Multi-class categorization based on cluster analysis and TFIDF

Sergio Bolasco[1], Pasquale Pavone[1]

[1]DSGLSSAR - Facoltà di Economia - SAPIENZA Università di Roma

I-00193 Roma – Italie

Abstract

In the logic of automatic text classification, this study presents a procedure based on the analysis of clusters and the TFIDF index to generate a multi-class categorization of documents. Starting from a set of appropriately selected terms, we proceed to a non-supervised disjunctive classification into groups. The set of terms characterizing each cluster identifies the thematic dictionary of the group. For each document, we calculate the TFIDF associated with each dictionary. The multi-class categorization is obtained on the basis of the higher values of TFIDF for each document. An example of this procedure applied to a corpus of some one-hundred reviews of restaurants is proposed.

Riassunto

In una logica di classificazione automatica di testi, questo lavoro propone una procedura basata sull'utilizzo in sequenza della cluster analysis e del TFIDF per generare una categorizzazione multi-classe di documenti. Sulla base di termini opportunamente selezionati si procede ad una classificazione non supervisionata in gruppi disgiunti, e come tale univoca. Il set di termini caratterizzanti ciascun cluster individua il dizionario tematico del gruppo. Per ciascun documento si calcolano i TFIDF associati a ciascun dizionario. La categorizzazione multi-classe è ottenuta sulla base dei valori più alti dei TFIDF per ciascun documento. Si propone un esempio della procedura sul corpus di trecento recensioni di ristoranti.

Keywords: automatic classification, multi-class categorization, clustering, TF IDF, gastronomic lexicon.

1. Introduction [1]

A basic problem in the automatic categorization of documents is a correct attribution of one or more themes to identify the content of the text. In general, in the absence of categories that have been predefined by the researcher, the documents are grouped according to their similarity, thus identifying their predominant theme.

In most of the procedures in (Bolasco et al., 2005), a classification of this type leads to disjunctive classes, even though the themes often have semantic elements or characteristics in common (for example, politics shares many terms with economics, food with agricultural products and so forth). The first situation leads to univocal classification, the second to fuzzy classification (Zadeh, 1977; Ricolfi, 1992). The purpose of this paper is to see how we can exploit the first type of classification to produce the second and, in particular, how to move from a non-supervised clustering to a supervised multi-class categorization.

[1] This paper has been financed by the funds from MIUR Facoltà 2005 – C26F059955 and is the result of a collaboration between the two authors. Paragraphs 1, 2.1, 3.1, 4 were written by Sergio Bolasco and 2, 2.2, 3 and 3.2 by Pasquale Pavone.

In information retrieval the TFIDF index (Salton, 1989) is often used to measure the weight of the words in a given source. This weight is used as an indicator of the importance of terms, as for example on the web, to measure the relevance of the contents of a document in relation to a specific query, which in most cases consists in a simple list or combination of words.

A classification of documents on the basis of the words contained in them filters the terms that are more sensitive or characteristic of that class. These characteristic elements can be measured by indices of homogeneity and selectivity and statistical tests in cluster analysis procedures (Bolasco, 1999). Together these characteristic words indicate the theme or category that will give its name to the class and each theme is defined by a list of terms (thematic dictionary).

Furthermore, these applications using have clearly shown (Lebart , 2003) that certain combinations of words create contexts with specific meanings. It is also evident that different combinations of some of these words produce different meanings (themes), that is, single words with the same meaning that are contextualised in different ways express different concepts. This also leads to the disambiguation of the occurrences of certain . It is therefore useful to find which combinations of words generate which themes and to what extent a document can be attributed to one theme rather than another, according to the logic of . As we shall see below, an appropriate procedure that uses both a cluster analysis and the TFIDF index facilitates the move from a univocal classification to a multi-class or multiple categorization.

2. The strategy of classification

The proposed procedure shows how it is possible to move from a non-supervised univocal classification to a supervised, but not exclusive classification. This will be achieved by passing through three phases: (IE), and . Here a very important role will be played by the TFIDF index, which is used as a selection factor in the first and third phases. It should be stressed that its use requires a corpus of at least several hundred, if not thousand, fragments of text or documents.

Before proceeding to the different stages of the strategy, it is necessary to carry out a of the text so that we can draw up a list of sensitive terms, which can then potentially become relevant terms. The pre-processing is made up of the following steps: lexical analysis, the elimination of stop-words and the stemming of some particularly interesting forms for the analysis.

The lexical analysis has two purposes: to recognize the grammatical category of words and to make an analysis of the repeated segments to identify and lexicalize some idiomatic expressions in the text. This means some polysemous forms will be identified (for example the word < >: (wine list) (dessert menu) (credit card) (à la carte).

The elimination of stop-words makes it possible to exclude those forms that are of little relevance to the content. Normally these will be conjunctions, articles and some prepositions. In the following application to the lexicon of wine and gastronomy, for example, it is interesting to note that the articulated prepositions < > are significant in distinguishing tastes, smells or ways of cooking (: with herbs, in butter, à la Bolognaise) and therefore should not be eliminated.

The stemming proves a useful way of reducing the number of elements to be analysed so that different forms will be considered as a single type, that is, they will have different morphological variations, but similar semantic meanings. The lexical morpheme, as the root of a word, expresses a concept regardless of its grammatical category and/or its gender/number of the inflected form. For example, the type < *> [equivalent to the English simpl*] includes , , , to represent the same concept.

After the lexical pre-processing has made a selection of the "sensitive" words to be considered as significant, a textual analysis will establish the categories of the documents.

This is done in the first phase of the strategy by extracting information through the TFIDF index (hereafter TI), to obtain a set of keywords, which will be used for the description of the content of the documents making up the corpus. As is known, the TI index (for details see Salton, 1989) assigns a weight to each form of the words according to the relative importance of that form in a document, but also to its ability to discriminate in relation to the entire corpus. In this way the TI attaches a weight to each form based on its frequency in individual documents and its distribution within the collection of documents. By reordering the terms of a decreasing TI value in the vocabulary of the corpus and establishing a threshold value[2] below which a word will be excluded, we can obtain a list of the forms that are relevant to the subsets of the documents.

2.1. The identification of thematic dictionaries

The second phase of the strategy groups the documents in clusters according to the similarity of the distribution of the relevant terms selected by TI (keywords). This phase of based purely on is a non-supervised classification of documents, which will nevertheless reflect the similarity of the documents at a semantic level. Conceptual homogeneity will indicate the theme or semantic domain prevailing in that group of documents, which can be summarized as a category that was not pre-defined, but will be univocal.

We will use a matrix of Documents × Keywords, which groups the various documents to be classified with the occurrences of their relevant terms. By submitting this matrix to the classic chain, "simple correspondence analysis + cluster analysis", we will get a classification in K groups of documents. The semantic field, which is an expression of this similarity, can be seen through the proximity of the corresponding terms on a factorial map. The K groups of documents obtained in this way are disjunctive classes, but the corresponding K lists of words (dictionaries defining the "theme" or category of the class) characterizing the groups are not totally different from each other, as common terms can often be found.

The presence of the same term in different dictionaries can lead to the disambiguation of its possible meanings. If a word has a different meaning in a different context, it will appear in different dictionaries. For example, in wine and food lexicon, the word can be present both in a dictionary of first courses: (fresh pasta) (egg pasta) ; and also in a dictionary of sweets: (marzipan) (puff pastry). However not all identical terms in different dictionaries are necessarily polysemous. They can be monosemous words taking on

[2] This value is chosen in order to consider a sufficiently broad group of words (about a few hundred) to guarantee the variability of the phenomenon.

a different meaning in different areas. For example, the keyword < >[3] exists in a thematic dictionary on the "poor" cuisine of a , but it is also found in a thematic dictionary on gastronomic excellence ( ), where the 'simplicity' of the dishes is a result of their sophistication and culinary art. This last point is part of the general principle that the greater the experience and knowledge in a sector, whether it be science or art or anything else, the greater the tendency to "remove" artefacts in its presentation and rely on the bare essentials and therefore simplify. An analysis of 2,228 restaurant reviews published by Gambero Rosso Editore shows that the concept of * appears in 21% of the reviews of , in 11% of haute cuisine restaurants and around 7% of the reviews of restaurants with an average scoring.

On the other hand, the theme which classifies the cluster depends on the terms that are "synonyms" in relation to the meaning of the class in the thematic dictionary. The word list of each cluster, in decreasing order according to the test-value which will define the group, is important in order to capture the weight of the word in that dictionary. The potential degree of conceptual exclusiveness of a term, for example, with a low selectivity value[4] reflects the polysemy of that term in the corpus.

2.2. Multi-class categorization

The third phase of the strategy deals with the re-classification of the documents. For this purpose, let us consider the K thematic dictionaries resulting from the cluster analysis as queries for which the TI has to be calculated for the documents in the corpus. These TI are the elements in the matrix Documents × Query. For each row of the matrix, the highest value of TI (hereafter TI_{MAX}) will determine the allocation (re-classification) of the document to the category associated with the thematic dictionary of the query that has produced that value, on the basis of the greatest relevance of the query to that document. By the same logic, where there is another significantly high value generated by another thematic query among the remaining TI for the document under consideration, an additional category can be attributed to the document, even if it is less relevant. The criterion for establishing a threshold for the allocation of a further thematic category to avoid excessive classification could be as follows:

- the differences between the TI_{MAX} ($^{+}$) and -th TI of each K query [$diff_{k+k(i)}$] are calculated for each document. In this way the difference corresponding to the -th query on the TI_{MAX} has a value equal to zero, while, for the other queries for the same document, the difference will have a higher or lower value depending on how close their TI is to TI_{MAX};

- the lowest value among the K-1 differences other than zero, [$diff_{min}(i)$] is selected for the i-th document;

- the average of these smallest differences for all the N documents, [$diff_{med}$] is calculated;

- only the values of TI with below average difference values, [$diff_{k+k(i)} < diff_{med}$] will be filtered in the matrix;

- the values filtered in this way assign other thematic categories to each document on the basis of the corresponding query. A document that does not have filtered values will have a single categorization corresponding to the query with the value of TI_{MAX}.

---

[3] Better still is the similar concept of the stem <semplic*> which presents: semplice/i/ità/emente.

[4] The selectivity of word in group is measured by n / n , where n is the frequency of the word in the group and n is the frequency of the word in the corpus.

A different criterion - for the selection of the TI that are useful for multi-categorization - might be to consider one by one the decreasing TI values of the queries for each document and attribute the category corresponding to the highest TI among those remaining. The criterion for the second highest TI value (the one after TIMAX) favours fragments with a second high TI value in absolute terms, thus penalizing fragments with low TI values.

The criterion of the smallest differences penalizes the fragments that have a very high TIMAX and a second highest TI value, but not so much as to compensate the difference with the first value. On the contrary, it favours the fragments that have a low TIMAX and a second TI value close to the TIMAX, but still low.

It is clear that other criteria are possible for the selection of assignment values, but we will consider only the criterion of the smallest differences.

3. An application of the procedure

This strategy is applied to a corpus of 276 reviews, taken from the 2004 Restaurant Guide of the "Gambero Rosso" for the regions of Piedmont and Sicily because they provide very different examples of gastronomy[5]. The reviews consist in brief descriptions of the restaurants with a list of the dishes and wines on offer, the characteristics and atmosphere of the place, the type of service being offered. The aim of the application is to obtain an automatic classification of these reviews without a predefined categorization (non-supervised thematic classes), but with the possibility of recognizing a number of characteristics that each restaurant may have, and thus re-classifying it in a supervised and, if possible, a multi-class manner. The initial exploratory phase of clustering will recognise the range of themes on catering through a classification of disjunctive groups (categories). Then, taking this range as a starting point, it can be seen to what extent a restaurant "responds" to each of these categories. This is done by converting each of these categories into a query and measuring the degree of adherence of the restaurant review to the category through a corresponding TI.

During the pre-processing of the text and after the grammatical tagging of the inflected forms ( ) in the corpus using the software TreeTagger[6], repeated segments were identified in order to select and lexicalize some of the significant idiomatic phrases: here 31 compound words and fixed collocations were found. This led to the disambiguation of some sensitive terms (for example: (wine list) (cheese trolley) (butter and sage) (white chocolate), (dark chocoate), (balsamic vinegar), (clams), , , and so on. The stop words were then identified. Given the peculiarities of a review, it was decided to remove all the words that were not grammatically categorized as nouns or adjectives. The result of the pre-processing of the text led to 3,369 forms being considered as sensitive forms, either simple or complex, nouns or adjectives, which could then be selected as keywords in the documents analyzed.

[5] This strategy has been applied by means of Taltac (www.taltac.it) and Spad software.

[6] TreeTagger marks the words in the texts with their appropriate grammatical category and lemma. It has been developed within the TC project (http://www.ims.unistuttgart.de/projekte/tc) at the Institute for Computational Linguistics of the University of Stuttgart.

3.1. Automatic classification for the definition of themes

In the second phase of the application a cluster analysis (starting from 3 factors) was carried out on this matrix, which ranked the restaurants in 6 groups. The words that characterize each group produced a dictionary of terms. The following themes or categories summarized below were assigned to the six categories:

CLU 1 - **traditional Piedmontese cuisine**: (agnolotti butter and sage sauce) (sauce) (game) alle (veal with herbs) (choice of boiled meat);

CLU 2 - **trattoria**: (family-run business), (prices) (set menu), (tomato) (pork) (sausage) (barbecued, grilled or roast meat) (salads);

CLU 3 - **location/region, atmosphere**: (bonus for the wonderful enchanting location) (lake) (vineyards) (good, polite service) (ingredients) (choice) (good excellent cheeses) (aromas) (carp) ;

CLU 4 - **haute cuisine**: (equipment) (appetizer), (stuffed pigeon) di (farmhouse) d' (breast of duck) (sturgeon) (sauce) (cream) (butter) (honey) (mille feuille) (pastries) (coffee) (ice-cream) (white chocolate) (dark chocolate);

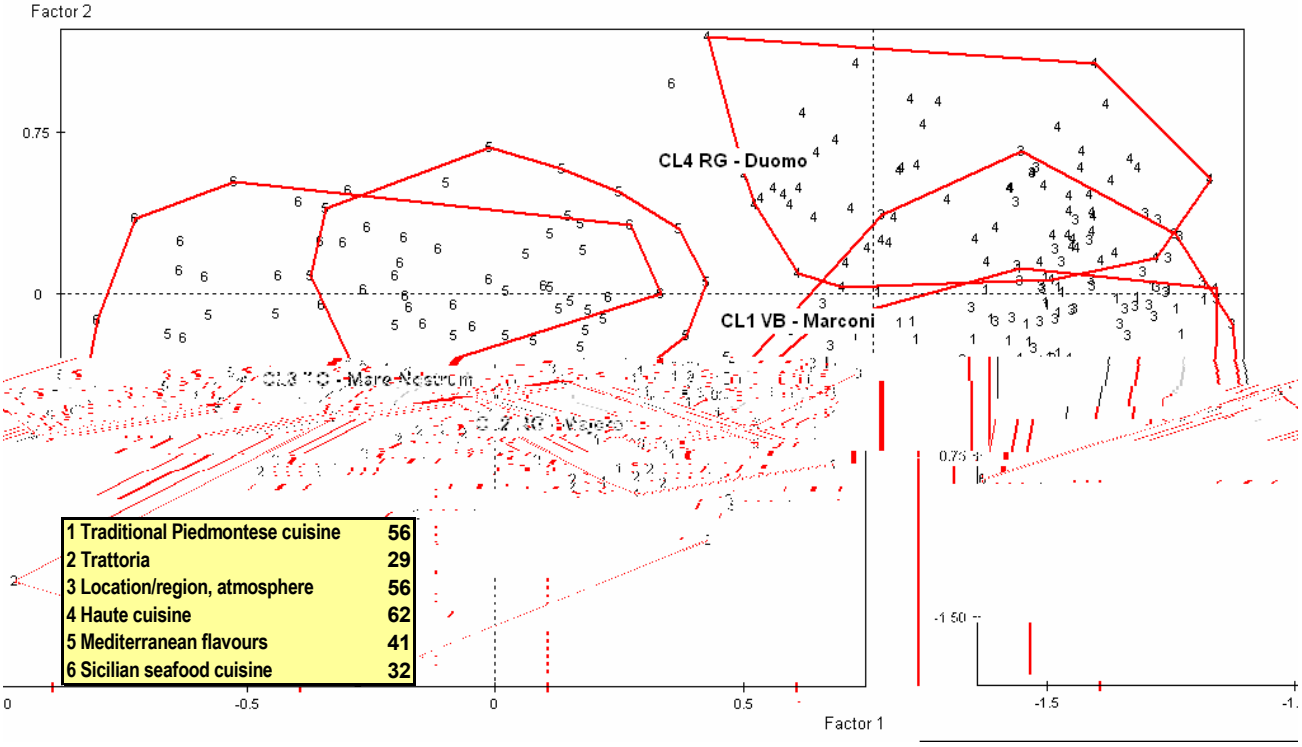
CLU 5 - **Mediterranean flavours**: (outdoor seating) (island) (swordfish) (squid) (octopus) (almonds) (pistachios) (citrus fruits) (capers) (aubergine) (wild fennel) (oils);

CLU 6 - **Sicilian seafood cuisine**: (courteous) (efficient) (terrace) (ingredients) (swordfish) (rock salmon) (tuna) (stone bass) (lobster) (shrimp) di (seafood), (mussels) (clams) (grilled) (fried) (sorbet) (lemon).

The factorial map shows, on the one hand, the contrast between types of cuisine (Mediterranean / non) or between different levels of quality (trattorias / haute cuisine) and, on the other hand, the similarities between the themes and categories (regional characteristics). There are a few examples of an exact overlapping of terms (, fish, barbecue, courtesy, day, ingredients, the atmosphere and a few others) and the synonymy of adjectives (wonderful, delightful, polite, courteous, impeccable, efficient) is quite evident. The restaurants can obviously present a mixture of these various elements, but the cluster analysis in disjunctive classes does not reveal this.

As can be seen from the factorial map of the groups, there is a clear overlapping of the classes (Figure 1), even if, for a better analysis, it should also take into account the third factorial dimension, as the clustering was performed starting from 3 factors. From a semantic point of view, the two factorial axes are easily interpreted. The horizontal axis referring to the first

factor puts the Mediterranean in contrast with the "Nordic" cuisine, or in other words, Sicilian restaurants are compared with those in Piedmont; however, the Sicilian are predominantly fish restaurants as opposed to the Piedmontese that mainly offer meat menus, whilst trattorias are somewhere between the two. The vertical axis referring to the second factor ranks the restaurants according to the quality of their cuisine: from the lowest level of trattorias up to the gastronomic excellence of the top restaurants.



3.2. The multi-class categorization of the restaurants on the basis of the TFIDF

In the last phase of the strategy, the TI for each thematic dictionary, thus 6 TI in all, were calculated for each restaurant. The highest value gives the dominant thematic category. It is then possible to assign other thematic categories to the documents which have sufficiently high TI values. In fact, the other TI in descending order for each restaurant represent a decreasing adherence of the document to the other themes (TI is zero when there is no element of the query contained in the review). The criterion of the smallest differences was applied in order to define a threshold value, below which the degree of adherence of the review to the category would not be considered interesting (see paragraph 2.2).

The difference between the TIMAX and the other values resulting from the query is calculated for each document. Taking the lowest value of the difference for each document, the average of the smallest differences for all documents was then calculated at 0,768. The TI with the smallest below-average value were considered significant. The reviews were then assigned categories corresponding to the query where the value of this difference was less than 0,768.

The results for this phase are as follows: in 75.3% (208 restaurants) of the cases, the categorization of the reviews on the first TI (TIMAX) coincides with that of the cluster analysis. Of these 49% (102 cases) have a single categorization. However, if we consider the second TI (with the smallest difference from the highest value), the overlapping reaches 94% (259 restaurants). In 10 cases, there is a different unique categorization from the one obtained with the cluster analysis. In the remaining 7 cases the categorization is different and multi-class.

Furthermore, the multi-class categorization describes all the intersections, of which only a part can be found on the factorial map in Figure 1. The quantitative distribution of the multi-class categorization can be seen as a whole in Table 1, which shows the ratio of multi-classes in relation to all the intersections of the classes attributed through the TFIDF[7].

	1 Traditional Piedmontese cuisine	2 Trattoria	3 Location/region atmosphere	4 Haute cuisine	5 Mediterranean flavours	6 Sicilian seafood cuisine
1 Traditional Piedmontese cuisine		0,053	**0,135**	0,063	*0,005*	*0,005*
2 Trattoria	0,053		0,068	*0,019*	*0,043*	*0,053*
3 Location/region atmosphere	0,135	0,068		**0,140**	*0,024*	*0,072*
4 Haute cuisine	0,063	0,019	0,140		*0,053*	*0,077*
5 Mediterranean flavours	0,005	0,043	0,024	0,053		**0,188**
6 Sicilian seafood cuisine	0,005	0,053	0,072	0,077	0,188	
% of overlapping of thematic pairing	0,261	0,237	0,440	0,353	0,314	0,396

The figures in bold indicate major intersections between classes (themes). From Table 1 it can be seen that there is a greater semantic similarity between Sicilian and Mediterranean cuisine restaurants (classes 6 and 5: 0,188), between haute cuisine and location/region and atmosphere aspects (classes 4 and 3: 0.140), and also between the traditional (Piedmontese) restaurants and those that offer local cuisine (classes 1 and 3: 0.135).

The column totals indicate the percentage of multi-classification for each query. The gastronomic offer under the location/region and atmosphere category (type 3) has a value of 44%, and is the class that intermingles the most with the others. In fact, the different reviews often provide descriptions of the local area. The second most frequent thematic class in terms of the co-presence of other classes is type 6 (39.6%) and it is, in fact, known that seafood cuisine can be present in all the categories. Finally, in the few cases that were not consistent with the clustering, the multi-class categorization highlights what the factorial level cannot explain, that is, the coexistence of 'distant themes' or the presence of outliers (marked in italics in the matrix). For example, the presence of top quality trattorias (pair 2-4), or Piedmontese trattorias with Mediterranean cuisine (1-5) or excellent Sicilian restaurants with meat menus (6-1), which have particularly low values in the table (respectively 0.019, 0.005, 0.005). The table as a whole thus illustrates the degree to which each type of restaurant may belong to each different thematic category.

The latter element is, in fuzzy logic, a kind of measure of the probability of belonging to a combination of two classes and therefore testifies the semantic similarity between the groups.

[7] The table quantifies the proportion of each pair of categories in relation to the total number of pairs of thematic classes. For example the value 0.135 at the intersection of TI 3 and TI 1 is obtained by putting 28 pairs of categories in relation to the total of 207 pairs (generated by 147 restaurants with double or triple categorizations). The restaurants that have more than 3 categories have not been included in this calculation, whilst those with a single attribution are excluded . The reading of a triangle of the matrix provides the probability of co-occurrence of each class with the others.

Similarly, for each restaurant it is possible to calculate the probability of belonging to each of the classes by putting its TI value in relation to the total of TI calculated for it in all the classes (see the example in the footnote).

Multi-categorization often highlights the bordering elements of a disjunctive clustering or outliers that cannot be explained at a single factorial plane. This case study of restaurant reviews is particularly suited for an understanding of the problem. It is obvious that restaurants can be classified under different aspects since, after the main classification, they have components of other themes. In the example in the footnote[8] there is clearly a co-presence of the semantic fields indicating (forms in bold) and (forms underlined).

4. Conclusion

As we have seen the proposed strategy makes it possible: i) to focus on the themes, using classic disjunctive clustering in the initial exploratory phase; ii) to identify the main semantic fields by means of the characteristic words of the classes obtained; iii) refine the corresponding dictionaries, as a basis for as many queries; iv) to re-launch an automatic classification using the TI for each query to obtain a multi-class categorization, where there is a substantial co-presence of themes. In the TI evaluation, a weight, as a function of the test-value of the elements in the cluster they belong to, could be assigned to the terms in the thematic dictionary forming the query.

This experimentation is able to produce even better results, if the composition of the thematic dictionaries is refined by introducing more idiomatic phrases (a few hundred are required) or by a more thorough analysis of the concepts created by the grouping of equivalent nouns (e.g. “synonyms” as the same type of food).

References

Bolasco S. (1999). Carocci, Roma.

Bolasco S., Canzonetti A., Capo F. M. (2005). . CISU, Roma.

Lebart L., Piron M., Steiner J. (2003). . Dunod, Paris.

Ricolfi L. (1992). . F. Angeli, Milano.

Salton G. (1989). Addison-Wesley.

Zadeh L.A. (1977). Fuzzy sets and their applications to pattern classification and clustering. In J. Van Ryzin (ed.), . Academic Press, N.Y.

[8] CL4 RG – Duomo (TI4: 2,53=33,8%; TI6: 1,93=25,8%): Il ristorante di ... e ... regala un' esperienza gastronomica fuori dal comune. il servizio è preciso, anche se un po' lento. la Carta dei vini contempla centinaia di **pregevoli** etichette. pane fatto in casa , ottimi **appetizer** offerti accompagnati da una flute di bollicine e quattro **menu degustazione**. un **capolavoro** il crudo_di_pesce, con otto assaggi. buono il tortino di ragusano in mantello di pecorino con verdure grigliate e confettura di azzeruole. squisiti gli spaghetti neri con crostacei seppie e **crema** di peperoni, così come quelli con **tartara** di pesce bottarga di tonno e succo di carote e i paccheri con **salsa** di pomodoro costoluto cipollotto fresco e spezzatino_di_pesce. tra i secondi, segnaliamo un ottimo turbante di spatola con mollicata ai pinoli e gnocchetti di verdura, le **eccellenti** costine di maialino nero dei Nebrodi ripiene con fricassea di borlotti, il cosciotto e costoletta di agnello locale flan di fagioli e frutta secca. da non perdere il **fondente** di cioccolato con **salsa** di **cioccolato bianco** confettura di pera e **gelato** di vaniglia; interessante la rivisitazione del cannolo siciliano . deliziosa **pasticceria** secca della casa.