UNIVERSITÀ DI PISA
A.A. 2017/2018

RECIPE MARKET SURVEY

TWM REPORT

Bachini Francesco
Cinquini Martina
Giannini Miriam
Laid Laura
Morini Virginia

PREMESSA

Il progetto *'Recipe market survey'* nasce con lo scopo di svolgere un sondaggio sulle query di ricerca di ricette culinarie effettuate su Google in Italia rispondendo a due domande:

1. "Qual è il volume totale delle query di ricette di cucina in Italia?"
2. "Quali sono i principali siti nel SERP e per quali gruppi di query?"

Il punto di partenza della nostra indagine è la valutazione della domanda di mercato prendendo come riferimento una potenziale azienda intenzionata ad aprire un sito web nel settore investendo su un topic competitivo in base alle query più ricercate.

Si è ritenuto opportuno osservare i trend di ricerca delle keyword, suddivisi in seguito per categoria, per rilevare le strategie di analisi possibili, esaminare i dati tramite distribuzioni riguardanti le preferenze degli utenti ed identificare eventuali criticità di corrispondenza tra la ricerca dell'utente e l'offerta.

STUDIO DELLE KEYWORD

Gli utenti utilizzano la barra di ricerca per trovare una serie di contenuti correlati in modo rapido ed efficiente. Infatti, l'intento di ricerca parte direttamente dall'utente, il quale, mentre digita la query dichiara, più o meno esplicitamente, cosa sta cercando.

La comprensione della terminologia e della tipologia di ricerca che effettuano gli utenti potenzialmente interessati è di fondamentale importanza anche per individuare i termini di ricerca correlati ovvero keyword simili, che possono presentare un numero di ricerche più alto, meno concorrenza, oppure una combinazione di entrambi questi vantaggi. I risultati ottenuti potrebbero riguardare diversi contesti in cui le parole hanno significato. Nel nostro caso, ciò che si propone di analizzare è l'interesse relativo alle informazioni delle ricette di cucina cercate su Google in Italia.

Inizialmente, al fine di effettuare uno studio approfondito di questi aspetti, si è scelto di utilizzare Answer The Public[1] , uno strumento SEO che consente di elencare, attraverso grafici circolari, tutte le informazioni relative ai suggerimenti di Google. Questi grafici vengono organizzati attraverso una serie di combinazioni: domande, preposizioni, comparazioni, correlate e ordine alfabetico. Questo approccio discorsivo ci ha consentito di individuare alcune macro categorie sulle quali improntare la pianificazione e il successivo sviluppo della nostra indagine. Infatti, ciò che si evidenzia effettuando la ricerca 'ricette italiane' nel tool, è la forte presenza di query basate sulla portata (es. 'primi piatti'), sulla categoria alimentare (es. 'carne'), sulla periodicità (es. 'natale') e sulle intolleranze ('senza glutine').

SUDDIVISIONE IN CATEGORIE

Seguendo questa linea, le categorie prese in considerazione sono:

1. Portata
2. Stagionalità
3. Festività
4. Intolleranze e scelte alimentari

[1] www.answerthepublic.com

Ognuna di queste è stata ampliata in base ai risultati più significativi in ulteriori sottocategorie, in particolare:

Portata	Stagionalità	Festività	Intolleranze e scelte alimentari
Antipasti	Primavera	Natale	Senza glutine
Primi piatti	Estate	Pasqua	Senza Lattosio
Secondi piatti	Autunno	Capodanno	Vegetariani e Vegani
Contorni	Inverno	Feste	
Dessert			

UBERSUGGEST

Ubersuggest[2] è un tool gratuito per la ricerca ottimizzata grazie al quale è possibile trovare tutte le keyword digitate dopo una determinata parola chiave avendo una panoramica delle voci correlate a quella ricerca, in ordine alfabetico. Per ogni keyword è possibile visualizzare (vedi Figura 2) il volume di ricerca, il CPC[3] e la competition. Inoltre, vi è la possibilità di filtrare i risultati (vedi Figura 3) e selezionare lo strumento con cui vengono effettuati i suggerimenti.

Con l'ausilio di Ubersuggest, ogni sottocategoria è stata quindi inserita come keyword di partenza per generare le query relative; si è effettuato un filtraggio dei termini ambigui e i risultati pertinenti sono stati scaricato in file CSV.

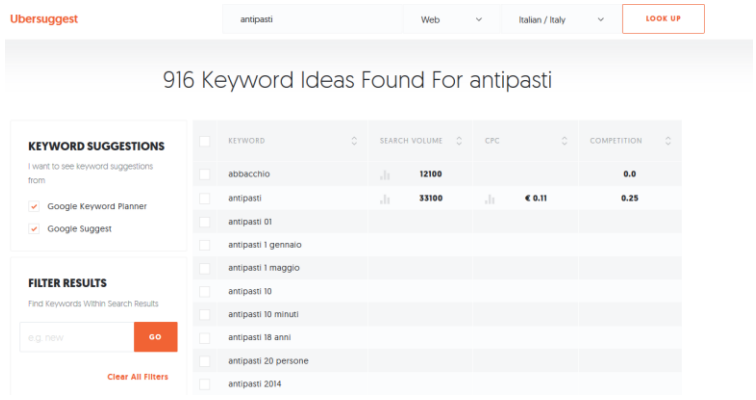

Figura 1: Keyword trovate per la categoria antipasti in Ubersuggest

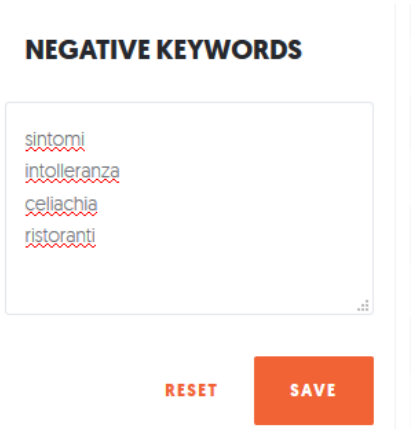

Figura 2: Filtro parole negative di Ubersuggest

L'unione dei file riguardanti ogni ricerca ha consentito la generazione di un dataset costituito da 9219 records. Tuttavia, il file conteneva duplicati e keyword non correlate alla nostra analisi. Ad esempio, per "ricette Capodanno", erano presenti record come "Hotel", "pernottamenti", "ristoranti", "capodanno idee", "case". Una motivazione è che il tool dispone di strumenti limitati per il filtraggio che non hanno consentito di eseguirlo completamente come, ad esempio, vi è la possibilità di inserire solamente 5 parole nelle *negative words.*

[2] www.ubersuggest.com
[3] costo per click

Inoltre, nei file erano presenti elementi che indicavano brand di prodotti, attrezzi da cucina, siti di cucina, chef, blog e programmi televisivi che, in base alle specifiche del progetto, non dovevano essere considerati.
Per questo motivo, si è eseguita una scansione manuale del file CSV in quanto non vi era la possibilità di conoscere a priori il set di parole da eliminare. Da questa scansione è stato creato un file di testo contenente una lista di keyword non corrette (es. 'Benedetta Parodi', 'giallo zafferano', 'misya') utilizzata per filtrare nuovamente il file CSV. Il dataset risultante conteneva 8609 records organizzati in base alle keyword.
Nonostante ciò, il dataset non aveva la conformazione prevista rispetto alla dimensione e alla qualità dei termini di ricerca. A causa di queste criticità, si è deciso di utilizzare Ubersuggest solamente per l'individuazione, nella maniera più accurata possibile, delle parole da filtrare e di non utilizzarlo come dataset definitivo per l'analisi. Perciò, si è scelto di basare la nostra indagine sul dataset di Google Trend Estimates fornitoci durante le prime fasi del progetto.

IL DATASET GOOGLE TRENDS ESTIMATES

Il file è costituito da 117721 record definiti da due attributi: query di ricette di cucina e volume di ricerca. Nella fase iniziale, il dataset è stato manipolato applicando il filtro con le parole non pertinenti che ha prodotto una riduzione di 3864 record. Pertanto, il totale dei record filtrato è 113858. Prendendo in considerazione tale dataset come quello di riferimento, sono stati seguiti i seguenti step:

1. Individuazione di set di parole necessari a effettuare la categorizzazione
2. Filtraggio per sottocategorie
3. Gestione delle ambiguità
4. Analisi delle distribuzioni
5. Considerazioni finali del primo task

1.Individuazione di set di parole necessari a effettuare la categorizzazione

Per evitare di dover scansionare il dataset manualmente come in precedenza, al fine di individuare i set di parole necessari a effettuare la categorizzazione, si è scelto di utilizzare Wikipedia.

In Wikipedia sono presenti alcuni portali aperti, suddivisi per area tematica, che hanno lo scopo di fornire un supporto per la navigazione ragionata delle voci dell'enciclopedia.[4] Nel nostro caso, si è preso in considerazione il portale della cucina che contiene la suddivisione nelle sottocategorie individuate. Per ognuna di esse, è stato creato un CSV, in seguito convertito in un file di testo, contenente le relative keyword.
Tale estrazione per le categorie 'Portata' e 'Scelte alimentari' è stata effettuata con lo strumento Data Miner[5], un'estensione di Google Chrome che permette di eseguire lo scraping dei dati da pagine web a file CSV.
Si è notato che alcuni elementi presenti nel file indicavano query troppo specifiche ad esempio tra i record di primi piatti vi erano 'Gnocchi alla romana', 'Gnocchi alla sorrentina', 'Gnocchi di malga', 'Gnocchi di zucca', 'Gnocchi ricci'.

[4] https://it.wikipedia.org/wiki/Portale:Portali
[5] https://data-miner.io/

Il problema era una perdita di generalità poiché il nostro scopo era effettuare un filtraggio basato su elementi legati alla tipologia di ricerca e non di una in particolare. Dunque, si è effettuato un troncamento della stringa selezionando i caratteri della prima parola sufficienti a individuare tutte le possibili query correlate senza essere vincolati dalla morfologia della parola stessa (singolare, plurale, diminutivi ecc.). Riprendendo il precedente esempio, abbiamo quindi considerato solo 'Gnocc'.

In relazione alle categorie Festività e Stagionalità, non è stato possibile affidarsi al portale di Wikipedia in quanto non erano presenti aree tematiche correlate. Quindi, per selezionare le relative query si sono utilizzati termini che inequivocabilmente appartengono a quella categoria in modo da individuare solo risultati *true positive* che soddisfano la *precision*[6].

2.Filtraggio per sottocategorie

I file contenenti le liste di keyword creati da Wikipedia sono stati utilizzati per generare le sottocategorie in base alla corrispondenza tra gli elementi presenti nelle diverse liste e le query del dataset.

3.Gestione delle ambiguità

Le ambiguità presenti nelle sottocategorie risultanti sono state gestite nei seguenti modi:

- Con le espressioni regolari sono stati rimossi termini specifici contenuti nel dataset risultante poiché il troncamento della parola utilizzata per il filtraggio li aveva inclusi. Ad esempio, per 'Primi piatti', tra gli elementi della lista delle keyword è presente la parola 'pasta' che ha considerato query appartenenti ad altre categorie. Di seguito, un esempio:

```
primoD[primoD['Keyword'].str.contains('sfoglia|frolla|fillo|brise|di zucchero') == False]
```

- Rimuovendo la lista delle keyword di una categoria dai risultati di quelle restanti. Ad esempio, le keyword di primi piatti sono state eliminate dai dataset degli antipasti, dei secondi piatti e dei contorni.

```
for primo in primi:
   secondiD = secondiD[secondiD['Keyword'].str.contains(primo) == False
```

- Rimuovendo le query che possono rientrare in più categorie.
  Ad esempio, il morfema 'insalat' ('insalata') contenuto nel file dei vegetariani non è necessariamente detto che indichi un piatto vegetariano, infatti una possibile query potrebbe essere 'insalata di pollo' che quindi risulterebbe non pertinente.
  Tenendo conto che, a causa della vastità dell'argomento, non è possibile individuare una classificazione netta, si è deciso di mantenere i duplicati inter categorie durante la fase di analisi e di rimuoverli solamente dopo aver generato il dataset completo.

[6] In un processo di classificazione statistica, la precisione per una classe è il rapporto tra il numero di veri positivi e il numero totale di elementi etichettati come appartenenti alla classe.
(https://it.wikipedia.org/wiki/Precisione_e_recupero)

4.a. Statistiche descrittive

Si sono svolte le seguenti statistiche descrittive sui risultati ottenuti dal filtraggio dei dati del dataset: deviazione standard, media e somma dei volumi e il conteggio delle keyword. I risultati sono stati rappresentati con grafici considerando nell'asse delle ascisse le sottocategorie e in quello delle ordinate i volumi. Le distribuzioni ottenute non sono particolarmente significative a causa dell'elevato numero di query a volume zero nel dataset che alterano gli andamenti.

Portate

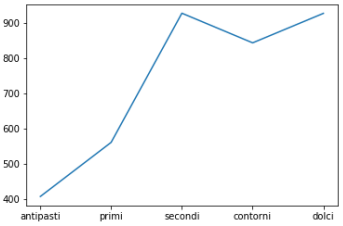

Figura 3: Standard Deviation

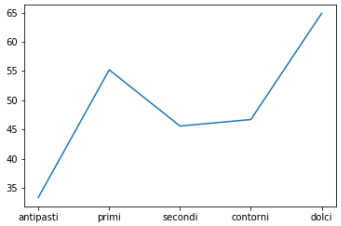

Figura 4: Media

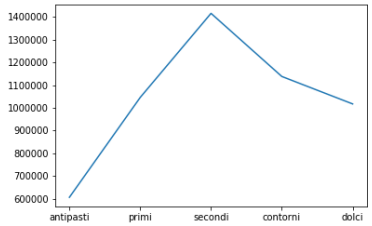

Figura 5: Somma dei volumi

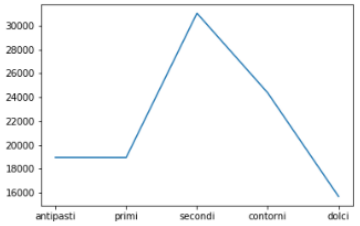

Figura 6: Conteggio delle keyword

Stagionalità

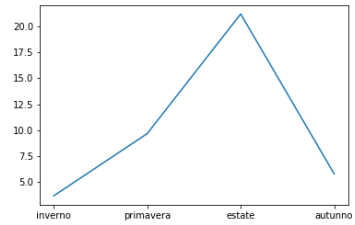

Figura 7: Standard Deviation

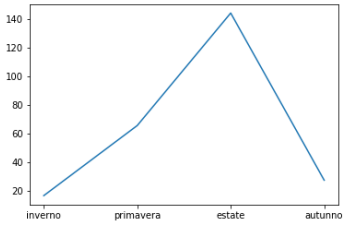

Figura 8: Media

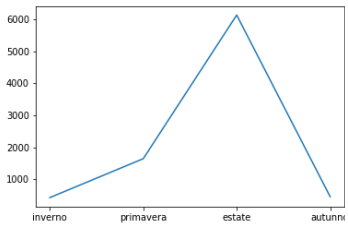

Figura 9: Somma dei volumi

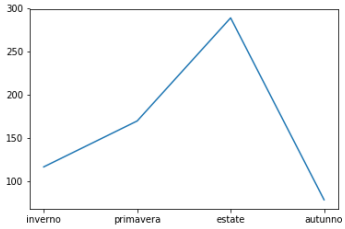

Figura 10: Conteggio delle keyword

Festività

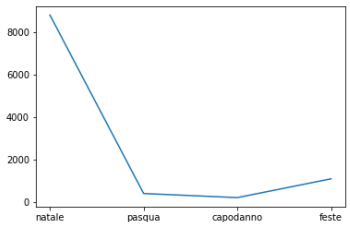

Figura 11: Standard Deviation

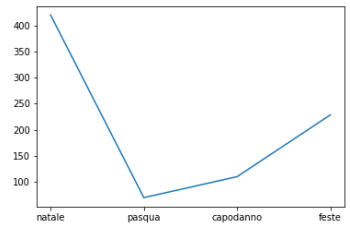

Figura 12: Media

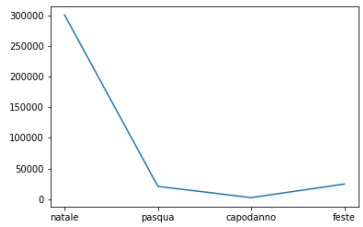

Figura 13: Somma dei volumi

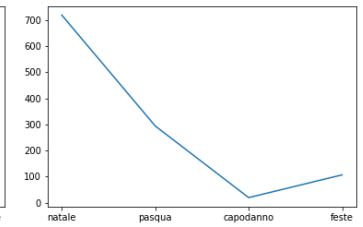

Figura 14: Conteggio delle keyword

Intolleranze e scelte alimentari

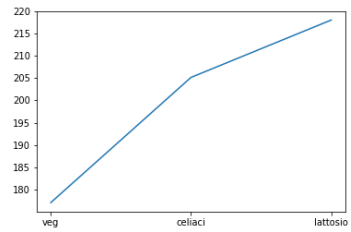

Figura 15: Standard Deviation

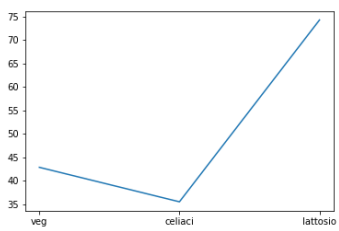

Figura 16: Media

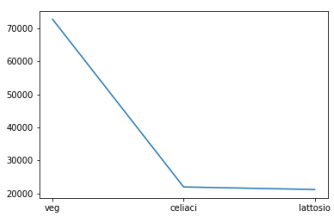

Figura 17: Somma dei volumi

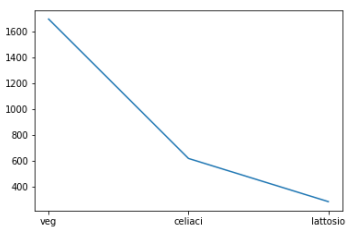

Figura 18: Conteggio delle parole

Nonostante ciò, possiamo osservare alcuni aspetti.

Ad esempio, in Portata, ‘secondi piatti’ ha il valore più alto della deviazione standard, ma una media inferiore a tutte le restanti sottocategorie tranne ‘antipasti’. Ciò significa che ci sono pochi valori vicini alla media mentre la maggior parte se ne discostano. Il valore medio più alto risulta invece essere quello dei dolci, ciò sta a indicare che la sottocategoria è composta da query ricercate con una distribuzione piuttosto regolare.

Inoltre, in Festività, ‘Natale’ ha valori molto elevati in ogni statistica, indice del fatto che le keywords di questa sottocategoria, nonostante siano relative ha un periodo limitato dell’anno, sono le più ricercate rispetto alle altre festività. Anche nella categoria Stagionalità si verifica una distribuzione simile, in particolare per ‘Estate’.

4.b. Parole e bigrammi frequenti

Al fine di approfondire la nostra indagine, si è pensato che, oltre a fornire al potenziale investitore statistiche descrittive dei dati, fosse interessante dotarlo di un set di parole e di bigrammi più frequenti di ogni sottocategoria. Nella creazione di una piattaforma web è infatti fondamentale avere a disposizione una serie di parole, 'keyword', frequenti su cui basarsi per diversi motivi:

- Il tema trattato in ogni pagina web dovrebbe ricondursi direttamente ad una parola o frase chiave.
- Le parole chiave aiutano i potenziali clienti e i motori di ricerca a capire lo scopo della pagina. Quando un motore di ricerca indicizza le pagine del sito, esso analizza le keyword per determinare qual è la funzione delle pagine.

Dopo aver generato le liste di parole e di bigrammi più frequenti per ogni sottocategoria si è deciso di visualizzarle tramite Wordcloud, una rappresentazione visiva di parole chiave con la peculiare caratteristica di attribuire un font di dimensioni più grandi alle parole più frequenti. Si tratta quindi di una lista pesata.

Di seguito alcuni esempi del lavoro svolto:

Figura 19: Parole frequenti per secondi piatti

Figura 20: Bigrammi frequenti per secondi piatti

Figura 21: Parole frequenti per celiaci

Figura 22: Bigrammi frequenti per celiaci

5.Considerazioni finali del primo task

L'indagine si poneva l'obiettivo di trovare il volume di query riguardanti le ricette culinarie ricercate su Google in Italia. Basandoci sulla categorizzazione del dataset effettuata, si è voluto analizzare il volume delle query sotto diversi punti di vista in modo da interpretare al meglio il risultato finale.

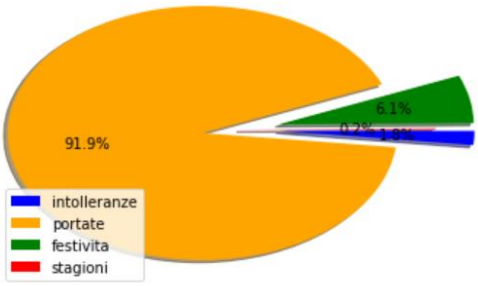

Figura 23: Volumi totali di ricerca delle query suddivisi per categorie

In primo luogo, si sono visualizzate le percentuali dei volumi totali di ogni categoria tramite un pie-chart. Da tale grafico, è risultata evidente la predominanza delle query appartenenti a 'Portata' con una percentuale del 91.9. Tale risultato non sorprende affatto in quanto quest'ultima è una categoria decisamente meno settoriale delle altre.

Alla luce di ciò, si è pensato di estrarre, per ogni sottocategoria, la query con il volume più alto per evidenziare se, nonostante la netta maggioranza in numero e in volume delle sottocagorie di 'Portata', altre presentassero query molto ricercate.

Le cinque query più frequenti (vedi Tabella 1) sono, come previsto, quelle che appartengono alla categoria 'Portata'. Risulta però inaspettato notare che la query 'panettone' facente parte di una sottocategoria ricercata solitamente in un ristretto periodo dell'anno abbia un volume pari a circa la metà di una ricetta non stagionale/periodica come 'Insalata'.

Tabella 1: Query con il volume più alto in ordine decrescete

Sottocategoria	Volume	Query
Secondi	134515.847	Pizza
Contorni	92956.453	Patate
Dessert	85207.296	Cioccolato
Primi Piatti	48103.284	Risotto
Antipasti	29891.430	Insalata
Natale	12180.644	Panettone
Feste	7850.912	Torta compleanno
Pasqua	4986.877	Uova pasqua
Intolleranti al glutine	4139.368	Grano saraceno
Vegani/Vegetariani	4101.549	Torta senza uova
Estate	2213.604	Ricette estive
Intolleranti al lattosio	1841.809	Biscotti senza burro
Primavera	793.191	Involtini primavera
Capodanno	732.137	Antipasti capodanno
Autunno	206.996	Ricette autunnali
Inverno	108.628	Insalata invernale

- Portate
- Stagionalità
- Festività
- Intolleranze/ scelte alimentari

Dunque, il dataset di partenza conteneva 113858 record. Di questi, ne sono stati categorizzati 78840, circa il 70% del totale. Poiché alcune query risultavano comuni a più sottocategorie, questo valore è stato ottenuto considerandone solo una occorrenza. (vedi paragrafo 'Gestione delle ambiguità').

Concludendo, in risposta alla domanda 'Qual è il volume di query riguardanti le ricette culinarie ricercate su Google in Italia?', in base alla nostra analisi, si rileva che il volume totale è 4616428.

SERP

Per rispondere alla seconda domanda del progetto, si è voluto individuare quali fossero i siti web più cercati per tipologia di portata utilizzando i file generati nella task precedente. Si è scelto di improntare la nostra analisi solo su questa categoria tenendo in considerazione il fatto che il volume di ricerca totale è il 97% del volume totale relativo alle categorizzazioni.

L'indagine si è sviluppata nelle seguenti fasi:

1. Analisi del file 'allgraphs.csv' contenente le keyword, il contatore dei siti web per ciascuna query e il dominio dei siti.
2. Assegnamento di un'etichetta identificativa della categoria di appartenenza (es. 'primo piatti', 'secondi piatti') ad ognuna delle keyword presenti.
   Per effettuare tale categorizzazione, si sono presi i file di testo utilizzati per il filtraggio e sono stati eseguiti in tale ordine *primi, antipasti, secondo, contorno, dolce* al fine di evitare ambiguità tra le keyword che potevano essere presenti in più categorie.

	Keyword	Sitoweb	Categoria
1618	spaghetti	ricette.giallozafferano.it	primo
1638	spaghetti agli asparagi e crema di carote	lastampa.it	primo
1658	spaghetti aglio	allrecipes.com	primo

Figura 24: Categorizzazione dei record

3. Eliminazione dei duplicati.
   È stato necessario eliminare tutti i record che contenevano gli stessi elementi.

	Keyword	Sitoweb
0	soufflè di funghi	ricette.giallozafferano.it
1	soufflè di funghi	cookaround.com
2	soufflè di funghi	blog.giallozafferano.it
3	soufflè di funghi	ricette.giallozafferano.it
4	soufflè di funghi	asiagofood.it

Figura 25: Eliminazione dei duplicati

4. Raggruppamento delle keyword relative ad ogni sottocategoria ordinate in base alla loro frequenza nei siti.
5. Calcolo della dimensione totale di ogni categoria e calcolo della probabilità in percentuale di trovare una determinata ricetta su ogni sito.

Concludendo, in risposta alla domanda 'Quali sono i principali siti nel SERP e per quali gruppi di query?' si sono evidenziati i primi tre risultati nella seguente tabella. Tra i maggiori competitor nel mercato italiano delle ricette si rileva, per ciascuna delle sottocategorie, al primo posto il sito *giallozafferano.it*, al secondo posto *youtube.com* e infine al terzo *lacucinaitaliana.it* e *cookaround.com*

Tabella 2: Risultati SERP

Ranking Sito	Antipasti	Primi	Secondi	Contorni	Dolci
1°	giallozafferano.it (44%)	giallozafferano.it (51%)	giallozafferano.it (37%)	giallozafferano.it (41%)	giallozafferano.it (41%)
2°	youtube.com (35%)	youtube.com (43%)	youtube.com (31%)	youtube.com (34%)	youtube.com (37%)
3°	lacucinaitaliana.it (29%)	cookaround.com (37%)	lacucinaitaliana.it cookaround.com (24%)	lacucinaitaliana.it (26%)	lacucinaitaliana.it (22%)