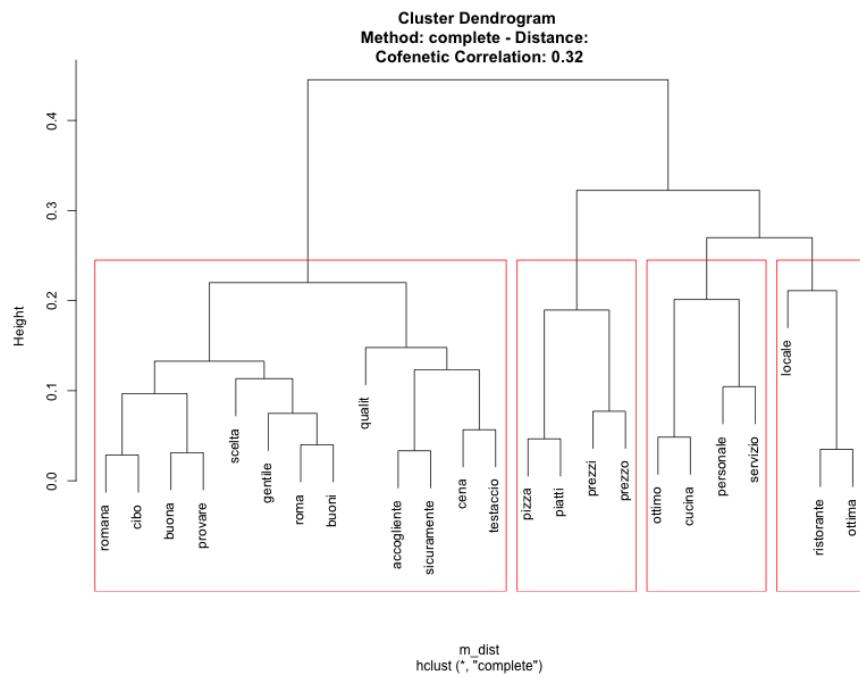


Figura 4: Cluster dendrogram – Method: complete – Distance: cosine

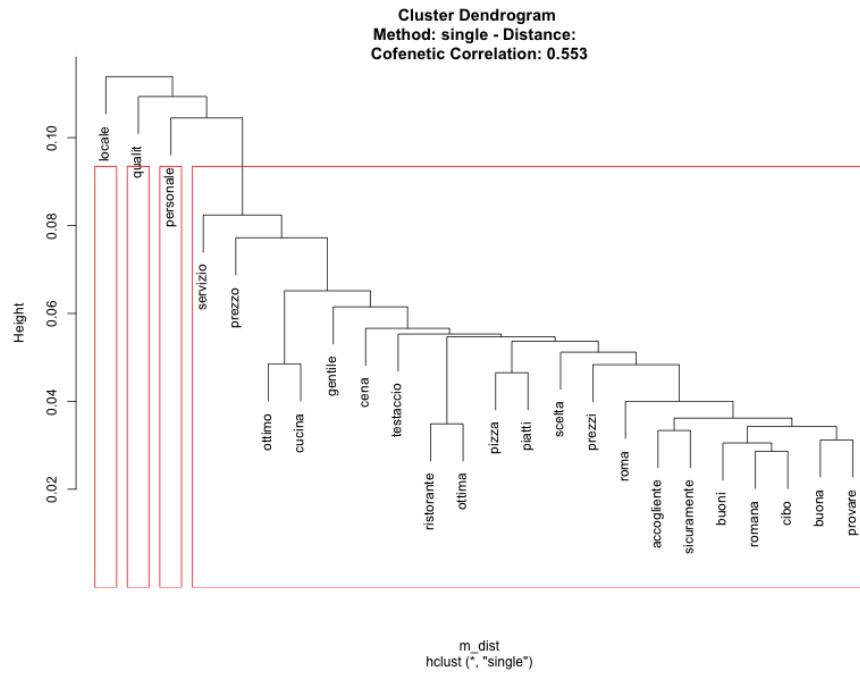


Figura 5: Cluster dendrogram – Method: single – Distance: cosine

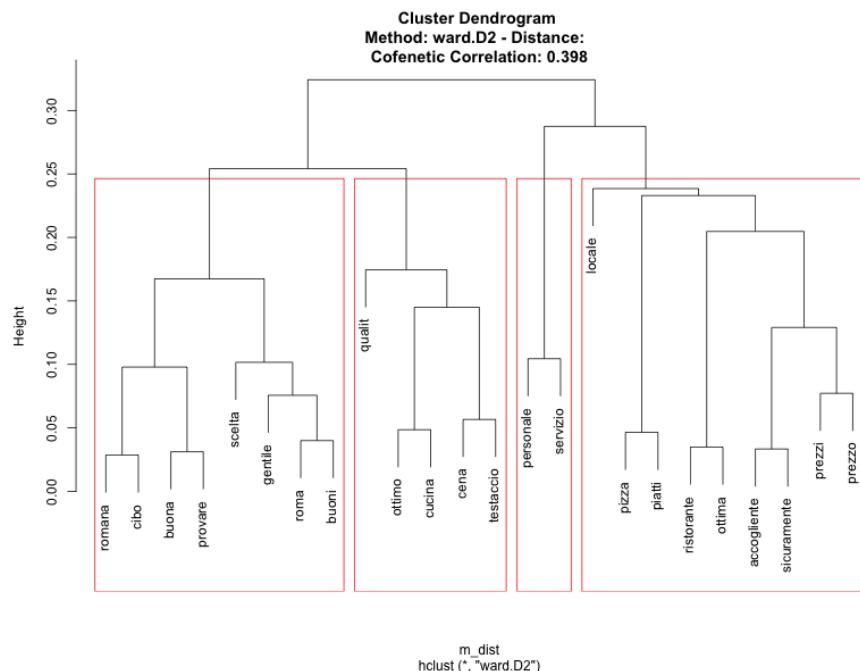


Figura 6: Cluster dendrogram – Method: ward.D2 – Distance: cosine

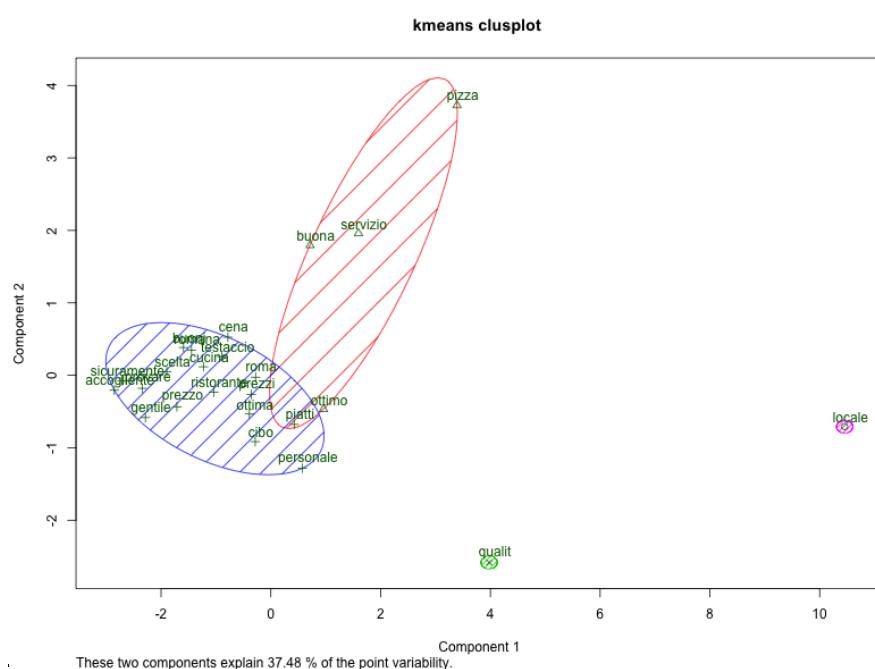


Figura 7: Kmeans cluster analysis

Le metodologie applicate forniscono risultati molto diversi tra loro. I cluster individuati utilizzando la distanza euclidea, se presentano un valore del coefficiente di correlazione cofenetico piuttosto alto, appaiono di difficile interpretazione. Al contrario, i cluster ottenuti calcolando la distanza del coseno, seppur con un coefficiente cofenetico più basso, identificano cluster maggiormente interpretabili. Ad esempio, il metodo del legame completo con distanza del coseno restituisce quattro gruppi: 1. le recensioni esprimono un giudizio ottimo sul locale nel complesso ; 2. si raggruppano giudizi positivi incentrati sul servizio, sul personale e la cucina; 3. appaiono recensioni che si focalizzano sul prezzo dei piatti, in particolare della pizza; 4. utenti che esprimono valutazioni su più aspetti che vanno dalla qualità, all'accoglienza, al cibo.

In generale, comunque, l'analisi svolta sulle singole recensioni presenta **forti limiti dovuti al fatto che la matrice TerminixDocumenti risulta molto sparsa** (con numerose frequenze nulle in corrispondenza di un gran numero di recensioni). Pertanto, si è ritenuto opportuno effettuare un'ulteriore analisi raggruppando le recensioni per i ristoranti. I dati sono stati elaborati utilizzando il software Iramuteq e i risultati verranno discussi nei paragrafi successivi.

Strategia 2 - analisi degli ultimi 50 commenti dei 20 ristoranti con più recensioni del quartiere Testaccio.

Barplot

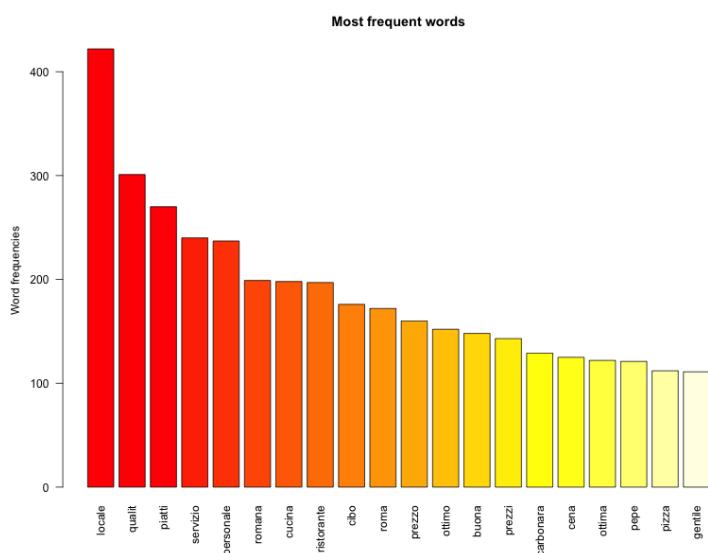


Figura: Parole più frequenti presenti all'interno del corpus della strategia 2 - senza stemming

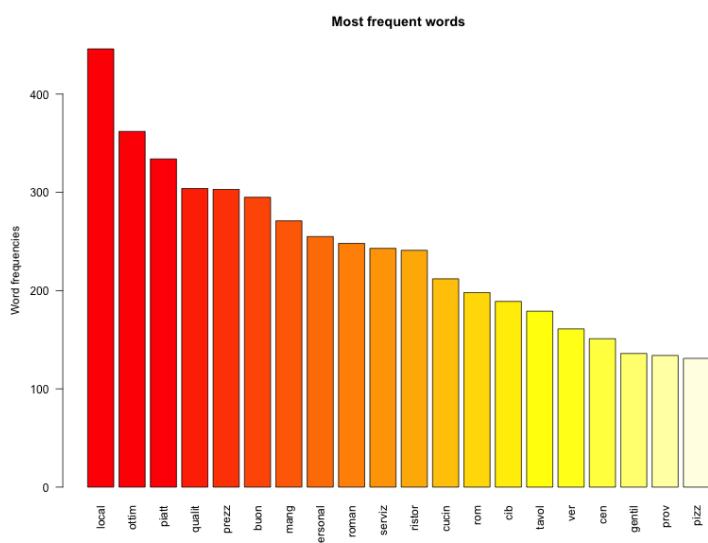
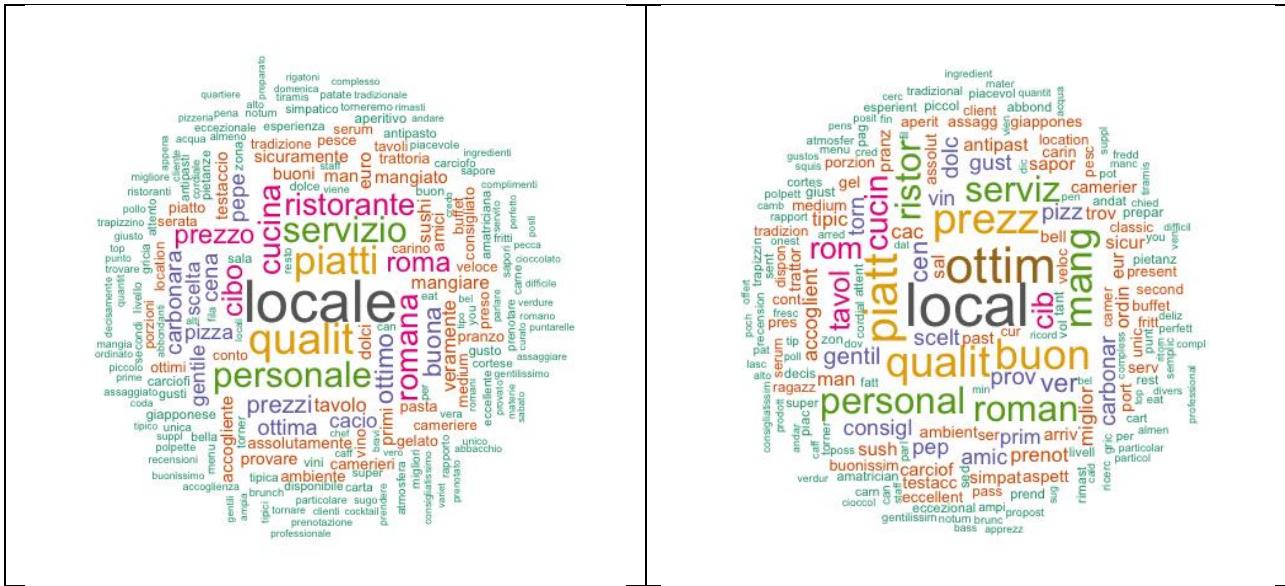


Figura: Parole più frequenti presenti all'interno del corpus della strategia 2 - con stemming

Wordcloud cumulata strategia 2



Dal confronto delle wordcloud prodotte nella nostra analisi sia per la strategia 1 che per la strategia 2 la **parola più ricorrente è "locale"**.

Possiamo notare che, confrontando la prima e la seconda strategia, nella nuvola di parole ci sono termini che variano la loro occorrenza, come ad esempio "piatti" che appare più spesso nella seconda analisi che nella prima. La parola "testaccio", che rispecchia una caratteristica comune a tutti i ristoranti analizzati, ha una ricorrenza decisamente maggiore nella prima analisi mentre nella seconda si riduce di molto in confronto alle altre parole utilizzate per le recensioni.

Inoltre, si può notare che nelle wordcloud della seconda strategia troviamo **un maggior numero di parole**, questo poiché analizzando un maggior numero di recensioni il vocabolario analizzato è di conseguenza maggiore.

Iistogramma dei rating

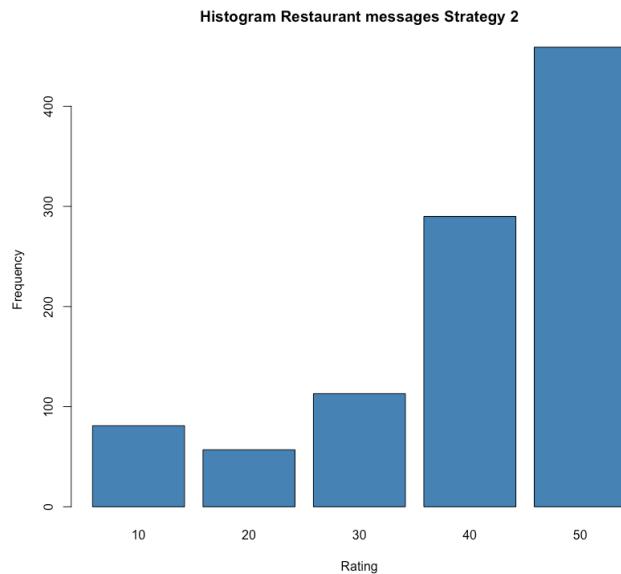


Figura: Distribuzione di frequenze delle recensioni in base al rating

Word cloud per rating



Figura: Wordcloud cumulate dei subset con rating 10 e 50

Comparison cloud per promoter e detractor

s2_low



s2_high

Commonality cloud per promoter e detractor



Cluster

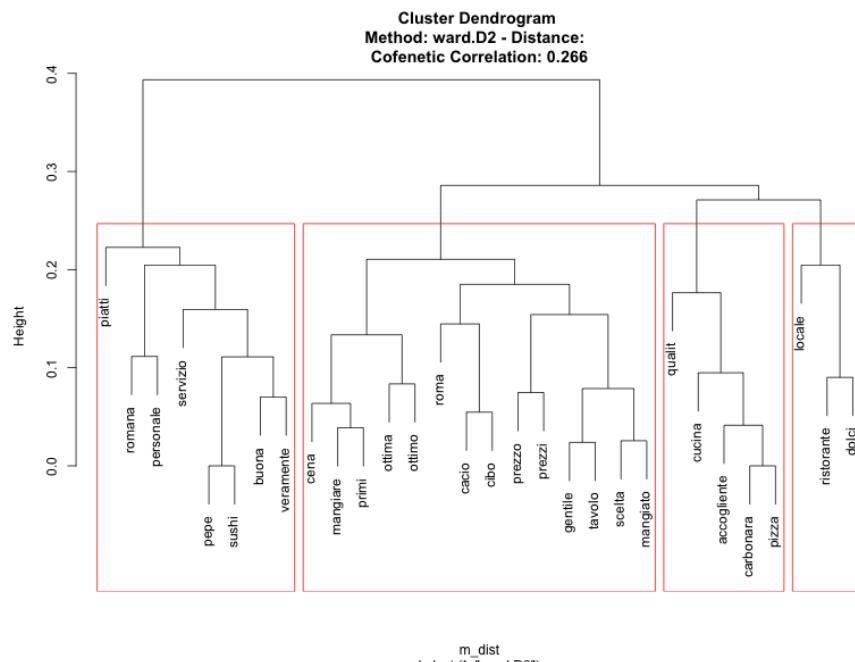


Figura: Cluster dendrogram – Method: complete – Distance: cosine

Analysis in Iramuteq

Strategia 1

La struttura del corpus

I file di input per Iramuteq devono essere in formato testo (.txt) e osservare specifiche regole di formattazione.

L'unità base si chiama "testo". Un testo può rappresentare un'intervista, un articolo, un libro o qualsiasi altro tipo di documento. Un corpus può contenere uno o più testi. I testi sono introdotti da quattro asterischi (*****) seguite da una serie di variabili stellate separate da uno spazio. È possibile inserire le variabili speciali nel testo introducendo l'inizio della riga con un trattino seguito da una stella (- *). Questo è noto come "temi". La riga deve contenere solo questa variabile.

Per il nostro formato corpus, abbiamo scelto un formato con tre variabili rappresentative.

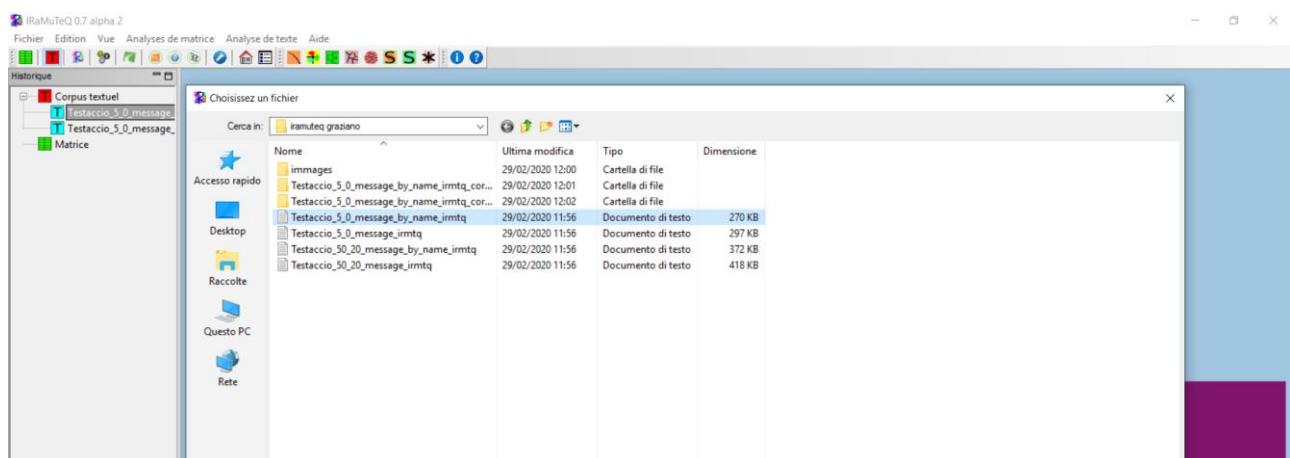
L'analisi è stata svolta accorpando le recensioni per ciascun ristorante, al quale poi è stato attribuito il rating medio delle recensioni.

**** *NAME_1.CasaManco *RATING_5

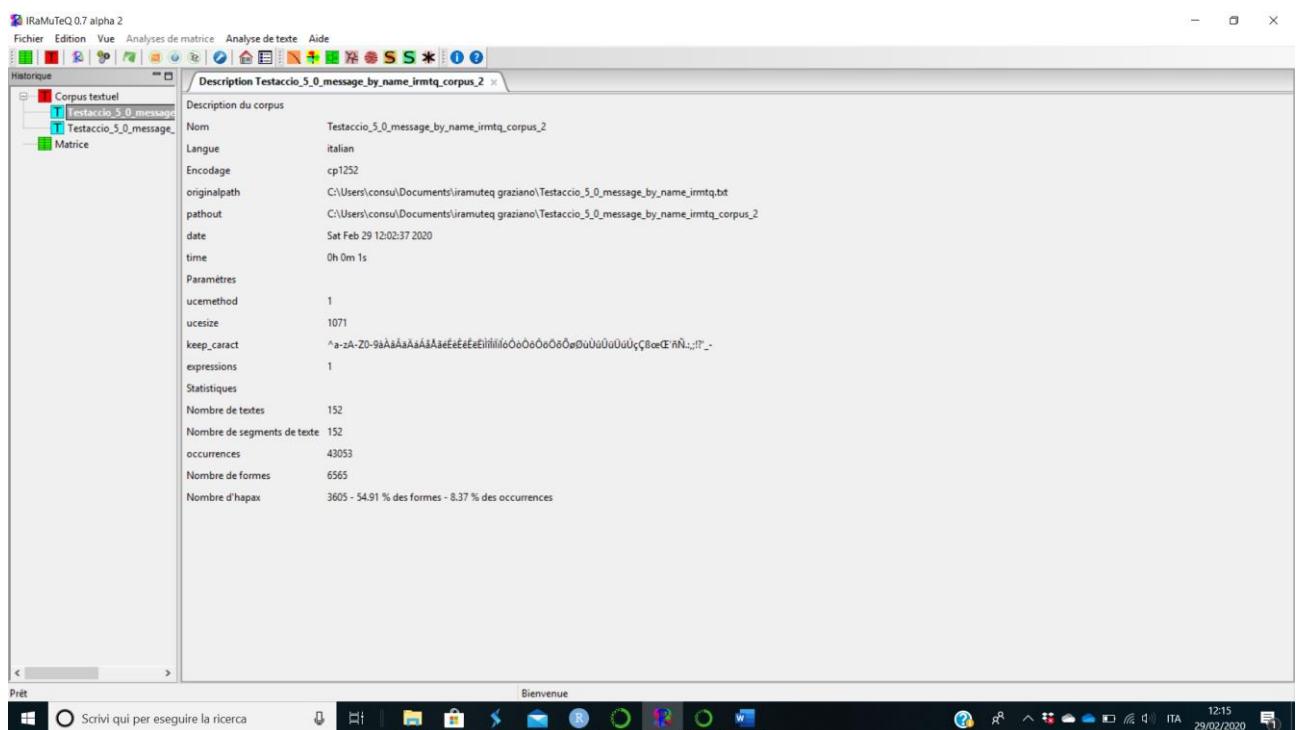
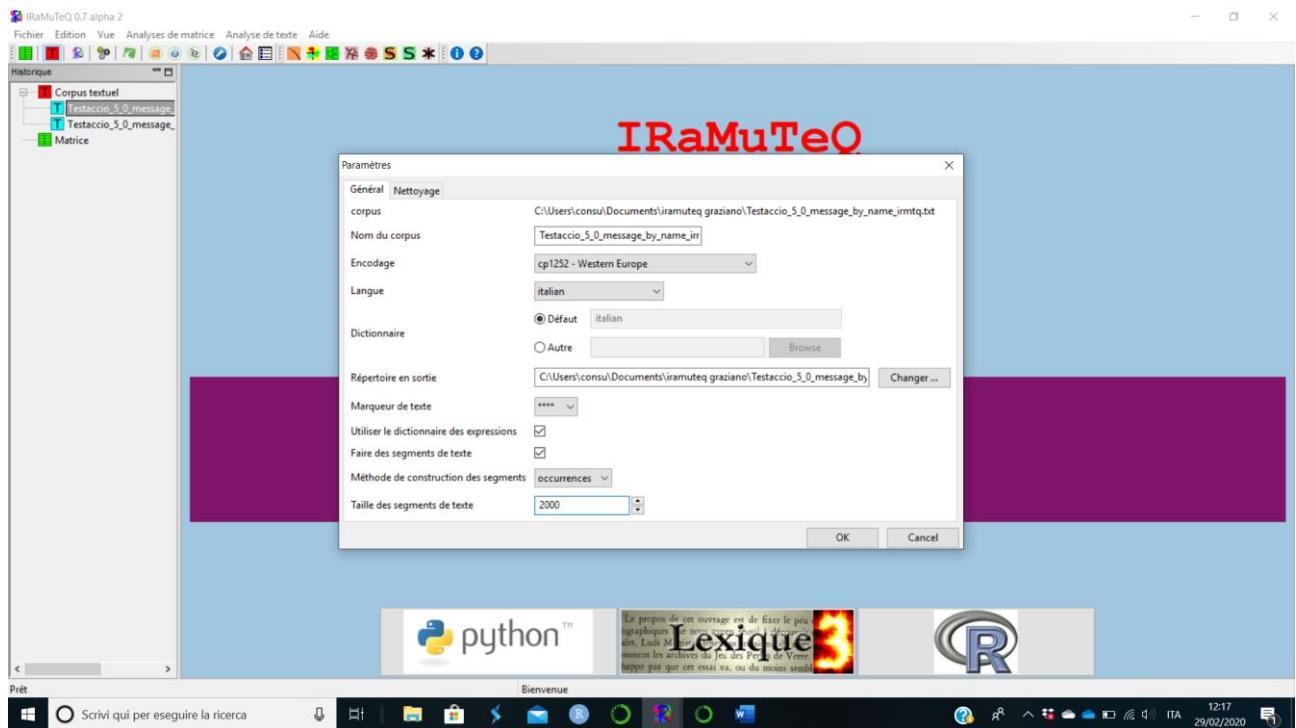
Alla fine capisci perchè un architetto lascia la professione per sfornare prodotti Perchè aveva un progetto ben preciso E in grado di realizzare impasti pazzeschi con forme strepitose che si baciano perfettamente con il nostro palato e si incastrano sapientemente nel nostro stomaco Fantastica fra tutti i premi affissi alle pareti manca il mio MIGLIOR ARCHIFETTA del mondo Ce ne sono tante diverse ma sempre con ottimi ingredienti Non potete sbagliare Da provare in particolare quella con la scarola uvetta e alici Trovata per caso dentro il mercato del testaccio di domenica ha dato un senso ad un viaggio milano roma che sarebbe stato altrimenti per i miei affari deludente Val la pena di venire da Milano apposta per mangiarsi queste pizze romane

Setting

Abbiamo caricato in Iramuteq un corpus costituito dai primi 5 messaggi di tutti i ristoranti del quartiere di Testaccio suddivisi per il nome del ristorante.



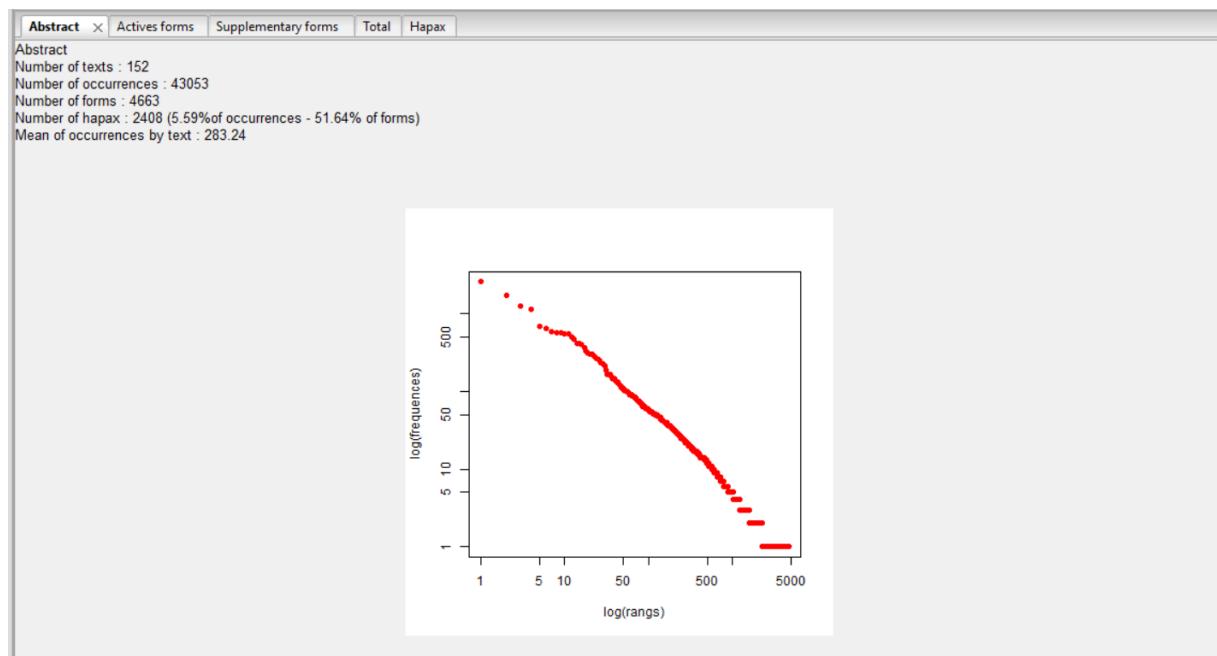
Nello step successivo sono stati aumentati i segmenti a 2000 per far combaciare il numero dei ristoranti presenti nel corpus con il numero di segmenti da prendere in esame.



Statistical summary

Dalla statistica di base abbiamo potuto notare che il corpus contiene 152 messaggi

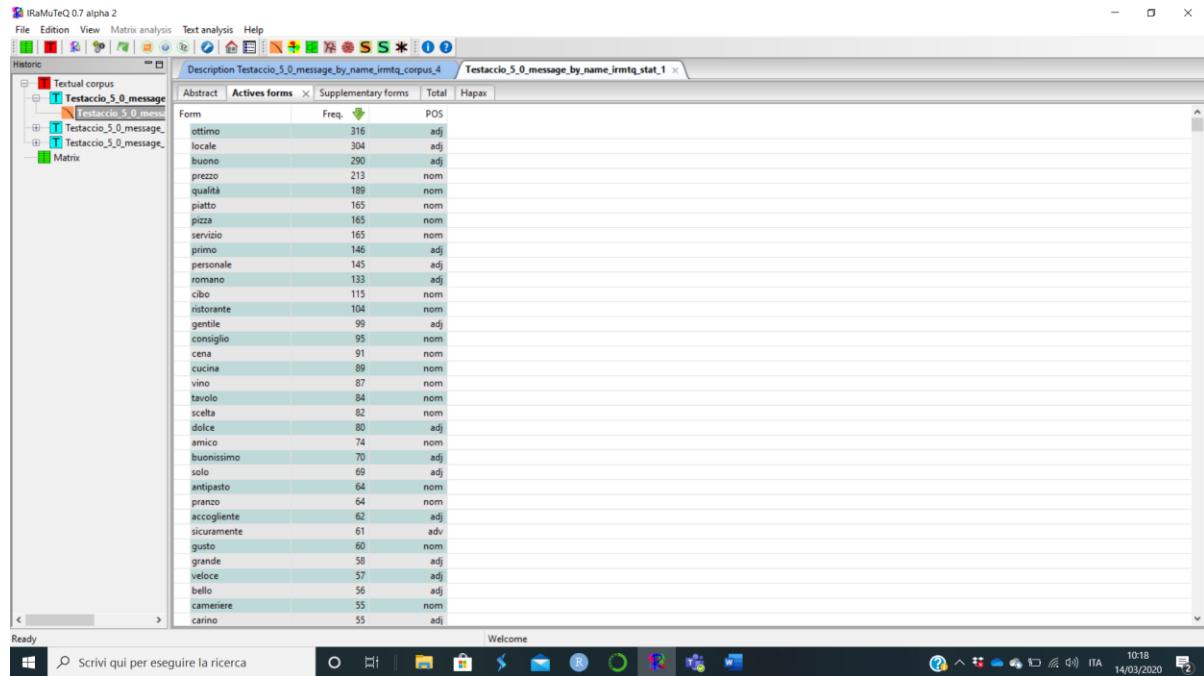
Gli hapax, ossia le parole che hanno una frequenza pari a 1 nel corpus, sono costituite da parole ordinarie ma anche da tipiche forme attaccate che si incontrano nei testi web e parole con errori di battitura o con “_”



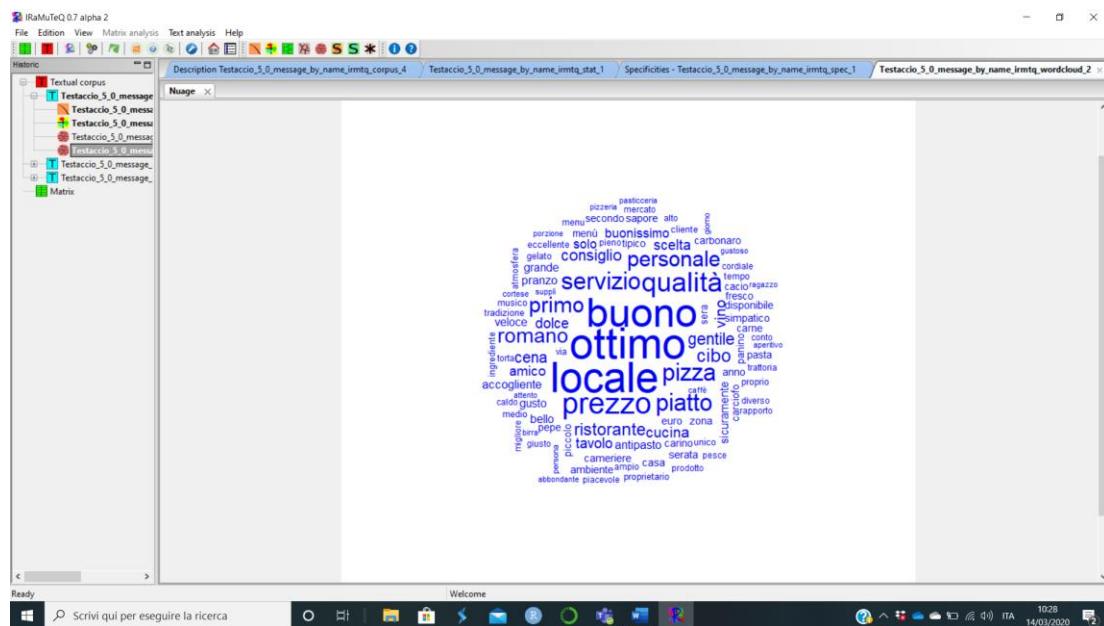
Description Testaccio_5_0_message_by_name_irmtq_corpus_4			Testaccio
Abstract	Actives forms	Supplementary forms	Total
Form	Freq.	POS	Hapax
zushi	1	nr	
zumare	1	ver	
zucchine_come	1	nr	
zuccheroso	1	adj	
zoé	1	nr	
zonz	1	nr	
zingarata	1	nr	
zighini	1	nr	
zializio	1	nom	
zereshk	1	nr	
zabaionee	1	nr	
zabaione_che	1	nr	
yakisoba	1	nr	
xvolevano	1	nr	
xvi	1	num	
wurstel	1	nr	
wok	1	nr	
willy	1	nr	
wifi	1	nr	
wi	1	nr	
whiskey	1	nr	
whatsapp	1	nr	
week_end	1	nr	
web	1	nr	
wasabi	1	nr	
vuole_poché	1	nr	
vomito	1	nom	
vomitare	1	ver	
volteggiare	1	ver	
volpetti	1	nr	
volare	1	ver	
voce	1	nom	

Description Testaccio_5_0_message_by_name_irmtq_corpus_4			Te
Abstract	Actives forms	Supplementary forms	Total
Form	Freq.	POS	Hapax
0	1	num	
09	1	num	
Olio	1	nr	
10euro	1	nr	
11	1	num	
150euro	1	nr	
157	1	num	
179	1	num	
200	1	num	
2020	1	num	
24h	1	nr	
26	1	num	
27	1	num	
2analcolici	1	nr	
2fratelli	1	nr	
3sicuramente	1	nr	
41	1	num	
43	1	num	
45	1	num	
51	1	num	
56	1	num	
57	1	num	
5l	1	nr	
72	1	num	
72h	1	nr	
76	1	num	
a_causa	1	sw	
a_dispetto	1	sw	
a_partire_dai	1	nr	
a_partire_dalla	1	nr	
abbaiare	1	ver	
abbaio	1	nom	

Tra le forme grammaticali più presenti nel testo troviamo gli aggettivi, tra i quali campeggia l'aggettivo “ottimo” segno evidente che, tendenzialmente, i ristoranti presi in esame hanno una buona considerazione. Si noti, altresì, come la forma nominale “locale” venga erroneamente identificata come un aggettivo pur essendo un nome. L’analisi delle parole evidenzia una stretta correlazione con il rating medio per locale come già si evinceva dagli istogrammi di output del codice in R.



Wordcloud



La wordcloud realizzata in Iramuteq è del tutto conforme con quella generata attraverso il software R studio ed evidenzia, quanto già preso in esame mediante le statistiche di base: "ottimo", "locale", "buono" e "prezzo" sono le parole con il più alto numero di frequenze presenti nel testo.

Analisi delle specificità per ristoranti

L'analisi sotto proposta ha lo scopo di valutare le occorrenze maggiormente frequenti per ciascun ristorante presente nella lista. La tabella sottostante riporta la schermata del Software con le indicazioni.

Il limite di questo approccio è relativo all'analisi lessico-metrica, ossia il numero di recensioni prese per ristorante sono troppo poche per fornire una overview corretta e precisa.

Nella strategia 1 il numero delle occorrenze è molto basso per ogni singola forma in quanto abbiamo a disposizione solo i primi 5 messaggi per ogni locale. Questo si riflette nelle statistiche di base.

Nelle forms frequencies in valori assoluti il fatto è molto evidente e suggerisce che tale tecnica sia molto più adatta se applicata sui rating e non per ristoranti oppure su una strategia diversa come la n.2 dove si riportano i primi 50 commenti dei primi 20 ristoranti.

iRaMuTeQ 0.7 alpha 2

File Edition View Matrix analysis Text analysis Help

Historic

Textual corpus

- SubTestaccio_5_0_message_by_name_irmtq_corpus
- Testaccio_5_0_message_by_name_irmtq_st
 - Testaccio_5_0_message_by_name_irmtq_st
- Matrix

Specificities - Testaccio_5_0_message_by_name_irmtq_spec_1 Clustering - Testaccio_5_0_message_by_name_irmtq_corpus_4 Description SubTestaccio_5_0_message_by_name_irmtq_corpus_4

Forms Banal forms POS Forms frequencies POS frequencies Forms relative frequencies POS relative frequencies CA

forms	*NAME_1.CasaManco	*NAME_10.TrattoriaPennestri	*NAME_100.CharroCafe	*NAME_101.CimagliaSilvana	*NAME_102.Yoshi
pizza	4	0	1	2	0
personale	3	0	0	0	0
ingrediente	2	0	0	0	0
ottimo	2	6	2	2	2
impasto	2	0	0	0	0
buono	2	0	3	1	2
pena	2	0	0	0	0
romano	2	1	0	2	0
migliore	2	2	0	0	0
sicuramente	1	1	0	0	1
palato	1	0	0	0	1
qualità	1	2	0	1	0
gentile	1	2	0	1	1
prezzo	1	3	0	3	2
onesto	1	0	0	0	0
mercato	1	0	0	0	0
perfetto	1	0	1	0	1
diverso	1	0	1	1	1
disponibile	1	1	0	0	0
primo	1	0	1	2	0
vivamente	1	0	0	0	0
simpatico	1	0	0	0	0
domenica	1	0	0	0	0
super	1	0	0	0	0
sapore	1	1	0	0	1
fantastico	1	0	3	0	1
consiglio	1	0	1	3	0
condimento	1	0	0	0	0
altro	1	0	1	0	1
gusto	1	0	0	0	1
lungo	1	1	0	0	0
prodotto	1	0	0	0	0
eccezionale	0	0	0	0	0

Ready Welcome

Scrivi qui per eseguire la ricerca

18:02 15/03/2020 ITA

Analisi delle similarità o Textometrical analysis

Per la realizzazione della seguente analisi **abbiamo scelto di togliere i verbi e mantenere solo i nomi, gli aggettivi e gli avverbi (anche supplementari)**. Tale scelta trova riscontro nei modelli di espressione linguistica usati dagli utenti per recensire i locali (si vedano a tal riguardo alcune riflessioni presenti nel testo “Lingua, discorso e Società” di Stefania Spina). Dall’analisi delle frequenze assolute e relative delle Part of speech, infatti, ci siamo resi conto che fossero queste forme sintattiche ad avere l’incidenza più alta. Inoltre, **un approccio riduzionista facilita la visualizzazione per cluster, communities e per grafi**.

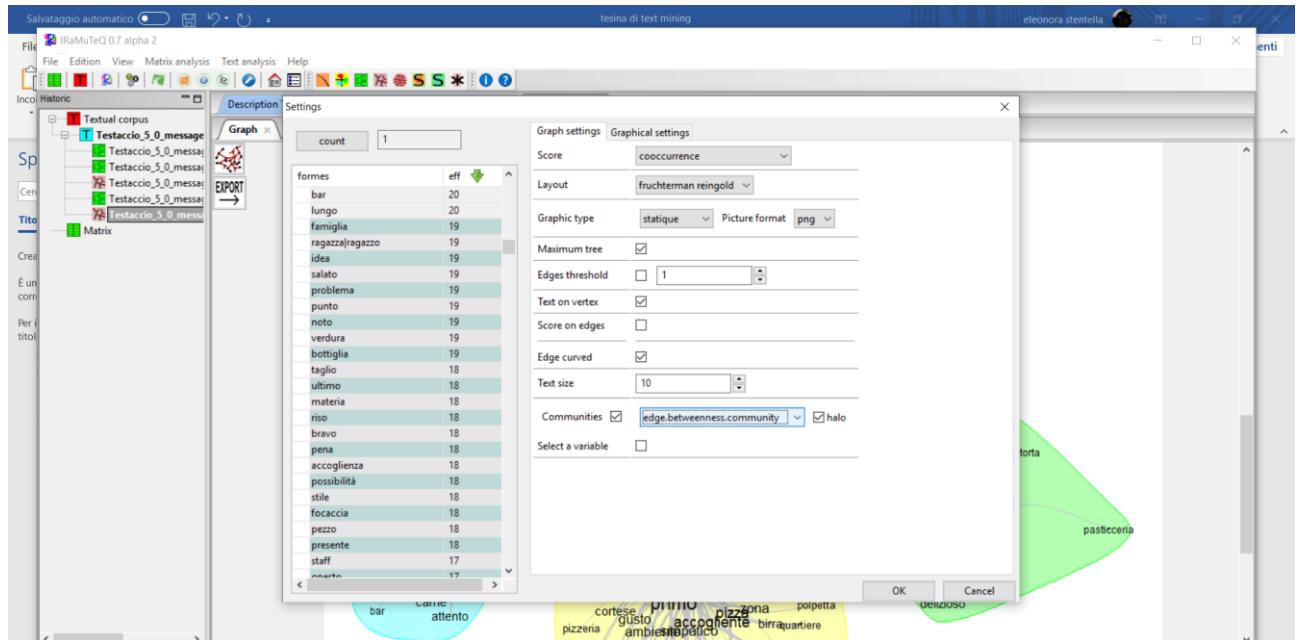
The screenshot shows the IRaMuTeQ 0.7 alpha software interface. The main window displays a network graph with various nodes representing words or concepts, such as "torta", "pasticcina", "palata", "fantastico", "caffè", "delizioso", "antipasto", "ambiente", "panino", "gusto", "cucina", "gentile", "ottimo", "primo", "romano", "pieno", "corse", "pollo", "cameriere", "eccellente", "carne", and "attento". The nodes are colored in different clusters: a cyan cluster at the bottom left, a yellow cluster in the center, and a green cluster at the bottom right.

A dialog box titled "Clés d'analyse" (Analysis keys) is open, showing a grid of 20 categories and their corresponding values:

	Choix des clés d'analyse	0=éliminé; 1=active; 2=supplémentaire
Adjectif	1	voir liste
Adjectif démonstratif	0	voir liste
Adjectif indéfini	0	voir liste
Adjectif interrogatif	0	voir liste
Adjectif numérique	0	voir liste
Adjectif possessif	0	voir liste
Adjectif supplémentaire	2	voir liste
Adverbe	1	voir liste
Adverbe supplémentaire	0	voir liste
Article défini	0	voir liste
Article indéfini	0	voir liste
Auxiliaire	0	voir liste
Chiffre	0	voir liste
Conjonction	0	voir liste
Formes non reconnues	0	voir liste
Nom commun	1	voir liste
Nom supplémentaire	0	voir liste
Onomatopée	0	voir liste
Pronom démonstratif	0	voir liste
Pronom indéfini	0	voir liste
Pronom personnel	0	voir liste
Pronom possessif	0	voir liste
Pronom relatif	0	voir liste
Préposition	0	voir liste
Verbe	0	voir liste
Verbe supplémentaire	0	voir liste

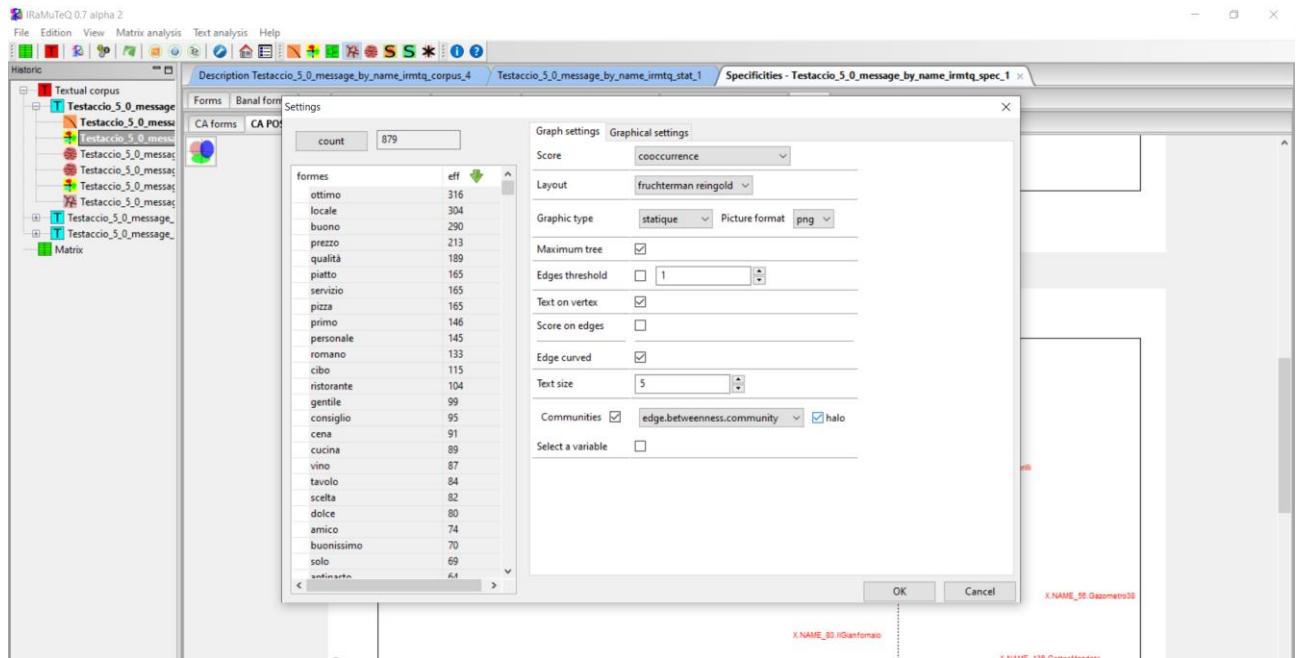
At the bottom of the dialog box, there is an "OK" button.

Abbiamo, inoltre, optato per un cut delle occorrenze.

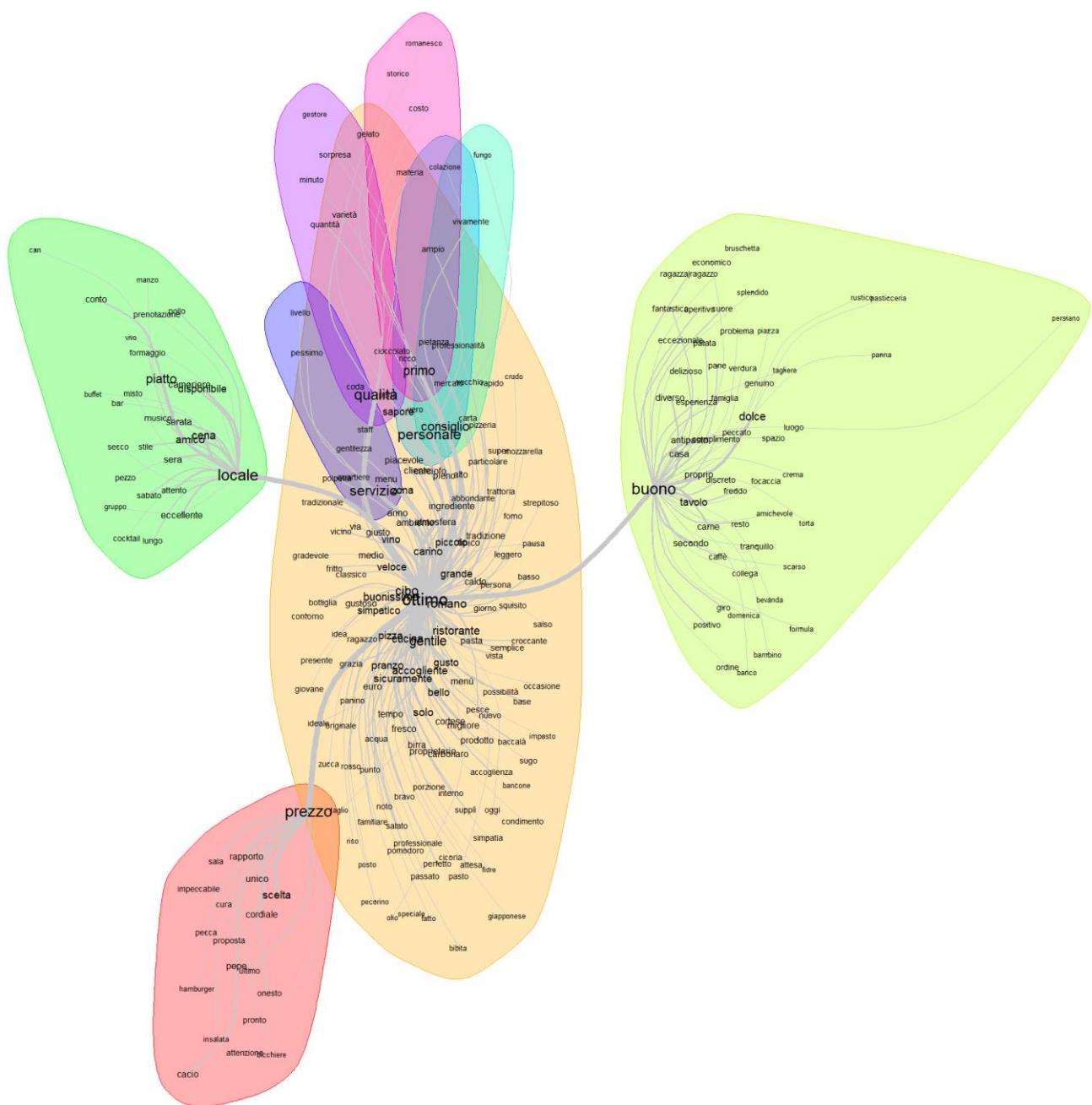


L'analisi delle similarità per la strategia 1 è stata eseguita selezionando solo le forme con un numero di occorrenze superiori a 12 e scegliendo come valore chiave **il grado di betweenness** in quanto permette di vedere chiaramente quali sono le parole che hanno un'interazione forte e che collegano due insiemi vicini tra loro. La betweenness, che in ambito delle Reti permette di identificare gli information brokers, in questo caso permette di capire quali servizi o piatti possano essere stati valutati tra il buono e l'ottimo. Di questo ne abbiamo già avuto intuizione mediante l'analisi grafica precedente dove abbiamo analizzato la distribuzione delle forme linguistiche nel corpus.

In questo caso si è scelto anche di togliere gli aggettivi supplementari al fine di ridurre il numero delle parole complessive ed avere un quadro più sintetico e chiaro.

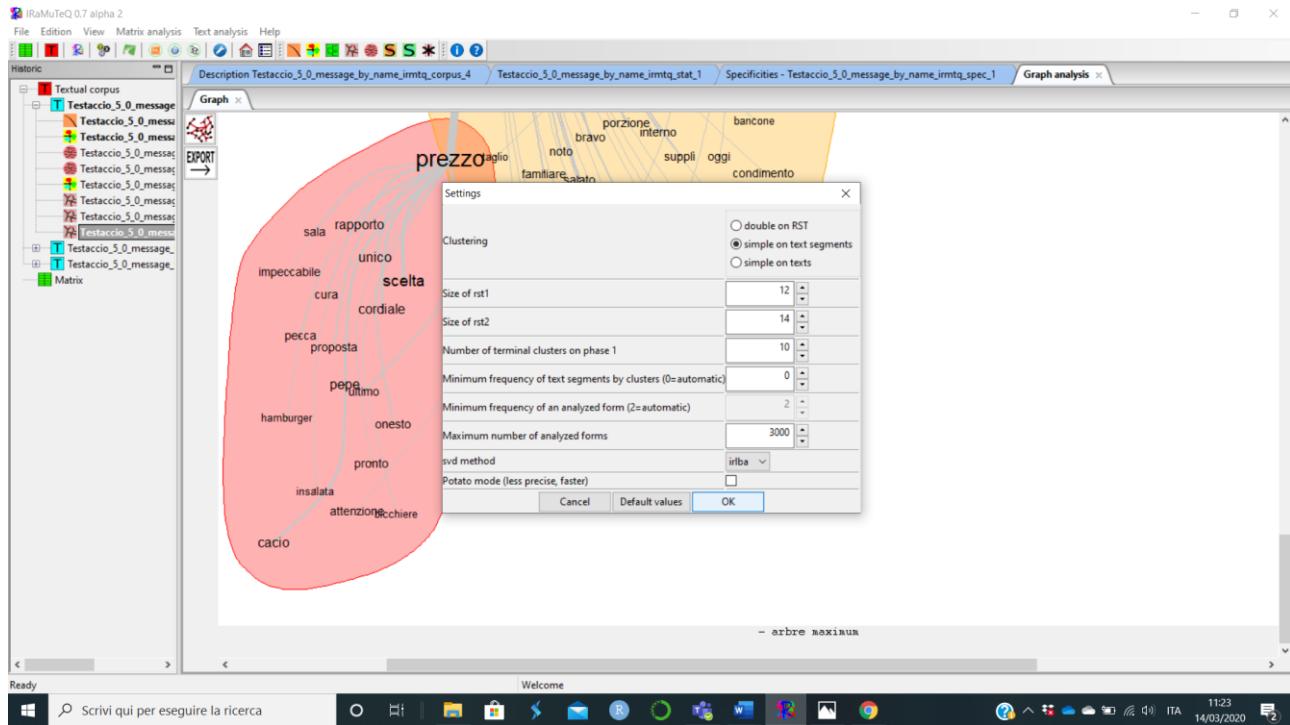


L'analisi delle somiglianze è una tecnica basata sulla **teoria dei grafi** (Flament, 1962). Presenta in un formato grafico la **struttura di un corpus**, distinguendo tra le parti condivise e le specificità delle variabili codificate. Ciò consente di far **emergere il collegamento tra le diverse forme nei segmenti di testo** (Marchand & Ratinaud, 2012).



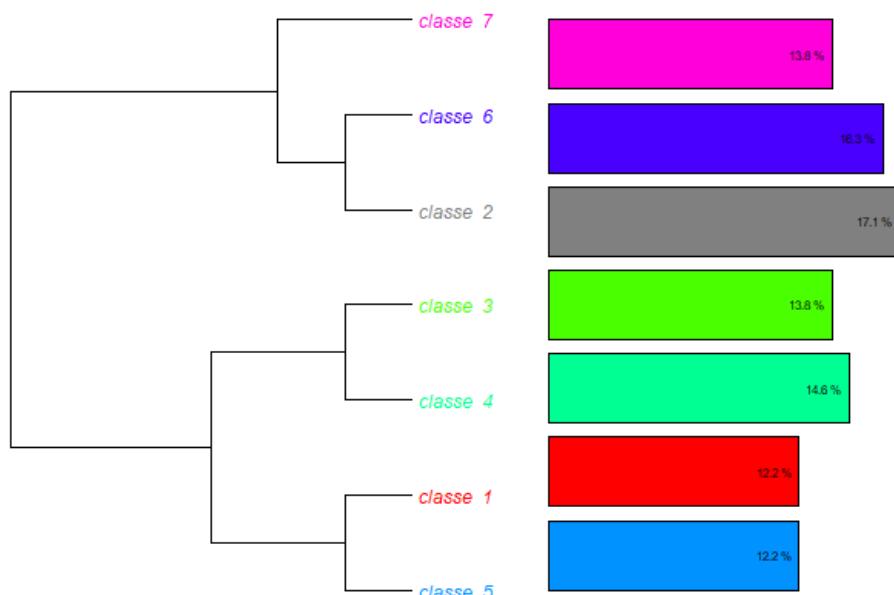
Cluster analysis

The Hierarchical Descending Classification



Un metodo utilizzato da Alceste (algoritmo di Iramuteq) è la classificazione gerarchica discendente. Questo metodo offre un approccio globale a un corpus. L'HDC dopo aver partizionato il corpus, identifica le classi di parole (moduli) statisticamente indipendenti. Queste classi sono interpretate attraverso i loro profili, che sono caratterizzati da specifiche forme correlate. L'HDC mostra i risultati attraverso un dendrogramma. Nel nostro caso abbiamo scelto di non stabilire il numero di cluster a monte bensì di lasciare che fosse l'algoritmo ad evidenziare i centroidi principali e a raccogliere i gruppi di parole in cluster. Ne escono 7 gruppi principali in relazione anche ai rating delle recensioni. Sotto i vari step di analisi effettuati.

Le classi derivano delle Chi squared table.



Volendo analizzare i cluster nel dettaglio si evince chiaramente che un rating è trasversale ai cluster identificati, per cui, ad eccezione dei cluster 4 e 5, gli altri risentono tutti del lessico usato nelle recensioni con rating tra il 4 e il 4.6. Questo è il motivo per cui abbiamo anche proceduto ad estrarre dei sub-corpus dividendoli in base ai rating in due gruppi ed escludendo il rating 4 (vedi paragrafo successivo). A rating simili corrisponde un uso similare degli aggettivi che inficiano la concreta demarcazione delle differenze.

Cluster 1 - include il lessico delle recensioni aventi rating medio pari a 4 e 4.2

Cluster 2 – 5 e 4.25

Cluster 3 – 4.6, 4.4 e 3,6

Cluster 4 – 2, 2.8, 3.2, 3, 3.4

Cluster 5 – 2.8 e 2.6

Cluster 6 – 4.8, 4.4, 4.6

Cluster 7 – 4 e 5

1 Cluster 1 15/123 12.2%	2 Cluster 2 21/123 17.07%	3 Cluster 3 17/123 13.82%	4 Cluster 4 18/123 14.63%	5 Cluster 5 15/123 12.2%	6 Cluster 6 20/123 16.26%	7 Cluster 7 17/123 13.82%	
n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	P
119	3	11	27.27	2.56		*RATING_4	NS (0.10927)
118	4	12	33.33	5.55		*RATING_4.2	0.01849
117	3	11	27.27	2.56	sw	dall	NS (0.10927)
116	2	5	40.0	3.76	sw	senz	NS (0.05239)
115	15	94	15.96	5.27	sw	I	0.02169
114	2	4	50.0	5.52	sw	quelli_che	0.01881
113	2	4	50.0	5.52	sw	per_finire	0.01881
112	3	12	25.0	2.04	nom	sugo	NS (0.15359)
111	3	12	25.0	2.04	adj	croccante	NS (0.15359)
110	3	12	25.0	2.04	adj	tradizionale	NS (0.15359)
109	3	12	25.0	2.04	adj	squisito	NS (0.15359)
108	6	30	20.0	2.26	nom	casa	NS (0.13299)
107	4	17	23.53	2.37	adj	lungo	NS (0.12395)
106	4	17	23.53	2.37	nom	vista	NS (0.12395)
105	4	17	23.53	2.37	adj	semplice	NS (0.12395)
104	14	95	14.74	2.52	adj	buono	NS (0.11256)
103	2	6	33.33	2.63	nom	cotto	NS (0.10472)
102	2	6	33.33	2.63	nom	prosciutto	NS (0.10472)
101	2	6	33.33	2.63	nom	ricordo	NS (0.10472)
100	2	6	33.33	2.63	nom	aspettativa	NS (0.10472)
99	2	6	33.33	2.63	nom	aspetto	NS (0.10472)
98	2	6	33.33	2.63	adj	efficiente	NS (0.10472)
97	2	6	33.33	2.63	nom	stella	NS (0.10472)
96	2	6	33.33	2.63	nom	amante	NS (0.10472)
95	2	6	33.33	2.63	nom	richiesta	NS (0.10472)
94	3	10	30.0	3.22	nom	luogo	NS (0.07263)
93	3	10	30.0	3.22	nom	spazio	NS (0.07263)
92	3	10	30.0	3.22	adj	originale	NS (0.07263)
91	5	20	25.0	3.66	adj	migliore	NS (0.05582)
90	2	5	40.0	3.76	nom	bisogno	NS (0.05239)
89	2	5	40.0	3.76	adj	azzecco	NS (0.05239)

Figura: Output del cluster n.1

1 Cluster 1 15/123 12.2%	2 Cluster 2 21/123 17.07%	3 Cluster 3 17/123 13.82%	4 Cluster 4 18/123 14.63%	5 Cluster 5 15/123 12.2%	6 Cluster 6 20/123 16.26%	7 Cluster 7 17/123 13.82%	
n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
70	4	11	36.36	3.18		*RATING_5	NS (0.07476)
69	1	1	100.0	4.9		*RATING_4.25	0.02690
68	3	6	50.0	4.83	sw	di,fronte	0.02796
67	5	17	29.41	2.12	adj	nuovo	NS (0.14527)
66	20	104	19.23	2.21	adj	ottimo	NS (0.13679)
65	6	21	28.57	2.36	adj	cordiale	NS (0.12410)
64	4	12	33.33	2.48	adj	squisito	NS (0.11507)
63	3	8	37.5	2.52	adj	sottile	NS (0.11229)
62	3	8	37.5	2.52	nom	strada	NS (0.11229)
61	3	8	37.5	2.52	nom	consegna	NS (0.11229)
60	3	8	37.5	2.52	nom	cura	NS (0.11229)
59	7	25	28.0	2.65	adj	gustoso	NS (0.10380)
58	7	25	28.0	2.65	nom	birra	NS (0.10380)
57	19	94	20.21	2.78	nom	prezzo	NS (0.09570)
56	6	20	30.0	2.82	adj	migliore	NS (0.09316)
55	9	34	26.47	2.93	adj	simpatico	NS (0.08689)
54	2	4	50.0	3.17	nom	avocado	NS (0.07518)
53	2	4	50.0	3.17	nom	box	NS (0.07518)
52	2	4	50.0	3.17	nom	freschezza	NS (0.07518)
51	2	4	50.0	3.17	nom	giornata	NS (0.07518)
50	2	4	50.0	3.17	nom	sgabello	NS (0.07518)
49	2	4	50.0	3.17	nom	viaggio	NS (0.07518)
48	2	4	50.0	3.17	adj	centrifugo	NS (0.07518)
47	2	4	50.0	3.17	adj	sano	NS (0.07518)
46	2	4	50.0	3.17	adj	breve	NS (0.07518)
45	2	4	50.0	3.17	nom	torno	NS (0.07518)
44	4	11	36.36	3.18	nom	impasto	NS (0.07476)
43	4	11	36.36	3.18	nom	sorpresa	NS (0.07476)
42	6	19	31.58	3.34	nom	varietà	NS (0.06762)
41	8	27	29.63	3.85	adj	caldo	0.04967
40	8	27	29.63	3.85	adj	fresco	0.04967

Figura: Output del cluster n.2

1 Cluster 1 15/123 12.2%	2 Cluster 2 21/123 17.07%	3 Cluster 3 17/123 13.82%	4 Cluster 4 18/123 14.63%	5 Cluster 5 15/123 12.2%	6 Cluster 6 20/123 16.26%	7 Cluster 7 17/123 13.82%	
n...	eff. s.t.	eff. total	pourcentage	chi2	Type	forme	p
111	4	15	26.67	2.37		*RATING_4.6	NS (0.12395)
110	4	12	33.33	4.25		*RATING_4.4	0.03924
109	4	7	57.14	11.7		*RATING_3.6	0.00062
108	2	6	33.33	2.02	sw	in_particolare	NS (0.15562)
107	2	5	40.0	3.0	sw	non_so	NS (0.08332)
106	6	24	25.0	3.13	sw	dell	NS (0.07693)
105	9	41	21.95	3.41	sw	in_questo	NS (0.06468)
104	5	16	31.25	4.69	sw	nell	0.03032
103	4	8	50.0	9.4	sw	é	0.00216
102	12	40	30.0	13.03	sw	all	0.00030
101	2	6	33.33	2.02	nom	arredamento	NS (0.15562)
100	2	6	33.33	2.02	nom	arredo	NS (0.15562)
99	2	6	33.33	2.02	nom	cotto	NS (0.15562)
98	2	6	33.33	2.02	nom	centro	NS (0.15562)
97	2	6	33.33	2.02	nom	sedia	NS (0.15562)
96	2	6	33.33	2.02	nom	spaghetti	NS (0.15562)
95	2	6	33.33	2.02	nom	amante	NS (0.15562)
94	2	6	33.33	2.02	nom	bancone	NS (0.15562)
93	8	39	20.51	2.15	adj	veloce	NS (0.14285)
92	13	74	17.57	2.19	nom	servizio	NS (0.13900)
91	3	10	30.0	2.39	adj	genuino	NS (0.12195)
90	3	10	30.0	2.39	adj	speciale	NS (0.12195)
89	3	10	30.0	2.39	nom	focaccia	NS (0.12195)
88	7	31	22.58	2.67	nom	serata	NS (0.10226)
87	4	14	28.57	2.89	nom	carta	NS (0.08936)
86	2	5	40.0	3.0	adj	impossibile	NS (0.08332)
85	2	5	40.0	3.0	nom	parcheggio	NS (0.08332)
84	2	5	40.0	3.0	adj	pazzesco	NS (0.08332)
83	2	5	40.0	3.0	nom	ingresso	NS (0.08332)
82	2	5	40.0	3.0	adj	popolare	NS (0.08332)
81	2	5	40.0	3.0	adj	caro	NS (0.08332)

Figura: Output del cluster n.2

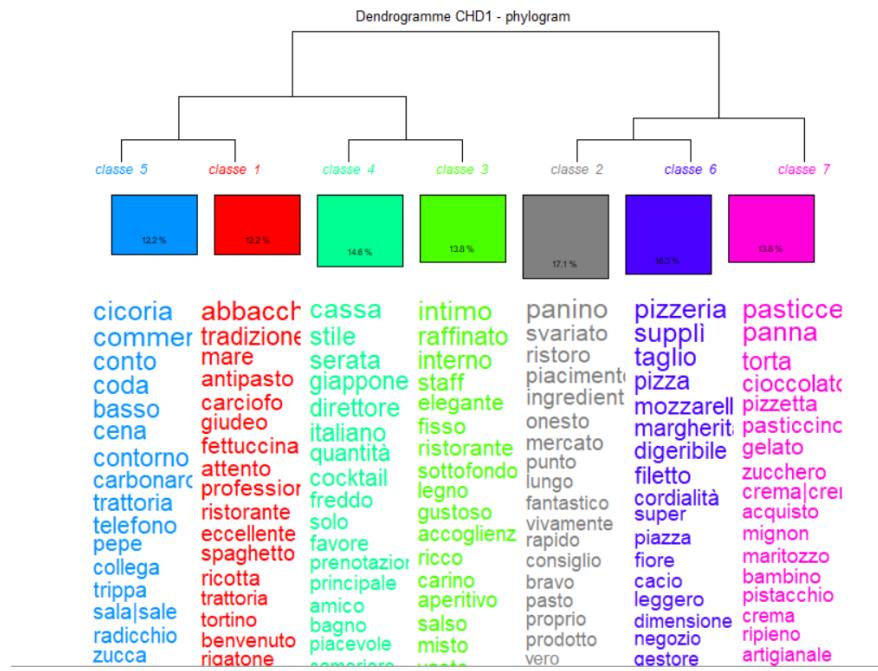


Figura: Risultato della Hierarchical Descending Classification

Nel caso successivo abbiamo creato una visualizzazione differente con nuvole di parole.



Rappresentazione dei cluster sui fattori

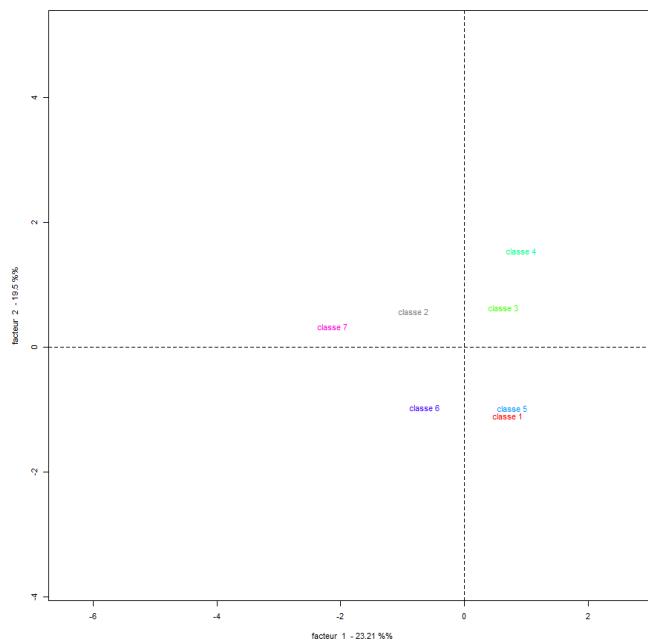
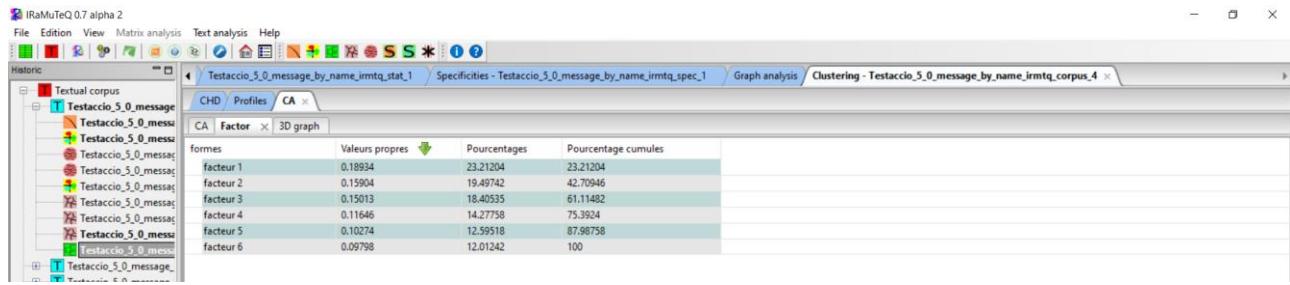


Figura: Quadranti del posizionamento dei cluster

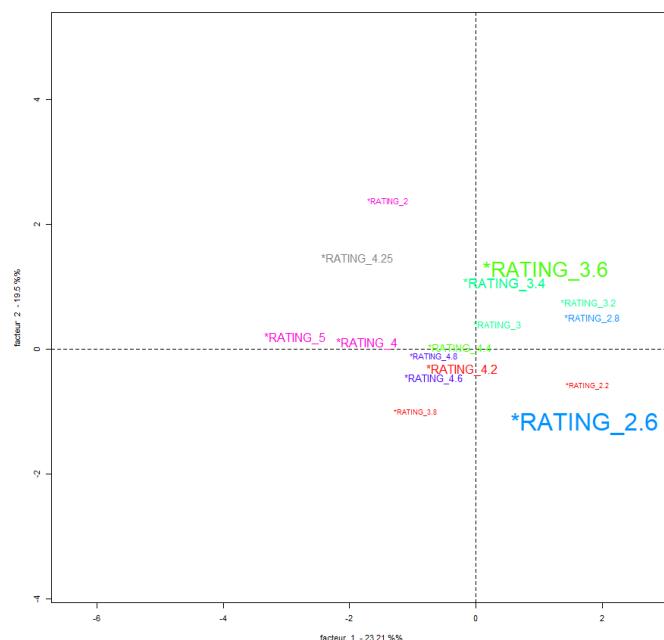


Figura: Quadranti dei rating

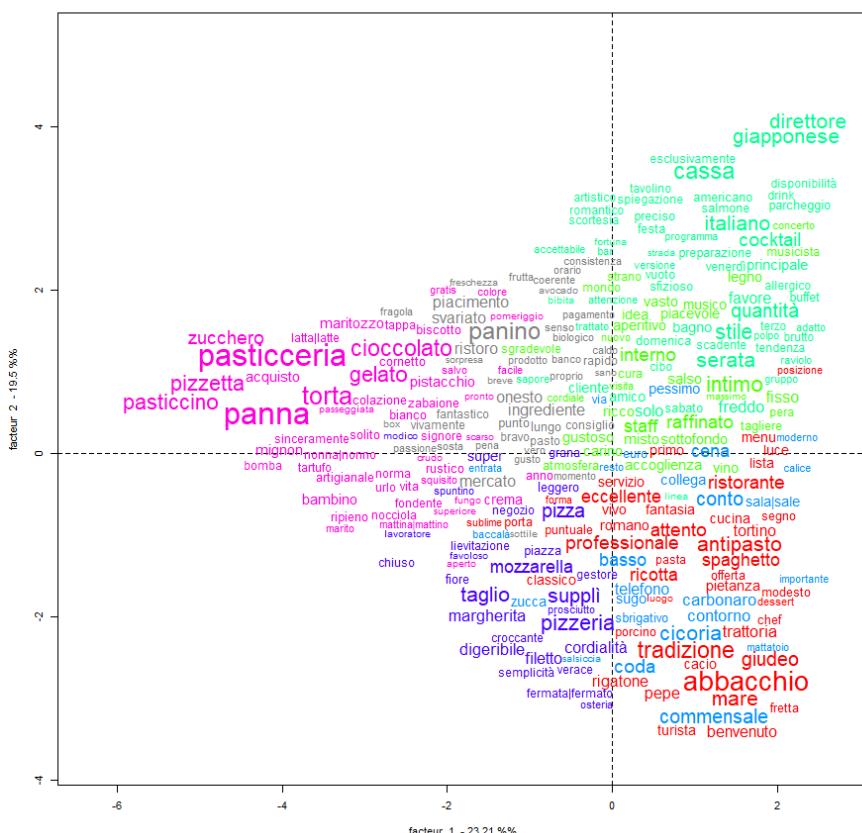


Figura: Quadranti delle variabili attive colorate in base all'appartenenza ai cluster

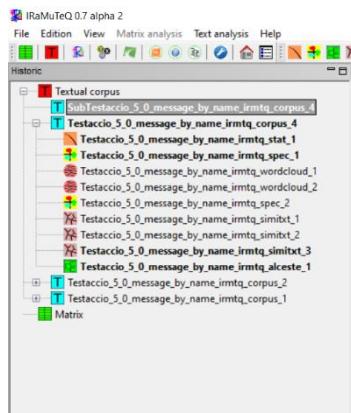
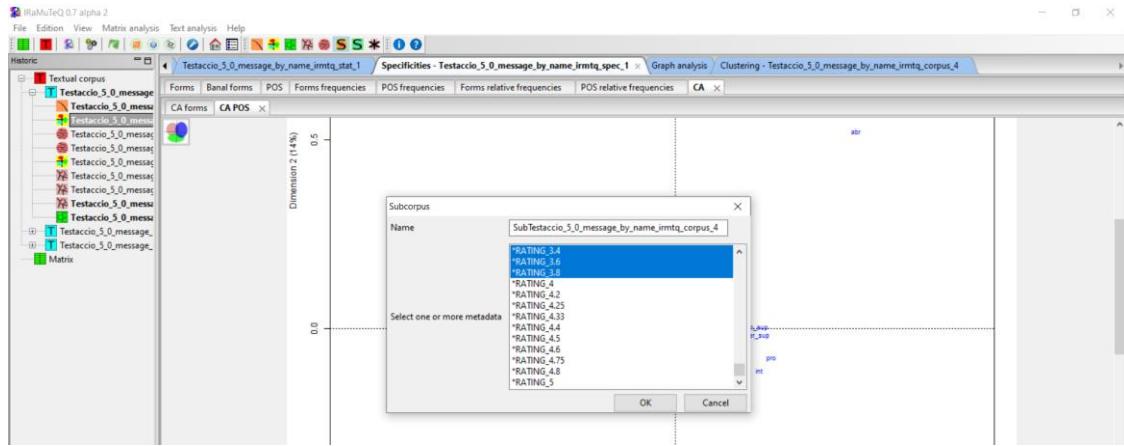
Si può osservare come i primi quattro fattori spieghino il 75,4% della varianza totale. Per facilità di lettura si sono rappresentati nei grafici solo i primi due fattori, che insieme spiegano il 42,7% di varianza.

L'osservazione del posizionamento dei cluster, dei rating e delle parole sui primi due fattori lascia emergere alcune considerazioni. Il primo fattore sembra essere caratterizzato molto bene dai rating: sui valori negativi si posizionano, infatti, i rating più alti e, mano a mano che ci si sposta sui valori positivi, si trovano quelli più bassi. In particolare, il cluster 7, che presenta i valori più elevati di rating, è quello che si distanzia maggiormente dagli altri posizionandosi nel primo quadrante in corrispondenza dei valori negativi del primo fattore ed è ben descritto dalle parole che rappresentano il tema dei dolci ("pasticceria", "panna", "torta", "pasticcino", "cioccolato", "gelato").

Leggendo i risultati lungo il secondo fattore, invece, sembrano emergere, in corrispondenza dei valori negativi, termini che caratterizzano le pizzerie (“pizzeria”, “taglio”, “margherita”) e i ristoranti della tradizione con piatti tipici della cucina romana (“abbacchio”, “giudeo”, “coda”, “supplì”), mentre nella direzione opposta, verso i valori positivi, si ritrovano parole che descrivono una cucina più variegata con offerte anche diverse dal classico ristorante (“cocktail”, “giapponese”, “panino”, “aperitivo”).

Subcorpus per metadata

Iramuteq consente di fare **ripartizioni di un corpus mediante l'utilizzo dei metadata o di altre informazioni presenti**. Nel nostro caso abbiamo scelto di fare un test con una ripartizione molto semplice prendendo un sub-corpus costituito da tutti i ristoranti della strategia 1 aventi un rating compreso tra 1 e 3.8 (come nella figura sottostante) e con rating uguale a 5. La scelta non è casuale e si ricollega all'analisi fattoriale realizzata in precedenza.

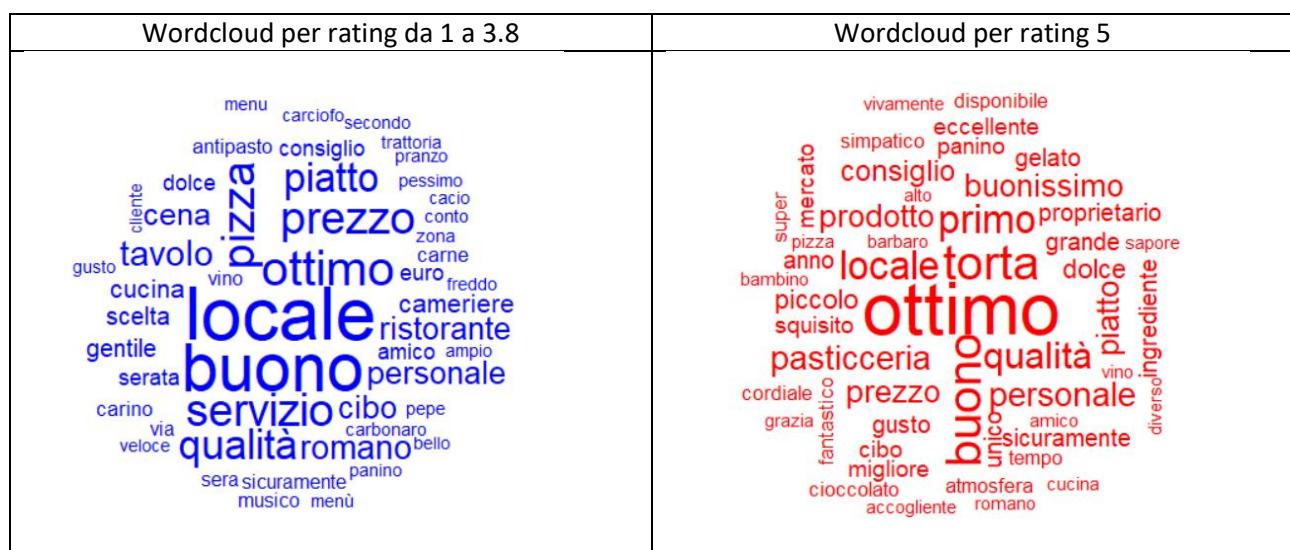


Iramuteq ha costituito un nuovo subcorpus che può essere analizzato come ripartizione a sé stante proveniente dal precedente. Tale operazione poteva essere eseguita anche selezionando una lista di ristoranti.

Wordcloud per sub-corpus

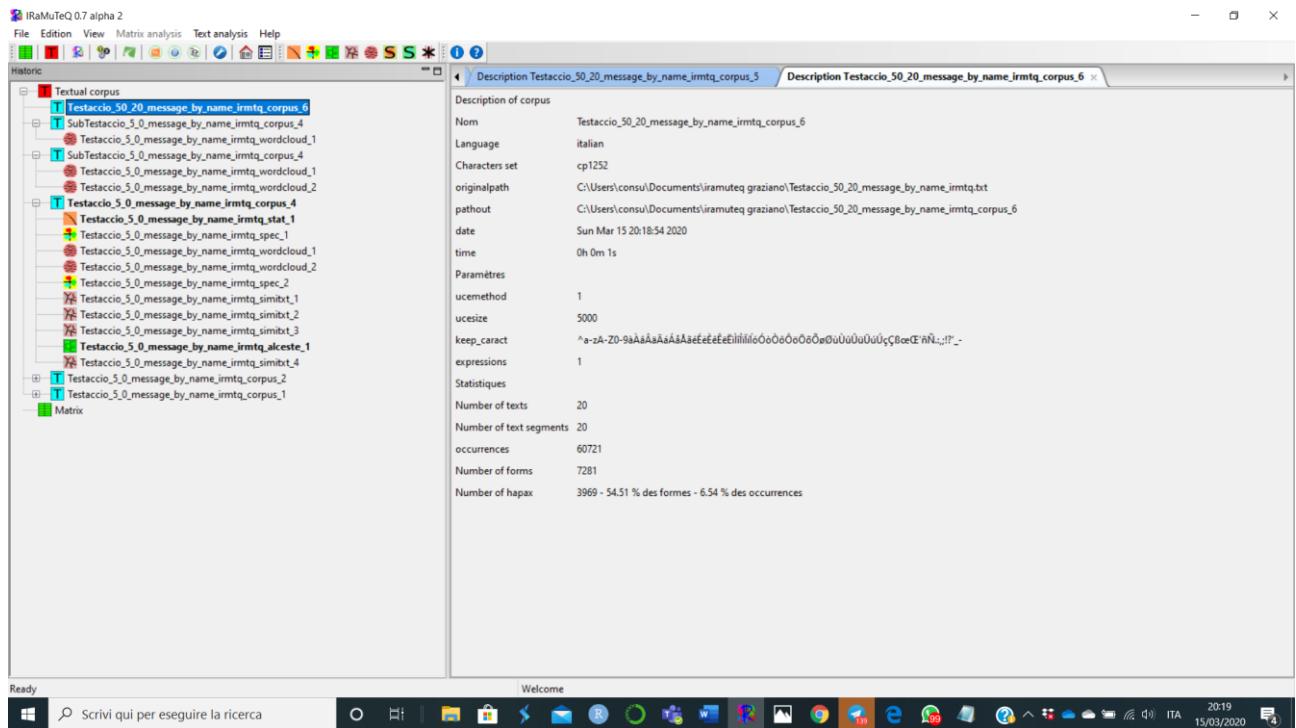
Analizzando le wordcloud sulla base dei sub-corpus per rating otteniamo le seguenti raffigurazioni.

Il limite stabilito è di massimo 50 parole a wordcloud. Abbiamo deciso di estromettere le recensioni con rating 4 in quanto, già dall'analisi fattoriale e dei cluster, era possibile scorgere l'utilizzo condiviso di molti aggettivi che non avrebbero permesso di evidenziare le differenze più sostanziali. Notiamo che nella wordcloud con subset rating 5 ha maggiori evidenze di "eccellenza". Compiono parole come "torta", "pasticceria", "gelato" non presenti nella wordcloud precedente.

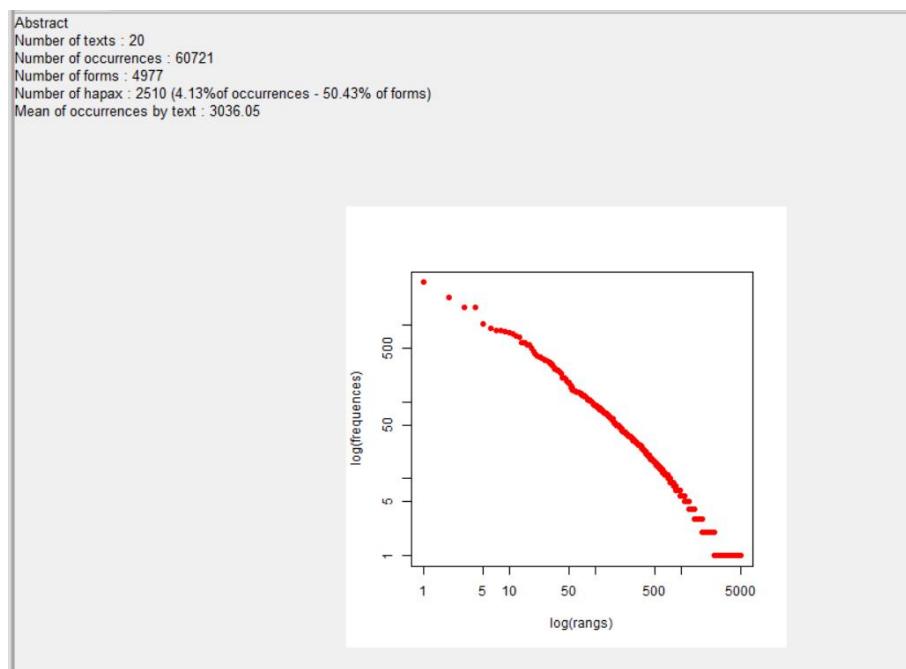


Strategia 2

Per fare un'analisi usando il file della strategia 2 è stato necessario rendere equipollenti il numero dei testi presenti nel corpus (=20 corrispondenti a 20 ristoranti) con il numero di segmenti. Per ottenere il risultato sotto riportato è stato fissato un limite massimo di occorrenze a 5000 per ogni testo.



Come fatto in precedenza è stata poi eseguita un'analisi statistica riepilogativa del testo.



Come già evidenziato nell'analisi per la strategia 1 anche nella 2 si evince che le forme più frequenti siano gli aggettivi e i nomi. Primi della lista "locale" (451), "buono" (417) e "ottimo" (374).

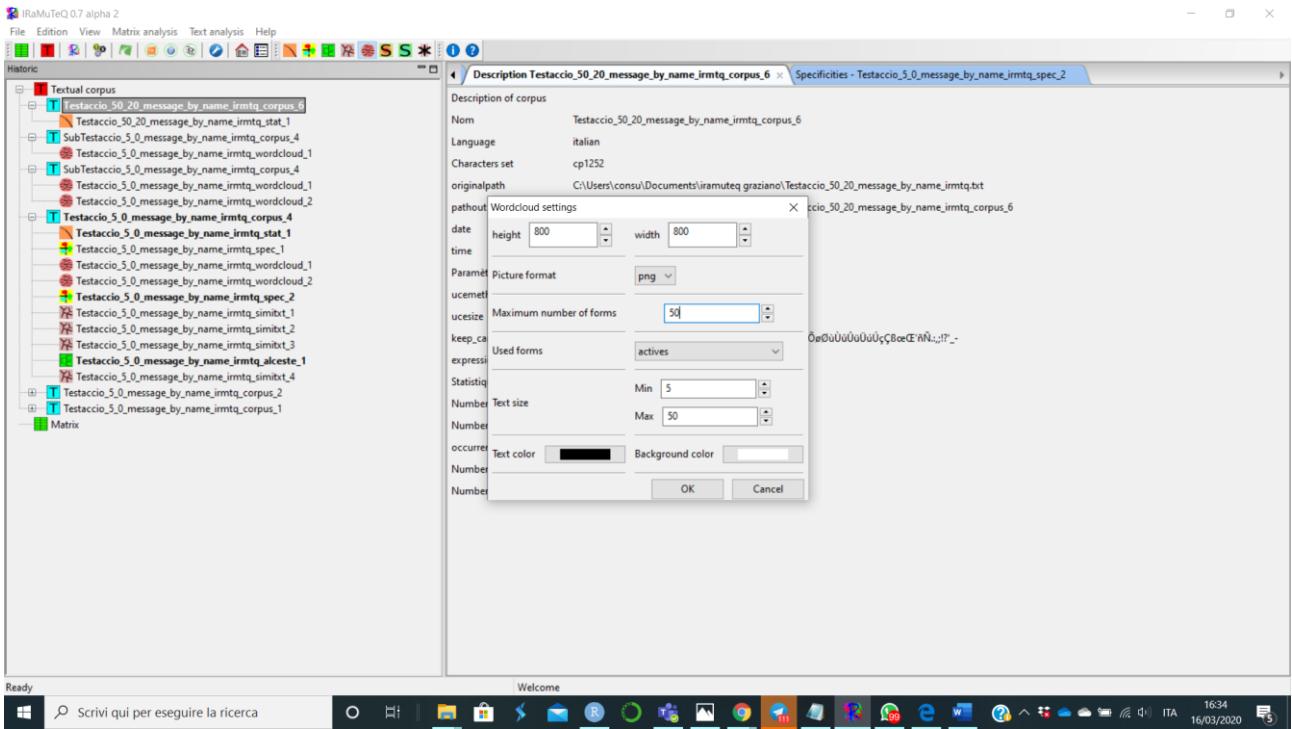
The screenshot shows the IRaMuTeQ 0.7 alpha 2 software interface. On the left, there is a tree view of a 'Textual corpus' containing several sub-folders like 'Testaccio_50_20_message_by_name_irmtq_corpus_6' and 'Testaccio_5_0_message_by_name_irmtq_corpus_4'. On the right, a table titled 'Testaccio_50_20_message_by_name_irmtq_stat_1' displays word statistics. The columns are 'Form', 'Freq.', and 'POS'. The top words are:

Form	Freq.	POS
locale	451	adj
buono	417	adj
ottimo	374	adj
piatto	343	nom
prezzo	324	nom
qualità	305	nom
romano	288	adj
servizio	255	nom
personale	249	adj
primo	240	adj
ristorante	233	nom
cucina	205	nom
cibo	199	nom
tavolo	165	nom
carbonaro	138	adj
cameriere	137	nom
gentile	136	adj
pizza	134	nom
dolce	133	adj
cena	131	nom
amico	127	nom
vino	126	nom
consiglio	122	nom
pepe	119	nom
scelta	118	nom
cacio	112	nom
gusto	105	nom
solo	102	adj
tipico	101	adj
antipasto	96	nom
secondo	92	adj
carciofo	91	nom
euro	90	adj
sicuramente	88	adv

Wordcloud

Anche la wordcloud, creata a partire delle stesse regole usate nel precedente studio con un massimo di 50 parole, coincide con la precedente.

The screenshot shows the IRaMuTeQ 0.7 alpha 2 software interface with a configuration dialog open. The dialog is titled 'Description Testaccio_50_20_message_by_name_irmtq_corpus_6' and includes fields for 'Nom' (Testaccio_50_20_message_by_name_irmtq_corpus_6), 'Language' (italian), 'Characters set' (cp1252), 'originalpath' (C:\Users\consu\Documents\iramuteq graziano\Testa), 'pathout' (C:\Users\consu\Documents\iramuteq graziano\Testa), and 'date' (Sun Aug 16 20:18:44 2020). A 'Settings' sub-dialog is also visible, showing parameters like 'Lemmatization' (yes), 'ucemethod' (no), 'ucesize', 'keep_carac', 'expressions', 'Statistiques' (Dictionary), 'Number of', 'Number of forms', 'Number of hapax', and file paths. To the right, a 'Clés d'analyse' (Analysis keys) dialog is open, showing a grid of keys and their counts. The grid includes categories like 'Adjectif', 'Conjonction', 'Adjectif démonstratif', 'Adjectif indéfini', 'Adjectif interrogatif', 'Adjectif numérique', 'Adjectif possessif', 'Adjectif supplémentaire', 'Adverb', 'Adverb supplémentaire', 'Article défini', 'Article indéfini', 'Auxiliaire', and 'Chiffre', each with a count from 0 to over 1000. The 'OK' button at the bottom right of the dialog is highlighted.



In questo caso abbiamo generato una sola wordcloud perché ci siamo resi conto che la media dei rating per ciascun ristorante era molto alta e compresa tra 3 stelle e 4 stelle.



Similarity analysis

Al fine di rendere l'analisi delle similarità più interessante, abbiamo deciso di applicare la suddetta tecnica sulla variabile rating e sui diversi ristoranti.

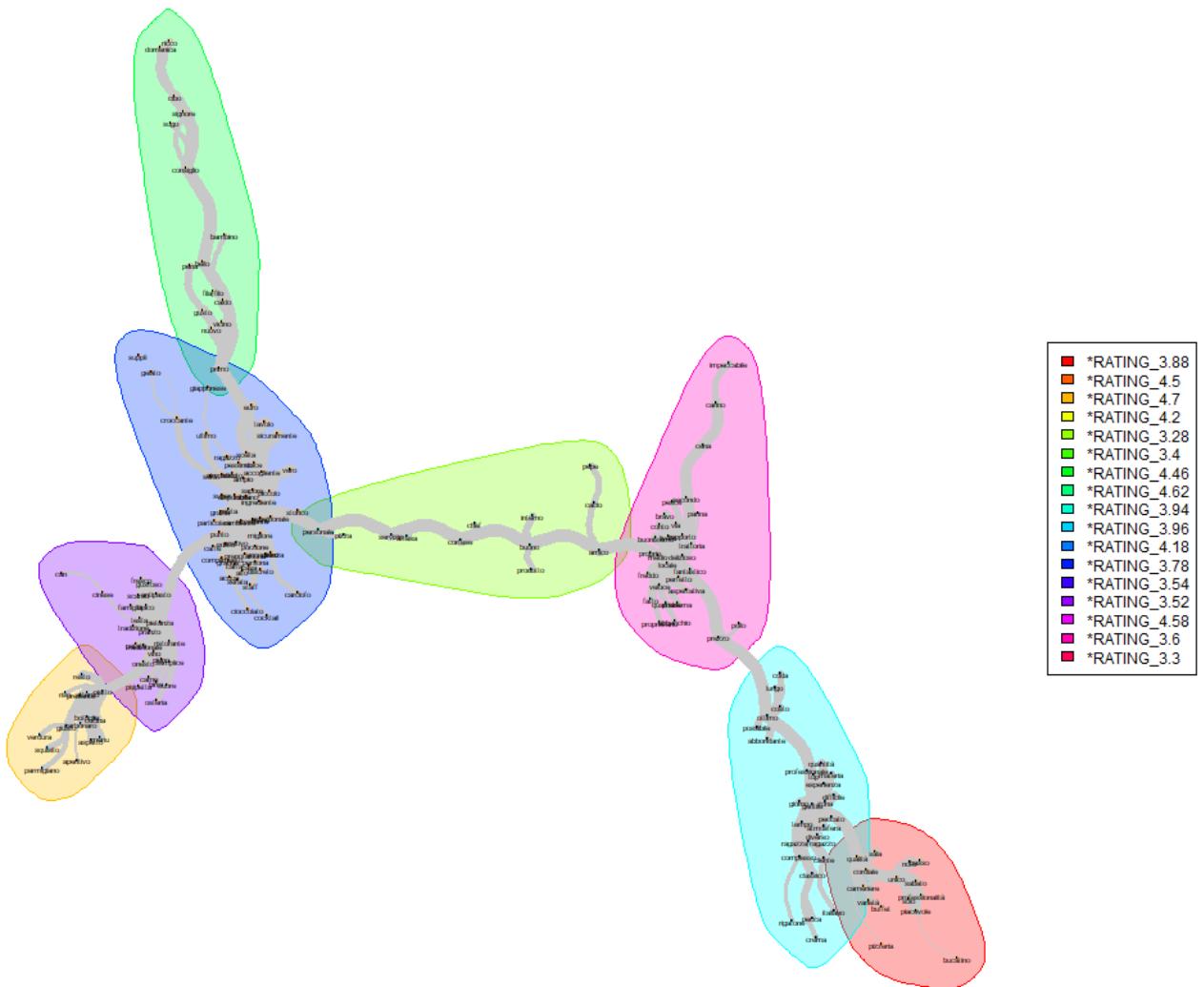


Figura: Analisi delle similarità, Grafico realizzato in Iramuteq

La difficoltà di lettura dell'immagine ottenuta sia con tale applicazione che con l'analogia suddivisione per ristoranti (n. 20) ci ha spinti a raggruppare i rating in 2 gruppi distinti: il primo fino a 3.99 di rating medio e il secondo con valore superiore a 4. E' stato poi caricato un diverso file di lavoro in Iramuteq ma i **risultati ottenuti non sono stati soddisfacenti**.

Similarity analysis in Gephi

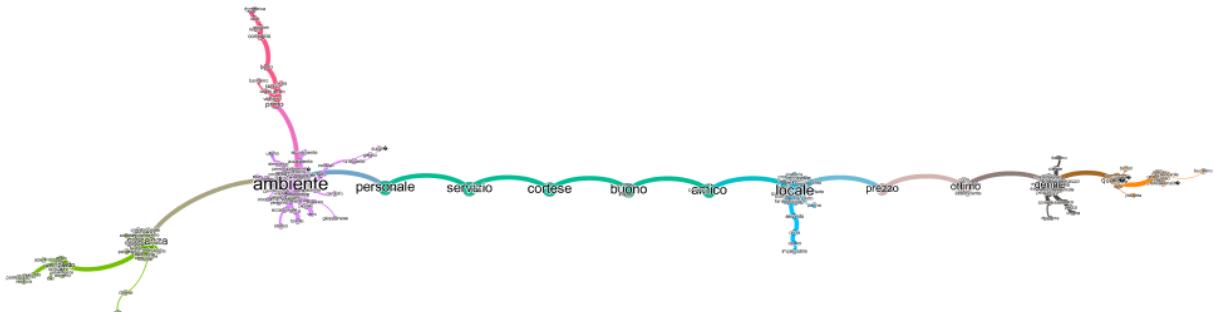
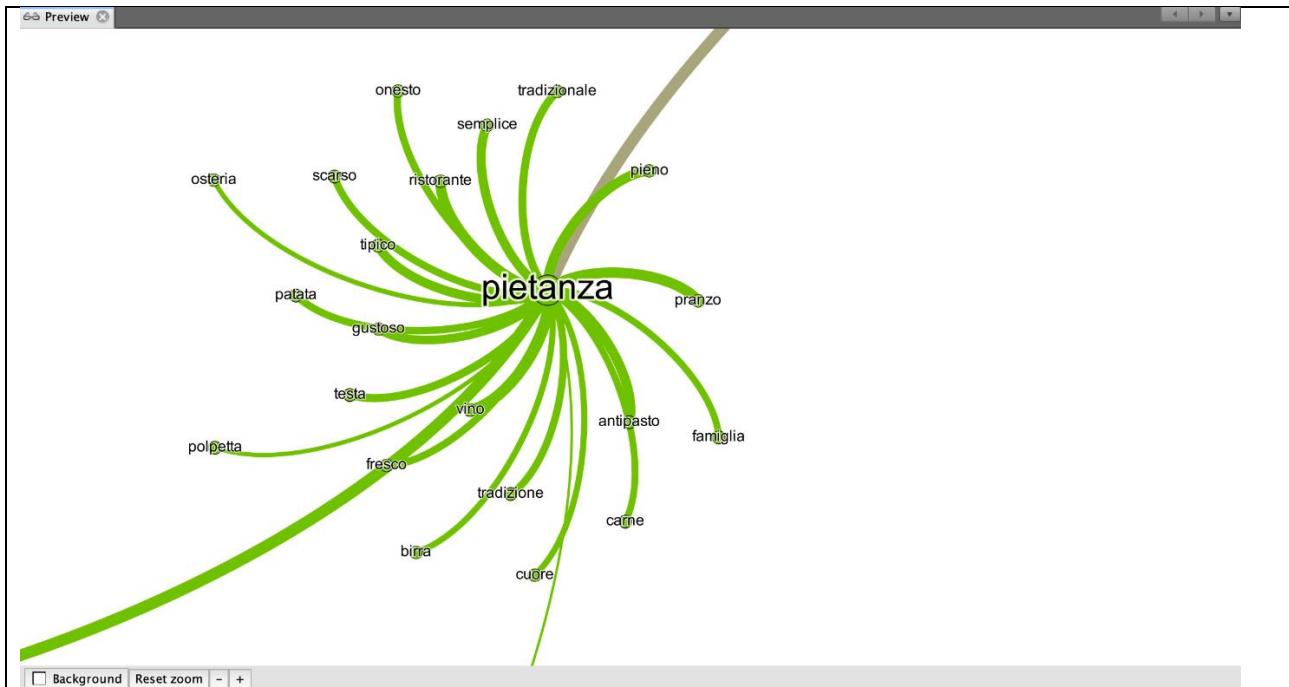


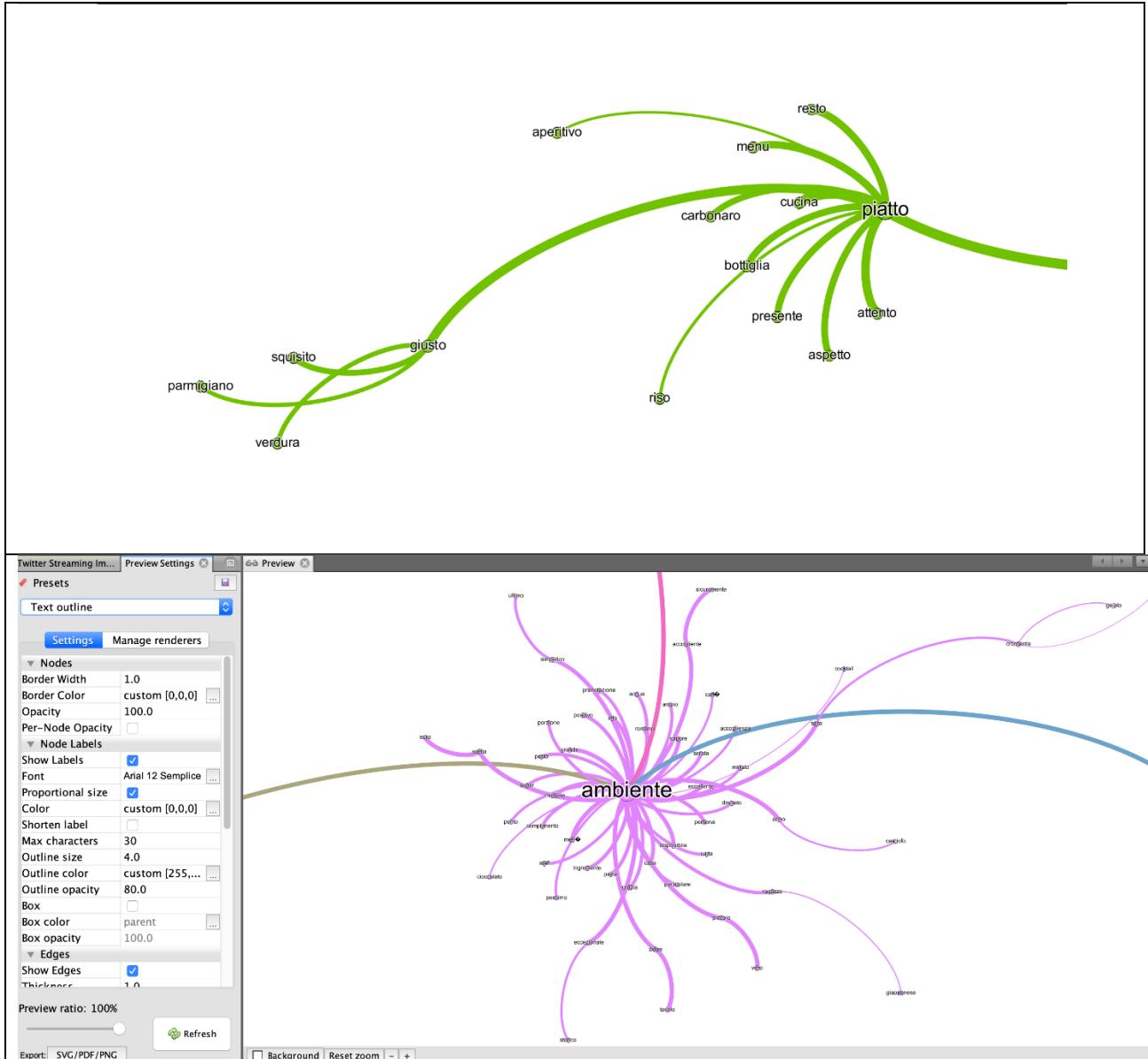
Figura: Analisi delle similarità, Grafico realizzato in Gephi utilizzando la modularity class per la clustering e gli algoritmi Atlas 2 ed Expansion

Come si può osservare dal grafico, “personale”, “servizio”, “cortese” e “buono” appartengono allo stesso cluster in verde.

Il cluster nel quale capeggia il nodo “ambiente” (ego-network) ha a corollario i termini “positivo”, “eccellente”, “discreto”, “persona”, “staff” e “posizione” ed anche l’aggettivo “pessimo” (in quanto in tale analisi **il corpus era stato analizzato per intero**).

Seguono alcuni dettagli.





Analysis in Voyant strategia 1

Il tool Voyant è attualmente usato nelle analisi di Text Mining da parte dell'università di Lione e trova ampio utilizzo anche da parte dei docenti della Miami University. Il tool permette di avere una panoramica esaustiva sui testi e una dashboard di controllo che include una parte di linguistica computazionale dove ogni singola parola viene contestualizzata.

Pre-processing su file di strategia 1

Ai fini dell'esercitazione sono state considerate tutte le recensioni senza includere i nomi dei ristoranti e i ratings, **creando un corpus unico**. Il corpus è stato caricato e sono state tolte le stopwords della lingua italiana, come nelle analisi precedenti, scegliendo però di tenere alcune parole (come ad esempio "molto"). Il software non consente di fare lo stemming.

Wordcloud e Summary statistics Voyant



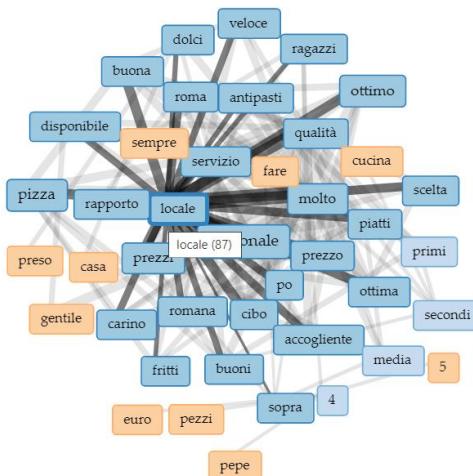
Il corpus contiene 1 documento con 43861 totale parole e con 6334 forme di parola uniche.

Densità del vocabolario: 0.144 (anche conosciuta come diversity del lessico)

Average Words Per Sentence: 43861.0

Parole più frequenti nel corpus [molto](#) (335); [locale](#) (290); [qualità](#) (189); [servizio](#) (163); [personale](#) (145)

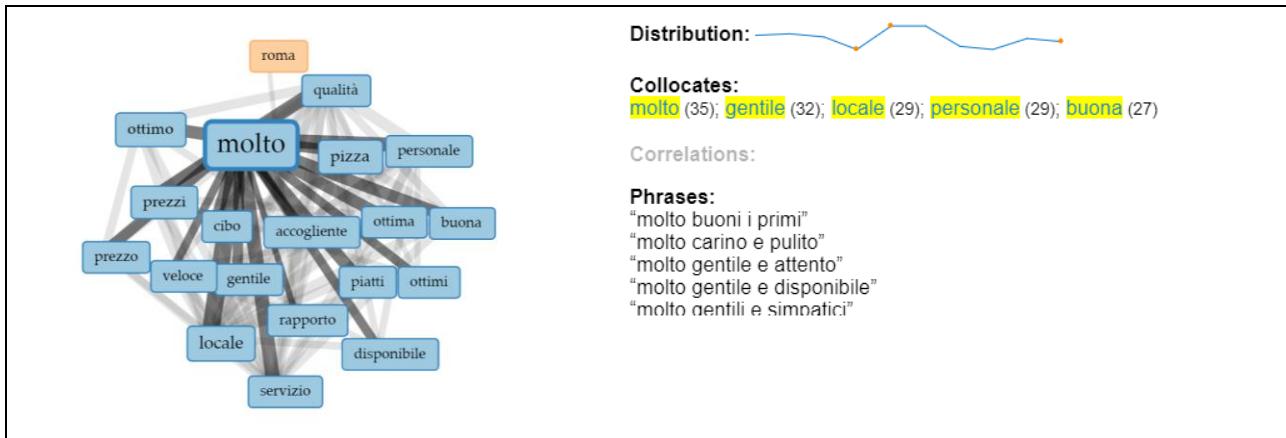
Analisi dei link all'interno del corpus



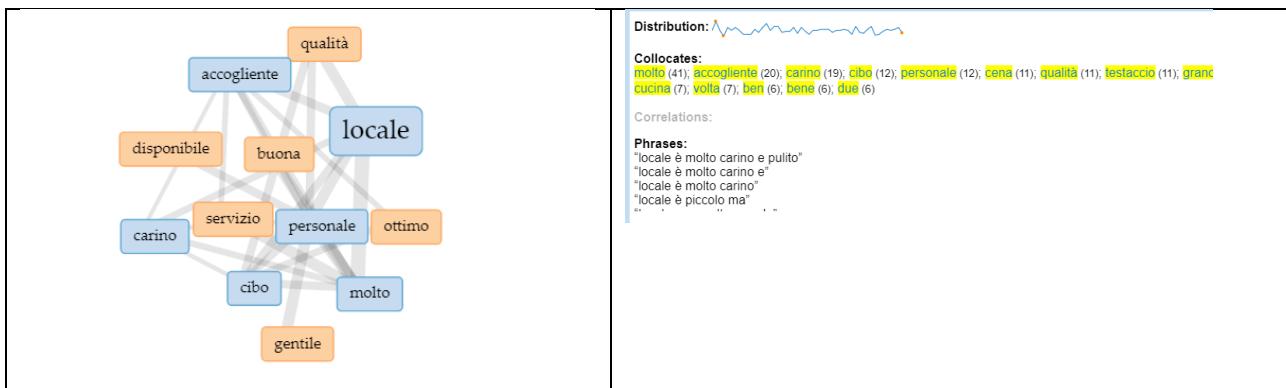
Il grafo delle collocazioni mostra un grafo a rete delle parole che con maggiore frequenza appaiono in prossimità di una parola data.

Analisi dei sintagmi per le principali parole presenti nel testo

Sono state prese le prossimità dell'aggetto "molto" fino a un massimo di quattro parole.



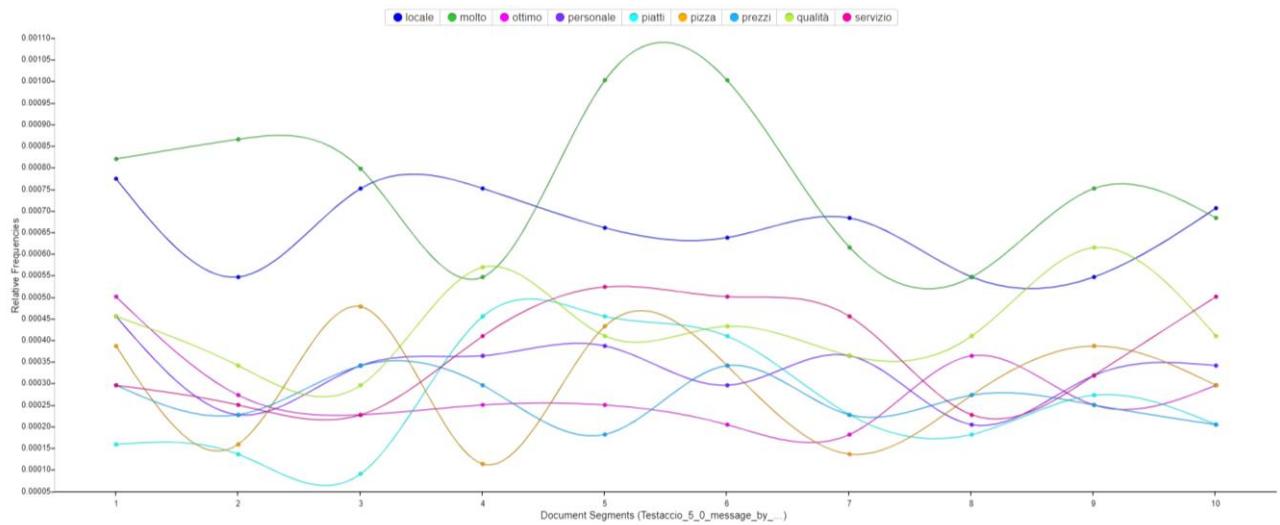
In questo esempio sono prese le prossimità della parola "locale" fino a un massimo di sette. In blu si possono vedere le parole più correlate alla parola oggetto di indagine.



In questo caso, sono state considerate le prossimità della parola "piatti", visualizzate con una tabella di frequenze delle frasi in base al numero di parole che contengono.

Frequency	Length	Phrase	Distributions
2	5	piatti della cucina romana e	2
2	4	piatti della tradizione e	2
6	4	piatti della tradizione romana	6
2	4	piatti della tradizione romanesca	2
2	3	piatti del giorno	2
2	3	piatti di pesce	2
3	3	piatti gustosi e	3
2	2	piatti a	2
3	2	piatti buoni	3
2	2	piatti che	2
2	2	piatti classici	2
2	2	piatti da	2
6	2	piatti e	6
2	2	piatti la	2
2	2	piatti non	2
2	2	piatti romani	2
4	2	piatti sono	4
2	2	piatti tipicamente	2

Analisi degli andamenti delle principali parole nel corpus



L'**analisi degli andamenti** delle parole con maggior frequenza mette in evidenza come si modifica la presenza delle parole lungo tutto il corpus (si ricorda che ai fini di questo esercizio si è scelto di trattare le recensioni come corpus unico). Ci sono, ad esempio, recensioni di ristoranti che, nella parte centrale del corpus, hanno determinato l'incremento relativo dell'aggettivo "molto" o della parola "qualità" soprattutto verso la fine; il termine "locale", invece, mantiene per lo più un andamento stabile, in particolare nella parte centrale del corpus del testo.

Analysis in Voyant strategia 2



Figura: Wordcloud

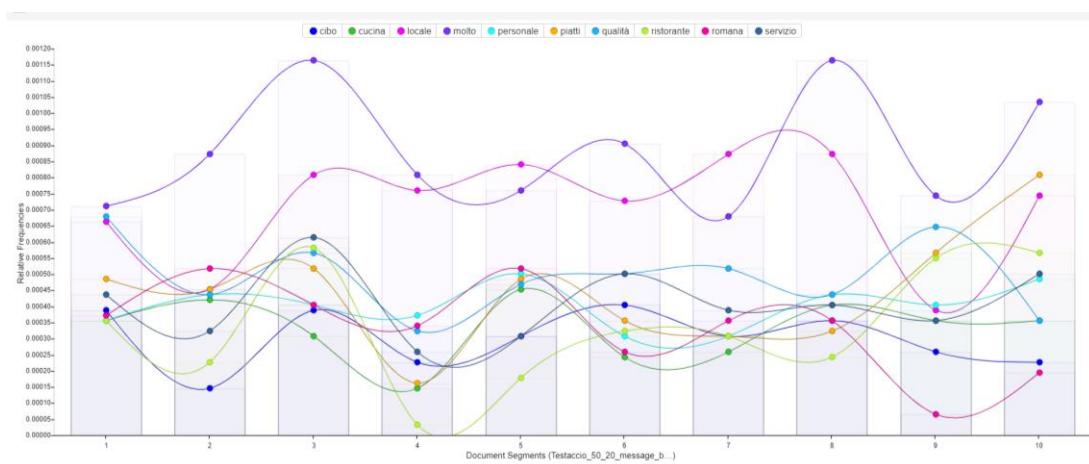
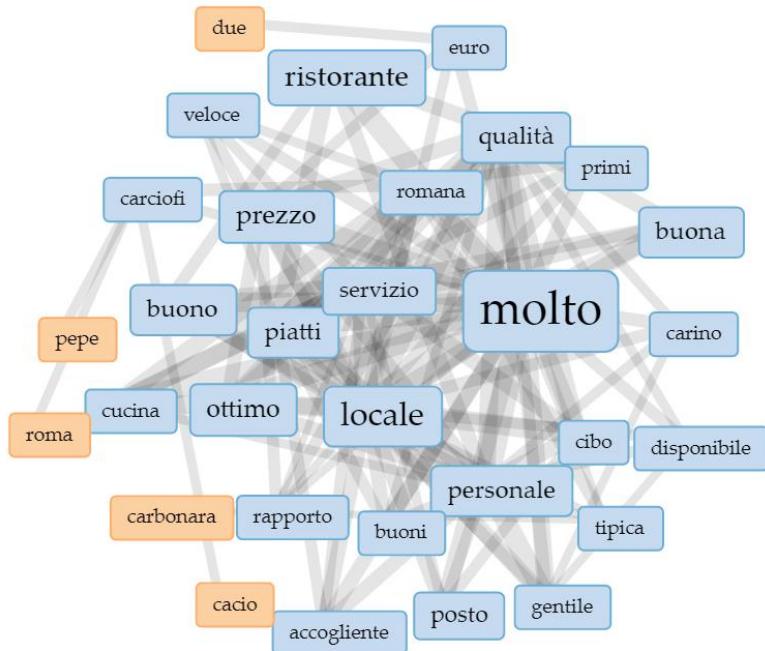


Figura: Analisi degli andamenti nel corpus

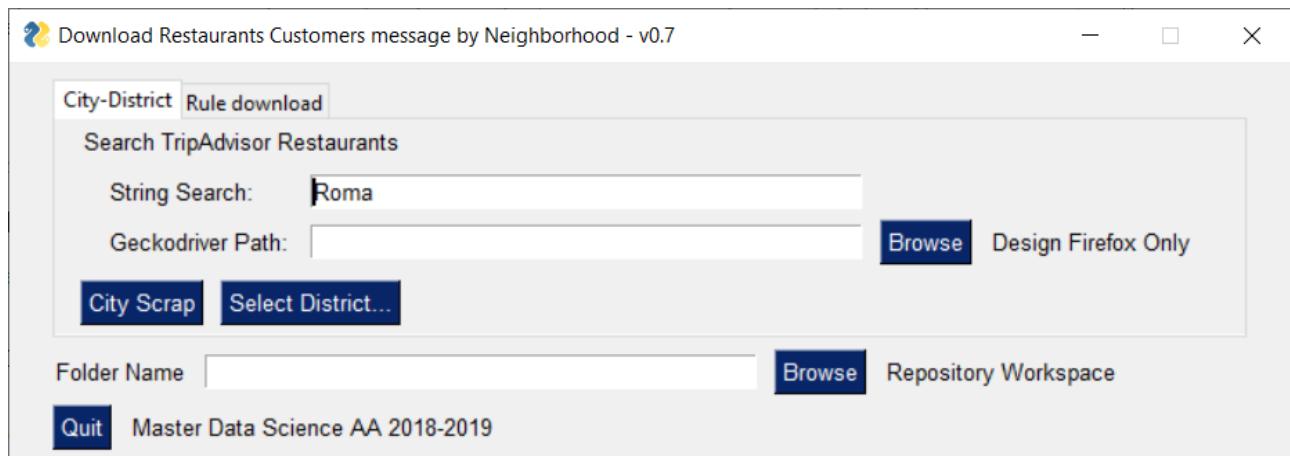
Appendice 1

Come accennato all'inizio, le operazioni di reperimento dei messaggi di testo sono state effettuate attraverso l'ausilio di strumenti informatici quali **python** con le librerie **selenium** e **BeautifulSoup**. Sono stati creati degli script che in modo automatico hanno effettuato l'accesso alle pagine web del sito di recensioni tripAdvisor per raccogliere i dati e salvarli in un file in formato csv. Le operazioni di scraping sono guidate attraverso un front end grafico che è stato realizzato con l'ausilio delle librerie grafiche **PySimpleGUI**, il quale ha permesso in pochi passi di realizzare delle form per i dati di input così da rendere l'operazione ripetibile attraverso pochissimi passi.

I pre-requisiti per il funzionamento del programma sono:

- aver prima caricato tutte le librerie necessarie presenti negli script,
- bisogna avere **firefox** installato sul pc,
- bisogna aver scaricato la versione corretta del driver **geckodriver** per firefox ed il sistema operativo ospitante.

E' possibile lanciare il programma attraverso lo script "tripA_main_gui.py", di seguito la prima schermata:

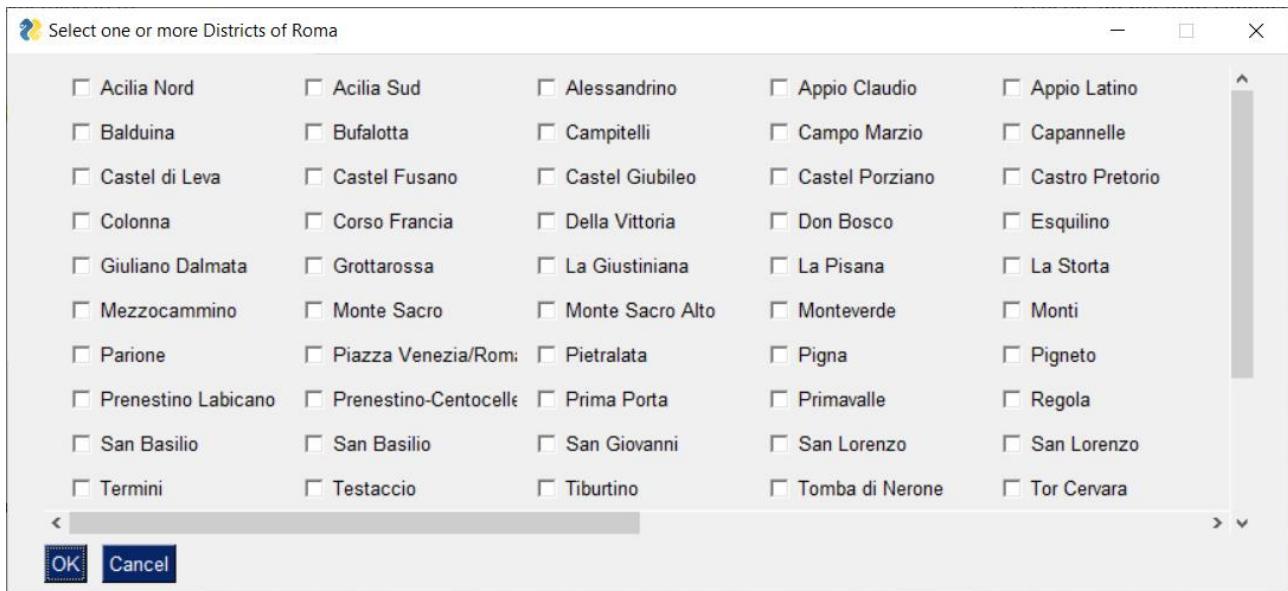


Nel primo tab "City-District" è possibile inserire i parametri di configurazione come i percorsi di input e di output nonché il nome della città su cui fare l'operazione di scraping:

- "String Search", deve contenere il nome della città per la ricerca su TripAdvisor,
- "Geckodriver Path", deve puntare alla cartella che contiene l'eseguibile,
- "Folder Name", deve contenere un percorso locale per l'output.

Una volta inserito il necessario, premendo il pulsante "City Scrap" si avvia lo scraping di tutti i quartieri della città selezionata. Nel nostro caso di Roma.

Finita l'operazione, viene generato come output un file in formato CSV contenente la lista dei quartieri di Roma e, premendo sul pulsante "Select District" tale lista viene caricata a sua volta all'interno della nuova finestra la quale permette la selezione di uno o più quartieri:



La finestra contiene la lista di tutti i quartieri trovati dalla ricerca precedente. Possono essere selezionati attraverso i **checkbox**, quindi premendo il pulsante “OK” si torna alla schermata precedente.

A questo punto, passando al secondo tab chiamato “Rule download” è possibile selezionare le due strategie singolarmente oppure lasciare al programma che vengano eseguite nello stesso processo, effettuando la doppia selezione.



Premendo sul pulsante “Execute”, il programma avvia il processo di scraping che, attraverso le librerie descritte precedentemente apre in piena autonomia le pagine di firefox per fare il recupero dei dati. Il risultato finale verrà salvato all’interno del percorso locale indicato in “Folder Name”, composto da due file in formato CSV; uno per strategia.

References

- Flament, C. (1962). L'analyse de similitude. Cahiers du centre de recherche opérationnelle, 4, pp. 63--97
- Lebart, L., Salem, A. (1994). Statistique textuelle. Paris: Dunod.
- Longhi, J. (2006). De intermittent du spectacle à intermittent: de la représentation à la nomination d'un objet du discours. Corela, 4 (2). URL: <http://corela.revues.org/457>.
- Longhi, J. (2008). Sens communs et dynamiques sémantiques : l'objet discursif intermittent. Langages, 170, pp. 109--124.
- Marchand, P., Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre/octobre 2011). Actes des 11èmes Journées internationales d'Analyse statistique des Données Textuelles. JADT, 2012, pp. 687--699.
- Pincemin, B. (2011). Sémantique interprétative et textométrie. Corpus, 10. URL: <http://corpus.revues.org/2121>.
- Reinert, M. (1998). Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste. Actes des 4èmes journées Internationales d'Analyse Statistiques des Données textuelles. URL : <http://lexicometrica.univparis3.fr/jadt/jadt1998/reinert.htm>.
- Reinert, M. (1999). Quelques interrogations à propos de l'objet d'une analyse de discours de type statistique et de la réponse « Alceste ». Langage et société, 90 (1), pp. 57--70.
- Corpus CoMeRe: <https://corpuscomere.wordpress.com>
- Iramuteq: www.iramuteq.org
- Ortolang: <https://www.ortolang.fr/market/home>
- Clement Levallois: lectures
- Stefania Spina, (2019). Fiumi di Parole