A Hybrid Method For The Extraction And Classification Of Product Features From User Generated Contents

S.Pelosi; A.Maisto; M.Stingo; R.Guarasci.

DIPARTIMENTO DI SCIENZE POLITICHE, SOCIALI E DELLA COMUNICAZIONE

SCENARIO D'INTERESSE

Commercio elettronico
(piattaforme B2C , online booking apps, internet marketing.)

&

Electronic Word of Mouth
(recensioni, commenti, etc.)

OBIETTIVI DI RICERCA

- Riconoscimento automatico delle caratteristiche fondamentali dei beni e servizi recensiti.
- Classificazione automatica di recensioni generate da utenti online.
- Sistematizzazione della conoscenza estratta in un’ontologia orientata verso i prodotti.

BACKGROUND TEORICO: **Lessico-Grammatica**

Gross (1971, 1975) riconosce nella descrizione sistematica del lessico il metodo cardine per un'analisi esaustiva delle proprietà sintattiche e semantiche di una lingua.

1. Raccolta estensiva e comparazione dei dati linguistici
2. L'unità minima d'analisi è la frase semplice: **N_0 V N_1**
3. Non soltanto i verbi sono **predicativi**

IPOTESI DI RICERCA: *AggVal* ed il loro potere predittivo

E' possibile estrarre le caratteristiche di beni/servizi e classificare le recensioni fornite degli utenti, basandosi sulla **polarità** degli **aggettivi** all'interno di particolari strutture frasali.

SentIta (Maisto&Pelosi 2014, Pelosi 2015)	
Adjectives	**Entries**
Positive Items in SentIta	**1,358**
Negative Items in SentIta	**3,385**
Intensifiers in SentIta	**638**
Neutral Adjectives in Sdic_it	**28,664**
Adjectives in Sdic_it	**34,045**

N_0 essere AggVal

L'idea iniziale era accettabile

(stare, diventare, rimanere, restare, rendere, sembrare, apparire, risultare, rivelarsi, dimostrarsi, mostrarsi)

V-inf essere AggVal

Vedere quel film è stato demoralizzante

N_0 essere AggVal di V-inf

La polizia sembra incapace di fare indagini

N_0 essere AggVal a N_1

La giocabilità è inferiore alla serie precedente

N_0 essere AggVal Per N_1

Per me questo film è stato noioso

COSTRUZIONE DEL CORPUS

Il corpus, distribuito su 6 domini, è composto da 600 recensioni (50 negative e 50 positive per ogni dominio) fornite da utenti sulle maggiori piattaforme di e-commerce e siti web d'opinione su servizi e prodotti.

Text features	Cars	Smartphones	Books	Movies	Hotels	Games	Tot
Neg docs	50	50	50	50	50	50	**300**
Pos docs	50	50	50	50	50	50	**300**
Word forms	17,163	19,226	8,903	37,213	12,553	5,597	**101,655**
Tokens	**21,663**	**24,979**	**10,845**	**45,397**	**16,230**	**7,070**	**126,184**

POS TAGGING e LEMMATIZZAZIONE DEL CORPUS

Entrambe le operazioni sono state eseguite utilizzando un modulo del software `LG-Starship` (Maisto 2017), denominato `Mr.Ling`.

`Mr.Ling`

- `Mr.Tag`: PoS tagger appositamente creato per l'italiano, utilizza il tagset dei dizionari DELA (Elia, 1995; Elia et al. 2010). Precisione del 91,5%
- `Mr.Lemmi`: Lemmatizatore basato sul DELAF (Italian Electronic Dictionary of Flexed Forms). Il dizionario è stato diviso in 6 sottoinsiemi per ogni parte del discorso (N, V, Adj, Prep, Det). Precisione del 92%

POTENZIARE LE RISORSE LESSICALI

Dizionario in formato Json

Informazioni di tipo:

- Semantico
- Trasformazionale
- Distribuzionale
- Strutturale
- Nominalizzazioni
- Aggettivazioni

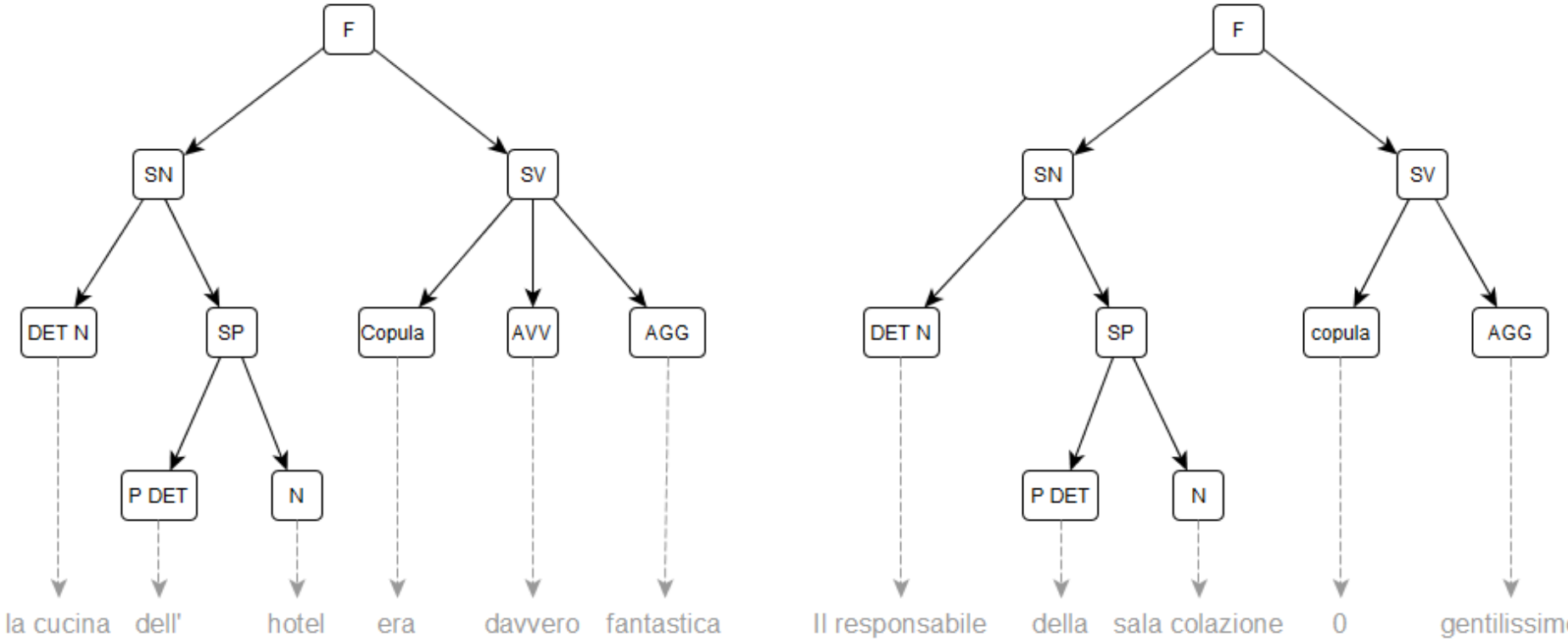
```
{
  lemma:"desiderare",
  type:"verb"
  role:"semantic predicate",
  lg_class: "43",
  structure:[
    N0:"experiencer",
    N1:"stimulus"
  ]
  trasformations:[
    nom:"des0",
    adj:["des01", "des02"]
    ]
},

{
  lemma:"desiderabile",
  id:"des01",
  type:"adj",
  structure:[
    N1:"experiencer",
    N0:"stimulus"
  ]
},

{
  lemma:"desideroso",
  id:"des02",
  type:"adj",
  structure:[
    N1:"stimulus",
    N0:"experiencer"
  ]
}
```

ESTRAZIONE AUTOMATICA DELLE CARATTERISTICHE DI BENI/SERVIZI

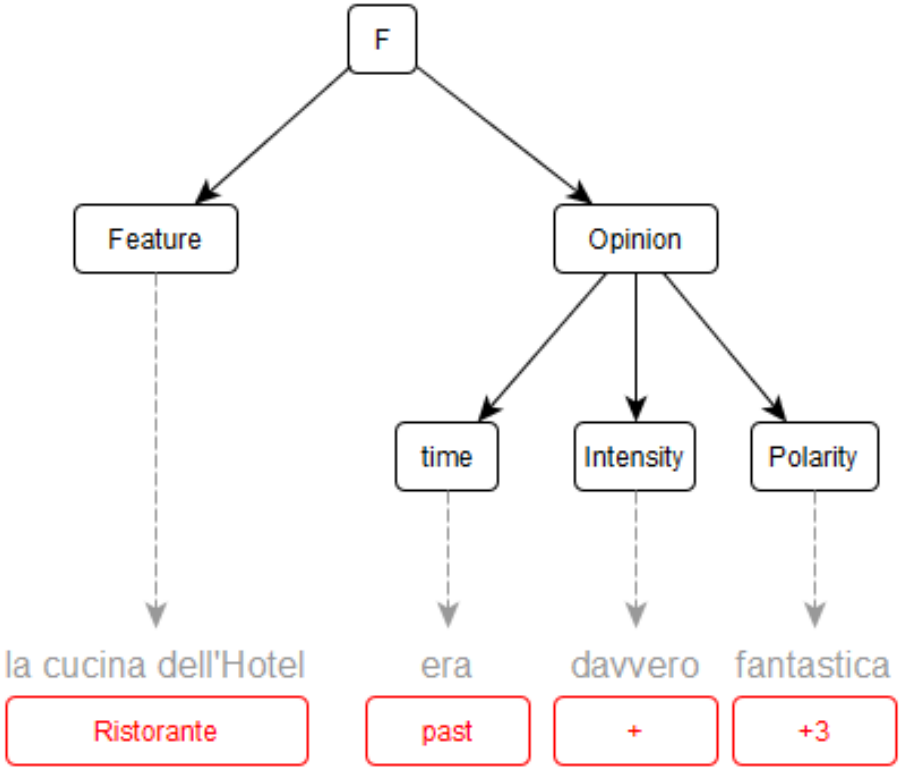
Gli *AggVal*, identificati come elementi predicativi, tracciano gli argomenti (fondamentali), i quali a loro volta identificano le features da estrarre.

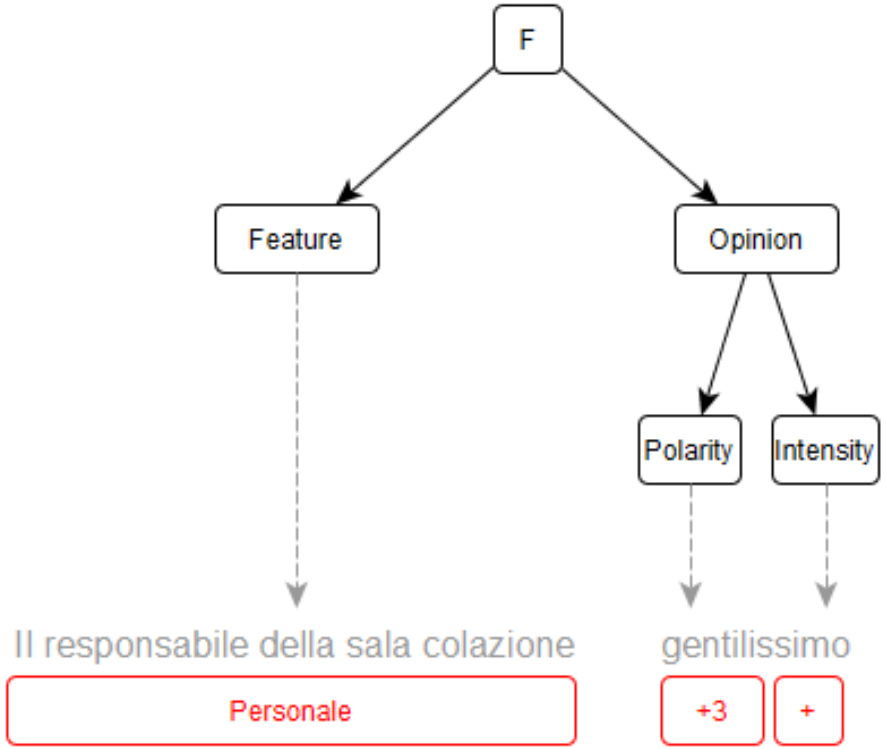


Rappresentazione grafica dell'outpout sintattico del parser

ESTRAZIONE AUTOMATICA DELLE CARATTERISTICHE DI BENI/SERVIZI

Gli *AggVal*, identificati come elementi predicativi, tracciano gli argomenti (fondamentali), i quali a loro volta identificano le features da estrarre.



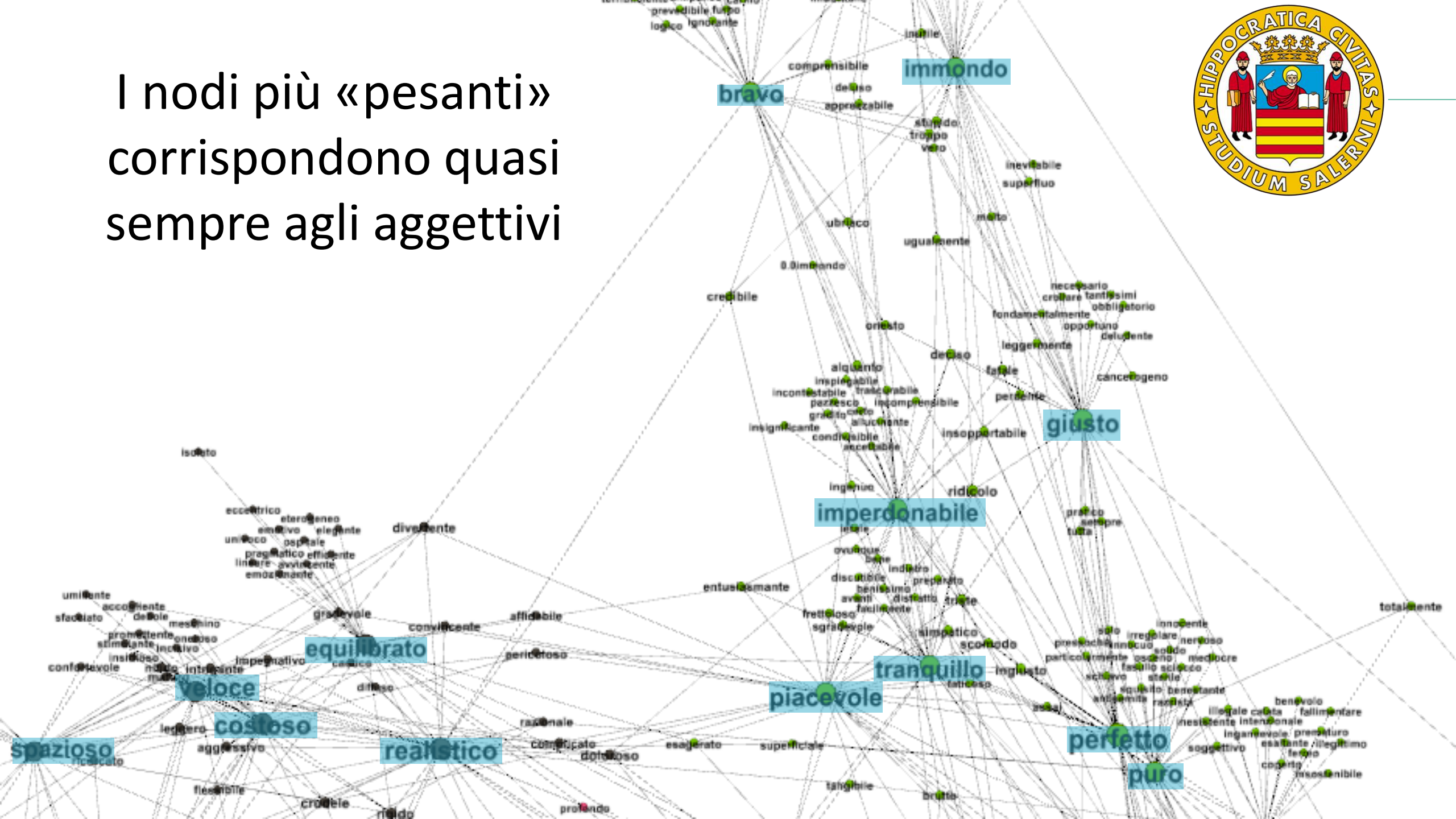


Rappresentazione grafica dei tag semantici assegnati dal parser

ANALISI SEMANTICA DELLE REVIEWS

1. Identificazione delle features.
2. Identificazione delle opinioni associate alle features.
3. Calcolo della Mutual Semantic Similiarity fra features
4. Creazione di un network semantico espanso

I nodi più «pesanti» corrispondono quasi sempre agli aggettivi



ESTRAZIONE AUTOMATICA DELLE CARATTERISTICHE DI BENI/SERVIZI

Similarità fra features estratte

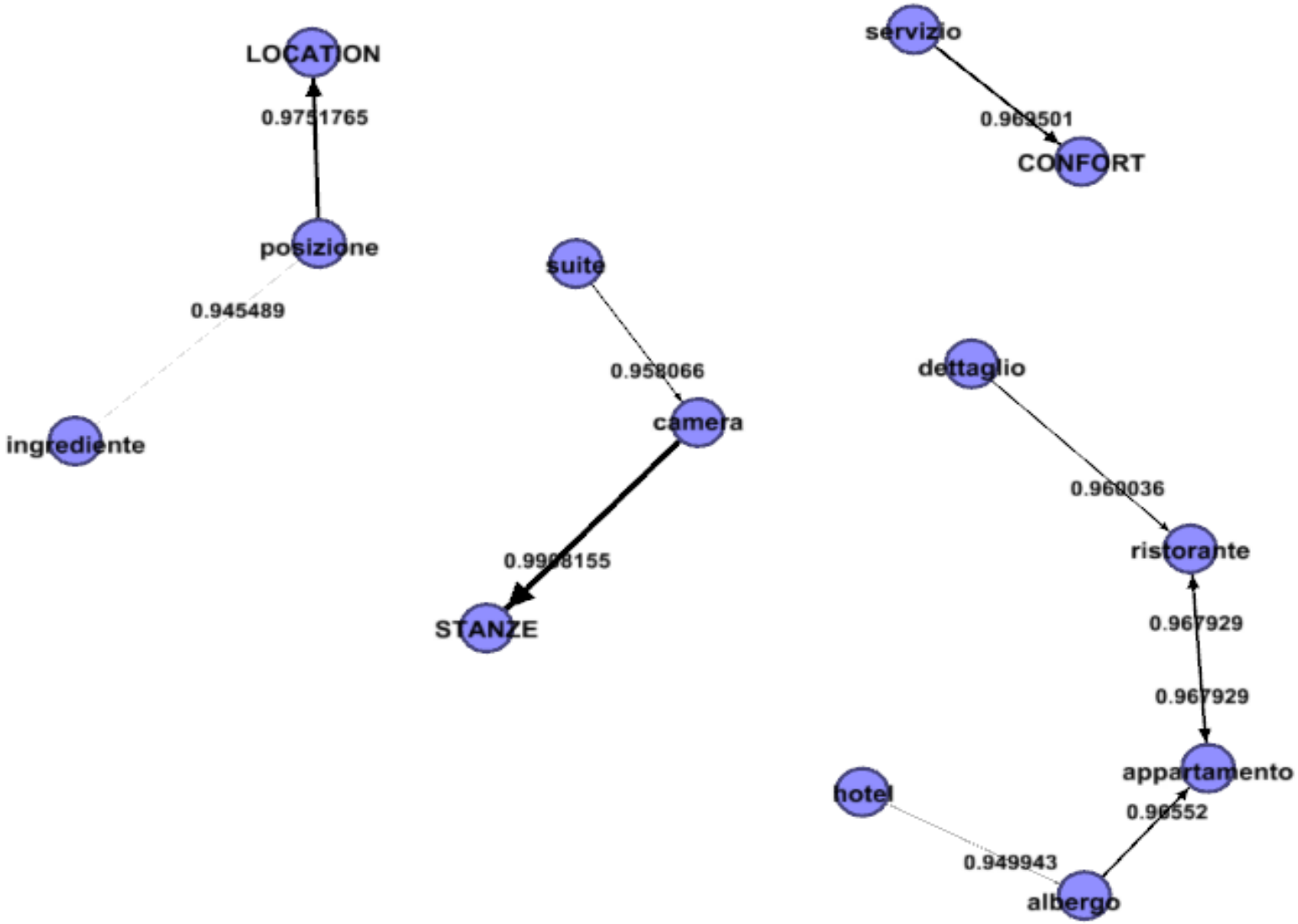
Feature 1	Feature 2	Similarità reciproca
Colazione «breakfast»	Ristorante «restaurant»	0,907
Colazione «breakfast»	Arredamento «furnitures»	0,828
Colazione «breakfast»	Vista «view»	0,751

ESTRAZIONE AUTOMATICA DELLE CARATTERISTICHE DI BENI/SERVIZI

Similarità fra features estratte e features generiche

Feature 1	**Feature 2**	**Similarità reciproca**
Suite	Camera	0,958
Suite	Stanza	0,949
Suite	**STANZE**	0,953

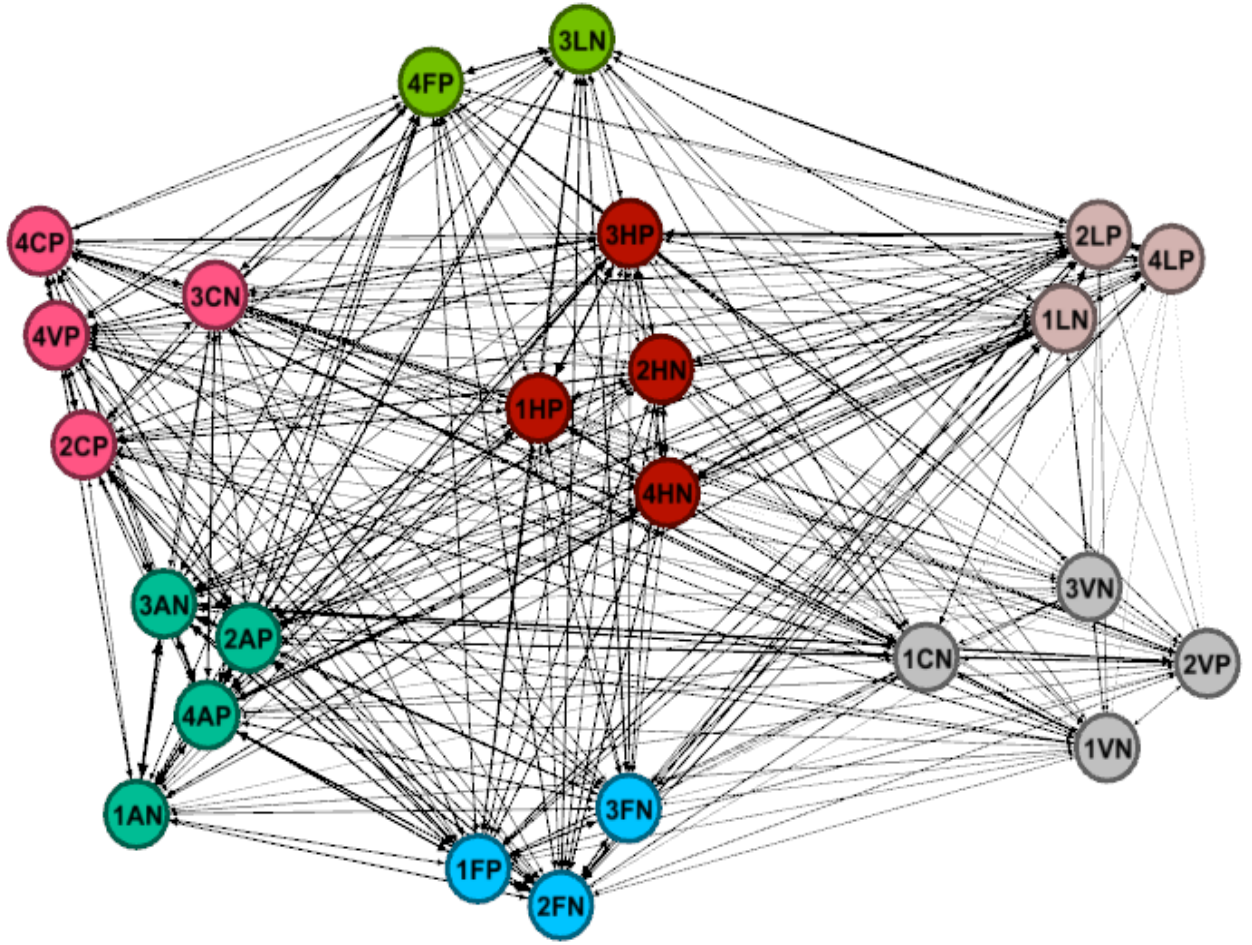
ESTRAZIONE AUTOMATICA DELLE CARATTERISTICHE DI BENI/SERVIZI



CLASSIFICAZIONE AUTOMATICA DELLE RECENSIONI

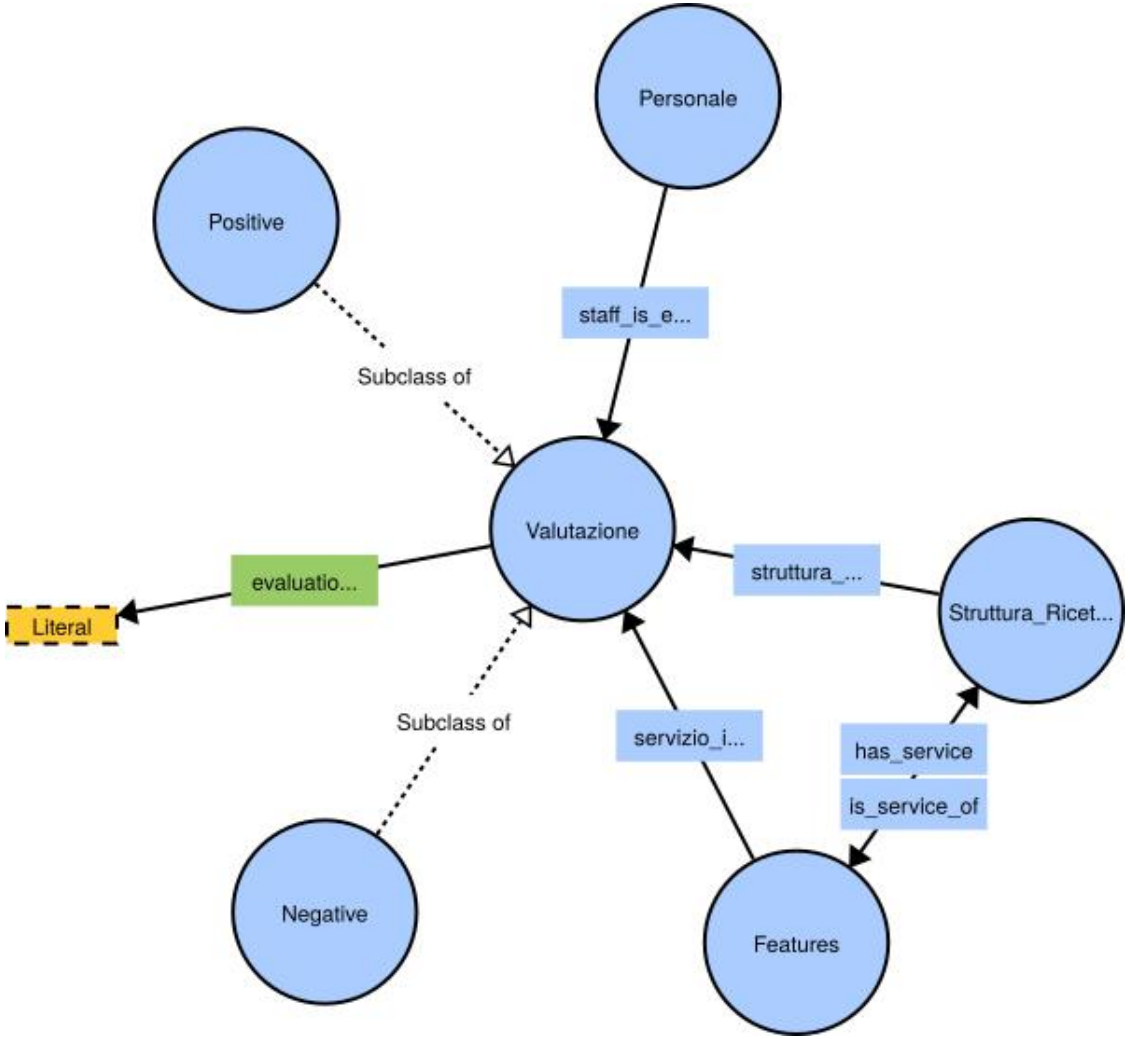
1 H N :

- 1 > numero del file
- H > lettera identificativa del dominio (in questo caso: Hotel)
- N > lettera indicante la polarità (in questo caso negativa)



PROTOTIPO D'ONTOLOGIA ORIENTATA AI PRODOTTI

- Struttura ricettiva: classe con object-property «has_service»
- Features: classe con object property «is_service_of».
- Personale: le posizioni lavorative coinvolte nel dominio di riferimento.
- Valutazione: classe divisa in positive/negative con data property «evaluation score» (---,--,-,+,++,+++). Le 3 classi sopra puntano a quest'ultima tramite proprietà «x_is_evaluated»



PROTOTIPO D'ONTOLOGIA ORIENTATA AI PRODOTTI

Codifica XML delle informazioni semantiche contenute all'interno di una porzione di recensione

```
…<BENEFIT SCORE="3" TYPE="PULIZIA">La pulizia è eccellente</BENEFIT>
<BENEFIT SCORE="3" TYPE="LOCATION">La vista sul mare è splendida</BENEFIT>
ma <DRAWBACK SCORE="2" TYPE="LOCATION">la notte purtroppo si sentiva qualche
schiamazzo</DRAWBACK >….
```

CONCLUSIONI: PROBLEMATICHE E SVILUPPI DI RICERCA

- Estensione delle risorse linguistiche Lessico-Grammaticali in formato Json
- Miglioramento della performance del parser sintattico di **LG-Starship**.
- Sperimentazione di una pipeline per il popolamento automatico delle ontologie orientate ai prodotti.
- Riconoscimento di valutazioni ironiche, sarcastiche, stereotipate, etc.

BIBLIOGRAFIA

Gross, M. 1971. *Transformational Analysis of French Verbal Constructions*. University of Pennsylvania.

Gross, M. 1975. *Méthodes en syntaxe. Régime des constructions complétives*. Hermann, Paris.

Maisto, A. and Pelosi, S. 2014. *A lexicon-based approach to sentiment analysis. The italian module for nooj*. In "Proceedings of the International Nooj 2014 Conference", University of Sassari, Italy. Cambridge Scholar Publishing.

Maisto, A. 2017. *A Hybrid Framework for Text Analysis*. Ph.D Thesis to be published. Department of Political, Social and Communication Sciences. University of Salerno, Italy.

Pelosi, S. 2015. *Sentita and doxa: Italian databases and tools for sentiment analysis purposes*. In "Proceedings of the Second Italian Conference on Computational Linguistics" CLiC-it 2015, pp. 226–231. Accademia University Press.

Elia, A. 1995. *Dizionari elettronici e applicazioni informatiche*. In "In III Giornate internazionali di Analisi Statistica dei dati Testuali, JADT", pp. 55-6, CISU, Roma.

Elia, A., Marano, F., Monteleone, M., Sabatino, S. and Vellutino, D. 2010. *Strutture lessicali delle informazioni comunitarie all'interno di domini specialistici*. In "Statistical Analysis of Textual Data, Proceedings of 10th International Conferences", "Journées D'Analyse Statistique des Données Textuelles", pp 9-11, Sapienza University, Rome, Italy.