TSRR

IFAC-TSRR vol. 7 (2015) 53-80

Elaborazione della ontologia ICF mediante tecnologie di analisi del linguaggio naturale e di "information retrieval"

N. Zoppetti[(1,*)], L. Burzagli[(1,**)], P.L. Emiliani[(1,***)]

([1]) Istituto di Fisica Applicata "Nello Carrara" del CNR (IFAC-CNR)

[(*)] N.Zoppetti@ifac.cnr.it
[(**)] L.Burzagli@ifac.cnr.it
[(***)] P.L.Emiliani@ifac.cnr.it

 ISSN 2035-5831

1 - Introduzione

Il progetto italiano D4All[1] prevede la progettazione di un ambiente domotico capace di seguire il profilo d'utente, piuttosto che rivolgersi ad uno stereotipo con caratteristiche fissate a priori e non corrispondenti alle diversità espresse dai vari individui. Ad esempio gli utenti possono avere una vista non perfetta, la loro mobilità può essere ridotta, oppure il loro udito può aver subito dei danni. Tale ambiente dovrà quindi fornire una personalizzazione delle funzionalità previste in accordo con le caratteristiche della persona. Tale metodologia di progettazione richiede in ogni fase del suo ciclo di vita (progetto, uso, controllo, monitoraggio, etc.) l'utilizzo di un insieme di strumenti eterogenei e un'attenzione particolare all'utente che fruirà dei servizi e delle tecnologie e al modo in cui interagirà con essi.

Al fine di ottimizzare la descrizione dell'utente in riferimento alle funzionalità richieste all'ambiente, l'IFAC, quale partner del progetto, ha proposto in seno al consorzio l'adozione di una classificazione internazionale delle funzionalità dell'individuo e delle sue attività. Si tratta della"InternationalClassification of Functioning,Disability and Health" [1], comunemente indicata come ICF, che l'Organizzazione Mondiale della Sanità ha pubblicato nel 2001, derivandola dalla precedente ICIDH del 1980. Tale classificazione struttura in modo estremamente dettagliato le funzioni del corpo, le strutture del corpo e le attività dell'individuo, per identificare situazioni di disabilità che emergono come prodotto di una condizione fisica, dell'attività condotta dall'individuo e dal contesto nel quale l'attività si attua.

In un ambiente che cerca di ottimizzare la personalizzazione dei servizi per ciascun individuo risulta di particolare importanza la relazione che esiste tra l'attività prescelta e le caratteristiche dell'individuo, alfine di determinare, per ogni funzionalità prevista, una realizzazione conforme ad essa.

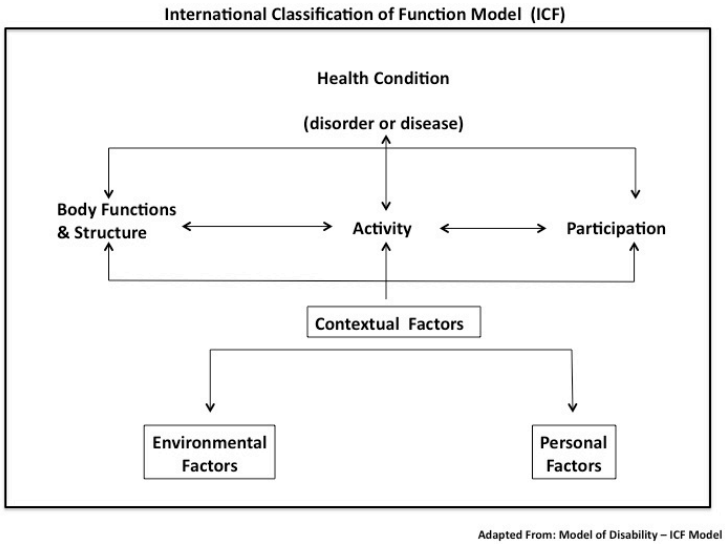

Fig.1–Struttura della ICF.

2 - Definizione del problema

Il lavoro descritto in questo report riguarda la determinazione di connessioni tra diverse categorie della ICF secondo la quale lo stato di ogni persona può essere descritto in base a quattro macro categorie che descrivono le sue caratteristiche fisiche ed abilità con il contesto in cui egli vive.

[1] http://www.d4all.eu/

- **b. BODY FUNCTIONS.** Body functions are the physiological functions of body systems (including psychological functions). Impairments are problems in body function or structure as a significant deviation or loss.
- **s. BODY STRUCTURES.** Body structures are anatomical parts of the body such as organs, limbs and their components. Impairments are problems in body function or structure as a significant deviation or loss.
- **d. ACTIVITIES AND PARTECIPATION.** Activity is the execution of a task or action by an individual. Participation is involvement in a life situation. Activity limitations are difficulties an individual may have in executing activities. Participation restrictions are problems an individual may experience in involvement in life situations.
- **e. ENVIRONMENTAL FACTORS.** Environmental factors make up the physical, social and attitudinal environment in which people live and conduct their lives.

Il problema di riferimento scelto consiste nel collegare in modo automatico ad una generica attività umana (tra quelle classificate nella ICF nella categoria d) le funzioni del corpo (classificate nella categoria b dell'ICF) potenzialmente coinvolte, in modo automatico. A questo fine è stato selezionato un metodo che si basa sulle descrizioni testuali contenute nella ICF stessa. Ad esempio, se selezioniamo l'attività **d6300 Preparing simple meals,** lo scopo è quello di selezionare tutte le "body functions" coinvolte, e dare ad esse una priorità. Stabilire cioè se le funzioni della vista (b210) siano coinvolte e se abbiano una priorità maggiore delle funzioni di movimento (b750-b789). Questo collegamento costituisce il punto di partenza per ogni eventuale adattamento da introdurre nell'ambiente per favorire l'utente.

La ricerca prende come elemento di riferimento una precedente attività portata avanti per stabilire la correlazione tra le attività relative alla "domestic life" (cap. 6) e le funzioni del corpo collegate, eseguita senza ausili di elaborazione automatica, ma solo frutto dell'esperienza umana descritta e pubblicata nel Deliverable 1.24 del progetto D4all [2]. In tale attività l'esperto ha deciso i collegamenti esistenti usando le sue conoscenze e le capacità umane di ragionamento. Il lavoro descritto in questo documento ha lo scopo di automatizzare questo processo.

2.1 - Cenni sulle ontologie

La classificazione ICF è una risorsa disponibile in vari formati e tra questi anche come ontologia[2,3].

Il termine ontologia è usato nel campo dell'intelligenza artificiale e della rappresentazione della conoscenza per descrivere un insieme di termini e di relazioni che definiscono un dominio strutturato in maniera gerarchica, che può essere usato come fondamento di una"knowledge base".

I principali componenti di una generica ontologia possono essere così classificati:

- **Individui**: detti anche istanze/oggetti,sono i componenti di base dell'informazione rappresentata.
- **Classi**: rappresentano le tipologie di oggetto contenute nell'ontologia. In relazione agli individui possono essere assimilate a dei concetti/astrazioni.
- **Attributi**: si tratta delle proprietà e delle caratteristiche che gli individui o le classi possono avere.
- **Relazioni**: modalità con cui individui e classi possono essere mutuamente collegati.

Le ontologie sono 'codificate' utilizzando varie tipologie di linguaggi tra i quali, vi sono RDF[4], RDFS[5] ed il loro discendente OWL[6], specificatamente introdotto per la diffusione delle ontologie sul web.

---

[2] http://en.wikipedia.org/wiki/Ontology_(information_science)

[3] http://bioportal.bioontology.org/ontologies/ICF

[4] http://en.wikipedia.org/wiki/Resource_Description_Framework

Secondo questi modelli di rappresentazione la conoscenza è rappresentabile come un insieme interconnesso di **affermazioni**, ciascuna delle quali è composta da tre elementi: un **soggetto**, un **predicato** ed un **oggetto**.

Con riferimento agli elementi delle ontologie precedentemente elencati, un generico predicato può essere la relazione che collega un generico individuo (il soggetto) ad un altro individuo o ad un attributo (che in questo caso hanno entrambi la funzione di oggetto). Oppure un'affermazione può specificare che una classe (soggetto) è una specificazione (predicato) di un'altra classe (oggetto) o che un individuo (soggetto) è della tipologia (predicato) descritta da una classe (oggetto).

Nelle Fig.2 e Fig.3 si riportano delle rappresentazioni (ottenute con il software Protégé[7]) di elementi contenuti nella ontologia ICF .

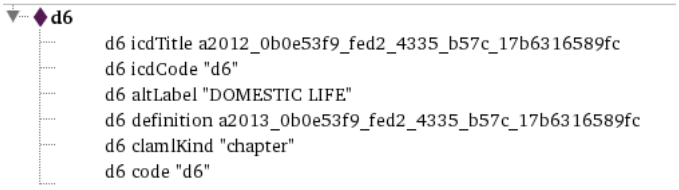

Fig.2– .relazioni tra d6 (individuo) ed altri individui/attributi con il software Protégé.

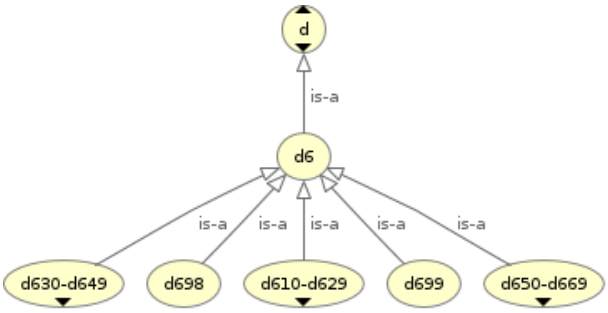

Fig.3–relazioni tra classi in ICF con Protégé

3 - Strumenti e tecnologie utilizzati

Il primo passo di questa attività è rappresentato dalla scelta degli strumenti e delle tecnologie da impiegare. Tale scelta è stata guidata da esigenze di carattere abbastanza diverso tra loro.

In particolare le caratteristiche desiderate per il linguaggio di programmazione erano le seguenti:

- adatto sia per lo sviluppo di applicazioni client, sia per applicazioni web;
- dotato di ampia disponibilità di strumenti dedicati per:
  - la gestione di ontologie;
  - l'elaborazione del linguaggio naturale, in relazione alla necessità di basare le elaborazioni sulle descrizioni testuali delle categorie ICF di interesse;
- diffuso in ambito scientifico e con ampia comunità di utilizzatori.

I requisiti elencati sopra hanno portato a restringere il campo ai due linguaggi di programmazione Python[8]e Java[9]. La scelta finale è ricaduta sul primo, sia in considerazione del fatto che già si possedevano competenze in merito, sia al fatto che, a differenze di Java, si tratta di un linguaggio open source.

La scelta della versione di tale software ha costituito il passo successivo. Le alternative erano costituite dalla versione 2, più matura, diffusa e supportata, oppure la più recente versione 3, più moderna ma al momento meno diffusa e con un minor numero di moduli a disposizione. La scelta è ricaduta su Python3 anche in

---

[5]http://en.wikipedia.org/wiki/RDF_Schema

[6]http://en.wikipedia.org/wiki/Web_Ontology_Language

[7]http://protege.stanford.edu/

[8]https://www.python.org/

[9]https://www.java.com/

considerazione del fatto che i due principali moduli individuati per l'elaborazione del linguaggio naturale (nltk[10], [3]).

e per la gestione di ontologie (rdflib[11,12]) sono disponibili sia per Python2 sia per Python3. In seguito si è verificato che anche il modulo gensim[13] [4], ampiamente utilizzato nelle soluzioni fino ad ora sviluppate, è disponibile per Python3.

Per quanto riguarda la strategia di sviluppo, si è optato per un percorso in tre passi, elencati nel seguito.

- Test dei moduli scelti e prime funzionalità implementate in notebook di IPython14, un ambiente di sviluppo basato su browser web, che permette di scrivere e provare il codice, combinandolo con testo formattato, immagini e con l'output stesso del codice.
- Implementazione di script (in genere più di uno che operano in serie) con semplice interfaccia testuale da riga di comando.
- Integrazione degli script in una piattaforma di prova basata sul modulo django15. La piattaforma mette a disposizione dell'utente un'interfaccia web con la quale provare le funzionalità implementate.

Il primo passaggio dei tre elencati è stato ispirato dalla necessità di disporre di una piattaforma di sviluppo che permettesse di integrare il codice con risorse documentali esterne.

4 - Descrizione del sistema di applicazioni realizzato

Per descrivere il sistema realizzato si farà riferimento all'implementazione delle soluzioni sotto forma di script (punto 2 paragrafo precedente) che si presta particolarmente a tale scopo. Le elaborazioni effettuate negli script si articolano nei passaggi elencati nel seguito, descritti anche nello schema di Fig.4.

- Consultazione ontologia ICF ed estrazione delle descrizioni testuali di interesse.
- Generazione dei corpora[16] testuali di base che possono basarsi sulle sole descrizioni testuali contenute nella ICF oppure su database testuali esterni, anche di grandi dimensioni (ad esempio DBpedia[17]e wordnet[18]) e rappresentazione dei corpora testuali di interesse sotto forma di vettori[19].
- Esecuzione delle *query di somiglianza* per determinare le affinità tra diversi documenti testuali.
- Postprocessing.

[10] http://www.nltk.org/

[11] www.rdflib.net/

[12] http://rdflib.readthedocs.org/en/latest/index.html

[13] http://radimrehurek.com/gensim/

[14] http://ipython.org/notebook.html

[15] https://www.djangoproject.com/

[16] http://it.wikipedia.org/wiki/Corpus

[17] http://en.wikipedia.org/wiki/DBpedia

[18] http://wordnet.princeton.edu/

[19] http://en.wikipedia.org/wiki/Vector_space_model

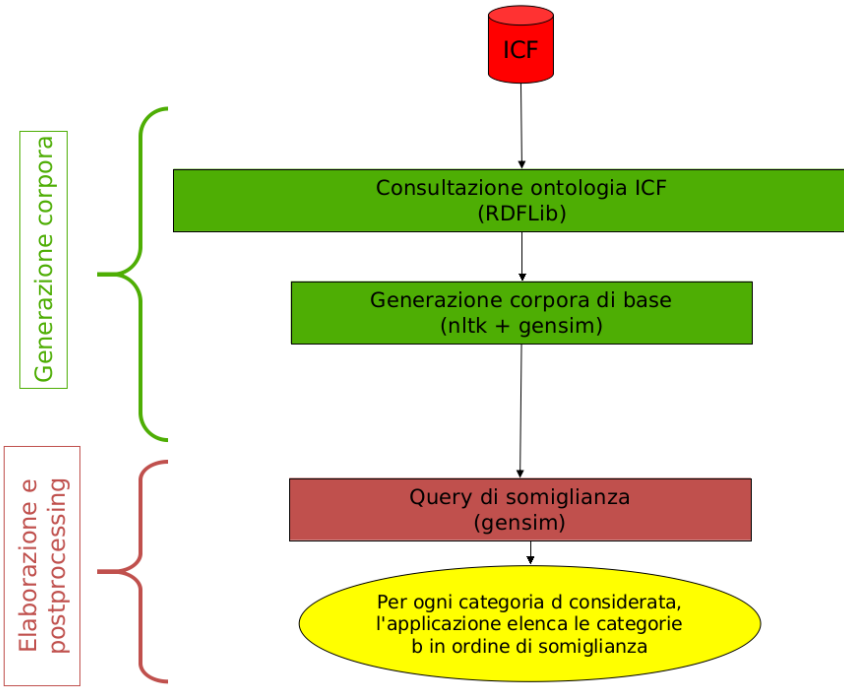

Fig.4–articolazione delle elaborazioni.

Nel seguito di questo capitolo si descriveranno brevemente le principali caratteristiche di ciascuno dei passaggi elencati.

4.1 - Estrazione da ICF delle descrizioni testuali di interesse (SPARQL)

Per consultare l'ontologia che rappresenta l'ICF (in formato owl) usando il linguaggio Python3 è stato utilizzato il già citato modulo rdflib[11], che mette a disposizione una specifica classe denominata `Graph` usata per rappresentare una generica ontologia nella memoria del calcolatore come una struttura a grafo formata da triple (soggetto,predicato, oggetto)[20].

Gli oggetti di tipo `Graph` supportano poi l'esecuzione di query SPARQL sulla ontologia rappresentata dove SPARQL[21]è un linguaggio di query specifico per le ontologie (standard W3C).

La query SPARQL che estrae i codici, i titoli e le descrizioni delle categorie con codice che inizia con "d6" è riportata nell'esempio di codice n.1.

```
1  PREFIX icd: <http://who.int/icd#>
2  SELECT  ?code  ?tit  ?def
3  WHERE {
4          ?node icd:icdCode ?code .
5  ?node icd:definition ?defref .
6  ?defreficd:label ?def .
7          ?node icd:icdTitle ?titleref .
8  ?titlereficd:label ?tit .
9             FILTER regex(?code,"^d6[0-9]{3}$")
10 }
```

Esempio di codice n.1

[20] http://rdflib.readthedocs.org/en/latest/intro_to_parsing.html

[21] http://www.w3.org/TR/sparql11-query/

Alla riga 1 della query la direttiva `PREFIX` definisce un abbreviazione usata poi nelle successive righe della query. Alla riga 2 si definiscono i tre 'campi' che si desidera estrarre che sono il codice della categoria selezionata `?code`, il relativo titolo `?tit` e la definizione `?def`. Tra la riga 3 e la riga 10, con la clausola `WHERE` si definiscono i vincoli imposti nella query ed in particolare si cercano gli elementi `?code,?tit` e `?def` dell'ontologia che soddisfano tutti i requisiti elencati nel seguito.

riga 4. `?node` ha un codice ICF icd:icdCode uguale a `?code`
riga 5. `?node` ha un riferimento a definizione icd:definition uguale a `?defref`
riga 6. `?defref` ha una stringa associata icd:label uguale a `?def`
riga 7. `?node` ha un riferimento a titolo icd:icdTitle uguale a `?titleref`
riga 8. ?titleref ha una stringa associata icd:label uguale a `?tit`
riga 9. `?code` sia conforme al pattern di espressione regolare ^d6[0-9]{3}$

La query SPARQL illustrata restituisce l'insieme delle terne `?code`, `?tit` e `?def` rappresentate nella ontologia ICF che soddisfano i requisiti considerati. I risultati sono riportati nell' Esempio di output n.1 (dove, per brevità, si è riportato in coda la descrizione di una sola delle categorie individuate dalla query).

```
---------------------------------------------------------------------------
?code -    ?tit
---------------------------------------------------------------------------
d6100  -  Buying a place to live
d6101  -  Renting a place to live
d6102  -  Furnishing a place to live
d6200  -  Shopping
d6201  -  Gathering daily necessities
d6300  -  Preparing simple meals
d6301  -  Preparing complex meals
d6400  -  Washing and drying clothes and garments
d6401  -  Cleaning cooking area and utensils
d6402  -  Cleaning living area
d6403  -  Using household appliances
d6404  -  Storing daily necessities
d6405  -  Disposing of garbage
d6500  -  Making and repairing clothes
d6501  -  Maintaining dwelling and furnishings
d6502  -  Maintaining domestic appliances
d6503  -  Maintaining vehicles
d6504  -  Maintaining assistive devices
d6505  -  Taking care of plants, indoors and outdoors
d6506  -  Taking care of animals
d6600  -  Assisting others with self-care
d6601  -  Assisting others in movement
d6602  -  Assisting others in communication
d6603  -  Assisting others in interpersonal relations
d6604  -  Assisting others in nutrition
d6605  -  Assisting others in health maintenance
---------------------------------------------------------------------------
?code -    ?def
---------------------------------------------------------------------------
d6201  -Obtaining, without exchange of money, goods and services
          required for daily living (including instructing and supervising
          an intermediate to gather daily necessities), such as by
          harvesting vegetables and fruits and getting water and fuel.
```

Esempio di output n.1

La gestione informatica delle risorse testuali di interesse risente della tipologia di elaborazione che si desidera effettuare. Come già accennato, per questa applicazione si è scelto di utilizzare tecniche di elaborazione basate sulla rappresentazione del testo come vettore numerico. Nei prossimi paragrafi si spiega inizialmente cosa significa rappresentare il testo con un vettore e quale è il senso e lo scopo di tale rappresentazione. In seguito si descriveranno i passaggi necessari ad ottenere tale rappresentazione a partire da collezioni di testi come quelli restituiti dalla query SPARQL riportata precedentemente.

4.2 - La rappresentazione del testo in vettori e la somiglianza tra testi

Il modello denominato "bag of words" (BOW)[22] è una rappresentazione semplificata utilizzata nell'ambito dell'elaborazione del linguaggio naturale e nell'information retrieval (IR). Secondo questo modello di rappresentazione un generico testo è rappresentato dall'insieme delle sue parole, trascurando la grammatica e anche l'ordine delle parole stesse, ma mantenendo la loro molteplicità. Questo insieme è esprimibile mediante un vettore di interi in cui l'i-esimo elemento rappresenta la molteplicità nel testo considerato della parola associata univocamente all'indice i. Le dimensioni del vettore saranno quelle del '*vocabolario*' di base cioè il numero di parole diverse rappresentate ed ogni parola del vocabolario è univocamente associata ad una dimensione dello spazio vettoriale utilizzato (e all'indice i che le corrisponde).

Una volta determinata la rappresentazione vettoriale dei testi considerati, una misura della loro somiglianza può essere ottenuta determinando il prodotto scalare tra i vettori stessi. Due parole che condividono le stesse parole con la stessa molteplicità sono quindi considerate identiche, indipendentemente dall'ordinamento delle parole nella frase.

È importante osservare sin da ora che il modello di rappresentazione BOW non è l'unico possibile ma che, a partire da esso, si possono ottenere (mediante opportune **trasformazioni)** rappresentazioni vettoriali dei testi più complesse. Ovviamente, usando diverse rappresentazioni vettoriali dei testi cambiano anche i risultati dei prodotti scalari che rappresentano le somiglianze tra i testi stessi.

Limitandosi per il momento al modello BOW, si applicano i concetti illustrati ad un semplice caso concreto definito nello script Python dell'esempio di codice n.2, nel quale:

(1) si definiscono 3 frasi
(righe 5-7);

(2) si dividono le frasi in parole
(righe 13-15);

(3) si creano dei dizionari che rappresentano il vocabolario
(righe 23, 26);

(4) si rappresentano le frasi in vettori secondo il modello BOW
(righe 32-34);

(5) si determinano gli indici di somiglianza tra le 3 coppie di frasi
(righe 40-42).

L'output dello script Python dell'esempio di codice n.2è riportato nell'esempio di output n.2

[22] http://en.wikipedia.org/wiki/Bag-of-words_model

```
1  Import numpy as np
2  ######################
3  #1         #
4  ######################
5  f1 = """Questa è la prima frase considerata"""
6  f2 = """Questa è la seconda frase considerata"""
7  f3 = """E infine ecco la terza frase che è una frase con una forma un
8  po’ diversa dalle altre"""
9
10 ######################
11 #         2          #
12 ######################
13 w1 = f1.lower().split(" ")
14 w2 = f2.lower().split(" ")
15 w3 = f3.lower().split(" ")
16 print(),print('w1 = ', w1),print('w2 = ',w2),print('w3 = ',w3)
17
18 ######################
19 #3         #
20 ######################
21 #i2w è un dizionario che associa la parola
22 #del vocabolario al rispettivo indice numerico
23 i2w = { i:w for i,w in enumerate(sorted(list(set(w1+w2+w3))))}
24 #w2i è un dizionario che associa un indice numerico
25 #ad ogni parola diversa contenuta nelle tre frasi
26 w2i = { w:i for i,w in enumerate(sorted(list(set(w1+w2+w3))))}
27 print(),print('w2i = ',w2i),print('i2w = ',i2w)
28
29 ######################
30 #         4          #
31 ######################
32 v1 = np.array([ w1.count(i2w[i]) for i in range(len(i2w))])
33 v2 = np.array([ w2.count(i2w[i]) for i in range(len(i2w))])
34 v3 = np.array([ w3.count(i2w[i]) for i in range(len(i2w))])
35 print(),print('v1 = ',v1),print('v2 = ',v2),print('v3 = ',v3)
36
37 ######################
38 #         5          #
39 ######################
40 sim12 = np.dot(v1,v2)
41 sim13 = np.dot(v1,v3)
42 sim23 = np.dot(v2,v3)
43 print(),print('sim12 = ',sim12),print('sim13 = ',sim13)
44 print('sim23 = ',sim23)
```

Esempio di codice n.2

```
w1 =  ['questa', 'è', 'la', 'prima', 'frase', 'considerata']
w2 =  ['questa', 'è', 'la', 'seconda', 'frase', 'considerata']
w3 =  ['e', 'infine', 'ecco', 'la', 'terza', 'frase', 'che', 'è', 'una', 'frase', 'con', 'una', 'forma', 'un', 'po'',
'diversa', 'dalle', 'altre']

w2i =  {'altre': 0, 'seconda': 15, 'diversa': 5, 'terza': 16, 'la': 11, 'è': 19, 'dalle': 4, 'un': 17, 'forma': 8,
'po'': 12, 'e': 6, 'considerata': 3, 'questa': 14, 'ecco': 7, 'con': 2, 'frase': 9, 'che': 1, 'prima': 13, 'una': 18,
'infine': 10}
i2w =  {0: 'altre', 1: 'che', 2: 'con', 3: 'considerata', 4: 'dalle', 5: 'diversa', 6: 'e', 7: 'ecco', 8: 'forma', 9:
'frase', 10: 'infine', 11: 'la', 12: 'po'', 13: 'prima', 14: 'questa', 15: 'seconda', 16: 'terza', 17: 'un', 18:
'una', 19: 'è'}

v1 =  [0 0 0 1 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 1]
v2 =  [0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 1 0 0 0 1]
v3 =  [1 1 1 0 1 1 1 1 1 2 1 1 1 0 0 0 1 1 2 1]

sim12 =  5
sim13 =  4
sim23 =  4
```

Esempio di output n.2

4.3 - Elaborazione di base dei testi

L'Esempio di codice n.2 semplifica molto la situazione sia per quanto riguarda la determinazione delle somiglianze tra frasi diverse, sia per quanto riguarda la elaborazione di base dei testi, che in tal caso si limita alla suddivisione delle frase in parole e nella successiva generazione dei vocabolari.

Per quanto riguarda l'elaborazione di base dei testi, un primo problema riguarda la presenza della punteggiatura (trascurata nell'esempio) e di parole molto comuni (spesso si tratta di articoli, e congiunzioni). Queste ultime non devono essere considerate nelle successive analisi in quanto da una parte sono poco significative per la determinazione delle somiglianze tra frasi/documenti e dall'altra aumentano le dimensioni del vocabolario rendendo più onerose le elaborazioni. Il primo passaggio effettuato nelle applicazioni è consistito quindi nell'eliminazione della punteggiatura e delle cosiddette *stopwords*.

Sempre allo scopo di aumentare la probabilità di distinguere le somiglianze tra testi, è opportuno *normalizzare* il testo in modo tale che due termini che differiscono ad esempio per tipologia di carattere (maiuscolo/minuscolo) o per declinazione (maschile/femminile singolare/plurale) siano ricondotti allo stesso termine nel vocabolario.

Se la distinzione maiuscolo/minuscolo può essere facilmente gestita trasformando tutto il testo in minuscolo, come nell'esempio, per quanto riguarda le declinazioni in numero e genere, ma anche, ad esempio, le coniugazioni dei verbi, il problema è più complesso e nel caso delle applicazioni realizzate è stato affrontato con due tecniche classiche nel NLP denominate *lemmatisation*[23]e *stemming*[24] di cui nel seguito si riportano due brevi estratti delle definizioni di wikipedia in italiano.

La *lemmatisation* o lemmatizzazione è il processo di riduzione di una forma flessa di una parola alla sua forma canonica (non marcata), detta lemma. In linguistica si dice lemma la citazione di una parola, ossia quella parola che per convenzione è scelta per rappresentare tutte le forme di una flessione. Nell'elaborazione del linguaggio naturale, la lemmatizzazione è il processo algoritmico che determina automaticamente il lemma di una data parola. Il processo può coinvolgere altre attività di elaborazione del linguaggio, quali ad esempio l'analisi morfologica e grammaticale.

[23] http://en.wikipedia.org/wiki/Lemmatisation

[24] http://en.wikipedia.org/wiki/Stemming

Lo *stemming* è il processo di riduzione della forma flessa di una parola alla sua forma radice, detta tema. Il tema non corrisponde necessariamente al lemma della parola: normalmente è sufficiente che le parole correlate siano mappate allo stesso tema (ad esempio, che andare, andai, andò mappino al tema and), anche se quest'ultimo non è un valido lemma per la parola.

Vedremo in seguito quanto si è utilizzato una tecnica piuttosto che l'altra.

4.4 - Documenti testuali di base

Per documenti testuali di base si intendono quelli estratti dalla ontologia ICF, nella forma in cui sono stati recuperati dalla opportuna query SPARQL.

Nell' esempio di output n.2 sono già state elencate le categorie d6, cioè quelle relative alle attività domestiche. In realtà nelle elaborazioni fatte si è preso in considerazione un sottoinsieme delle d6 che corrisponde all'insieme delle attività legate all'ambiente cucina che è quello di interesse per il progetto all'interno del quale si inserisce questa attività.

Nel seguito si elencano i codici titoli (evidenziati in grassetto) e le descrizioni delle sole categorie d6 considerate.

- **d6102 - Furnishing a place to live** - Equipping and arranging a living space with furniture, fixtures and other fittings and decorating rooms.
- **d6300 - Preparing simple meals** - Organizing, cooking and serving meals with a small number of ingredients that require easy methods of preparation and serving, such as making a snack or small meal, and transforming food ingredients by cutting and stirring, boiling and heating food such as rice or potatoes.
- **d6301 - Preparing complex meals** - Planning, organizing, cooking and serving meals with a large number of ingredients that require complex methods of preparation and serving, such as planning a meal with several dishes, and transforming food ingredients by combined actions of peeling, slicing, mixing, kneading, stirring, presenting and serving food in a manner appropriate to the occasion and culture.
- **d6400 - Washing and drying clothes and garments** - Washing clothes and garments by hand and hanging them out to dry in the air.
- **d6401 - Cleaning cooking area and utensils** - Cleaning up after cooking, such as by washing dishes, pans, pots and cooking utensils, and cleaning tables and floors around cooking and eating area.
- **d6402 - Cleaning living area** - Cleaning the living areas of the household, such as by tidying and dusting, sweeping, swabbing, mopping floors, cleaning windows and walls, cleaning bathrooms and toilets, cleaning household furnishings.
- **d6403 - Using household appliances** - Using all kinds of household appliances, such as washing machines, driers, irons, vacuum cleaners and dishwashers.
- **d6404 - Storing daily necessities** - Storing food, drinks, clothes and other household goods required for daily living; preparing food for conservation by canning, salting or refrigerating, keeping food fresh and out of the reach of animals.
- **d6405 - Disposing of garbage** - Disposing of household garbage such as by collecting trash and rubbish around the house, preparing garbage for disposal, using garbage disposal appliances; burning garbage.
- **d6502 - Maintaining domestic appliances** - Repairing and taking care of all domestic appliances for cooking, cleaning and repairing, such as by oiling and repairing tools and maintaining the washing machine.
- **d6600 - Assisting others with self-care** - Assisting household members and others in performing self-care, including helping others with eating, bathing and dressing; taking care of children or members of the household who are sick or have difficulties with basic self-care; helping others with their toileting.
- **d6604 - Assisting others in nutrition** - Assisting household members and others with their nutrition, such as by helping them to prepare and eat meals.

Per quanto riguarda le funzioni corporee si è scelto di considerare un livello di classificazione ridotto a19 tipologie di funzioni corporee di cui nel seguito si riportano codici e titoli.

- **b110-b139** - Global mental functions
- **b140-b189** - Specific mental functions
- **b210-b229** - Seeing and related functions
- **b230-b249** - Hearing and vestibular functions
- **b250-b279** - Additional sensory functions
- **b280-b289** - Pain
- **b410-b429** - Functions of the cardiovascular system
- **b430-b439** - Functions of the haematological and immunological systems
- **b440-b449** - Functions of the respiratory system
- **b450-b469** - Additional functions and sensations of the cardiovascular and respiratory systems
- **b510-b539** - Functions related to the digestive system
- **b540-b559** - Functions related to metabolism and the endocrine system
- **b610-b639** - Urinary functions
- **b640-b679** - Genital and reproductive functions
- **b710-b729** - Functions of the joints and bones
- **b730-b749** - Muscle functions
- **b750-b789** - Movement functions
- **b810-b849** - Functions of the skin
- **b850-b869** - Functions of the hair and nails

Tuttavia, nell'ICF, il livello di raggruppamento scelto per le funzioni corporee è privo di definizioni estese, per cui per ottenere dei documenti più ricchi è necessario estrarre le definizioni testuali relative a livelli di classificazione più fini. A titolo di esempio nell'elenco che segue si riportano codici, titoli e definizioni di tutte le categorie ICF che sono classificate come **b110-b139 - Global mental functions**.

- **b110-b139 - Global mental functions**
  - **b110 - Consciousness functions** - General mental functions of the state of awareness and alertness, including the clarity and continuity of the wakeful state.
  - **b114 - Orientation functions** - General mental functions of knowing and ascertaining one's relation to self, to others, to time and to one's surroundings.
  - **b117 - Intellectual functions** - General mental functions, required to understand and constructively integrate the various mental functions, including all cognitive functions and their development over the life span.
  - **b122 - Global psychosocial functions** - General mental functions, as they develop over the life span, required to understand and constructively integrate the mental functions that lead to the formation of the interpersonal skills needed to establish reciprocal social interactions, in terms of both meaning and purpose.
  - **b126 - Temperament and personality functions** - General mental functions of constitutional disposition of the individual to react in a particular way to situations, including the set of mental characteristics that makes the individual distinct from others.
  - **b130 - Energy and drive functions** - General mental functions of physiological and psychological mechanisms that cause the individual to move towards satisfying specific needs and general goals in a persistent manner.
  - **b134 - Sleep functions** - General mental functions of periodic, reversible and selective physical and mental disengagement from one's immediate environment accompanied by characteristic physiological changes.
  - **b139 - Global mental functions, other specified and unspecified**

È opportuno notare che l'ultima categoria, la **b139** è un contenitore generico per tutto ciò che non rientra nelle categorie precedenti e che questo tipo di categorie, peraltro prive di definizione estesa, sono state escluse dalle successive elaborazioni.

Per le successive elaborazioni si è quindi assunto come descrizione testuale di ciascuna delle macro categorie **b***-b***** elencate in precedenza, l'insieme di tutte le definizioni ed i titoli delle sotto categorie che la compongono, ad esclusione di quelle prive di definizione, citate nel precedente paragrafo.

In alcuni casi si è scelto di ampliare ulteriormente la descrizione testuale associata a ciascuna macro categoria **b***-b***** usando le cosiddette *inclusion strings* dell'ICF che ampliano le definizioni delle categorie ICF fornendo un elenco di termini collegati. Nel seguito si elencano codici, titoli ed inclusioni per le categorie classificate come **b110-b139 - Global mental functions.**

- **b110-b139 - Global mental functions**
  - **b110 - Consciousness functions** - functions of the state, continuity and quality of consciousness; loss of consciousness, coma, vegetative states, fugues, trance states, possession states, drug-induced altered consciousness, delirium, stupor
  - **b114 - Orientation functions** - functions of orientation to time, place and person; orientation to self and others; disorientation to time, place and person
  - **b117 - Intellectual functions** - functions of intellectual growth; intellectual retardation, mental retardation, dementia
  - **b122 - Global psychosocial functions** - such as in autism
  - **b126 - Temperament and personality functions** - functions of extraversion, introversion, agreeableness, conscientiousness, psychic and emotional stability, and openness to experience; optimism; novelty seeking; confidence; trustworthiness
  - **b130 - Energy and drive functions** - functions of energy level, motivation, appetite, craving (including craving for substances that can be abused), and impulse control
  - **b134 - Sleep functions** - functions of amount of sleeping, and onset, maintenance and quality of sleep; functions involving the sleep cycle, such as in insomnia, hypersomnia and narcolepsy

4.5 - I corpora usati per le elaborazioni

Una volta definite le principali risorse testuali di interesse è possibile formulare in modo più specifico il problema di base che le applicazioni intendono risolvere che può essere rappresentato con la seguente domanda.

*Quali sono le funzioni corporee, scelte tra quelle definite dalle macro categorie b***-b*** di ICF, che sono maggiormente coinvolte in una data attività, scelta tra quelle appartenenti alla categoria ICF d6 ed elencate precedentemente?*

Se si intende risolvere il problema usando la già illustrata tecnologia delle query di somiglianza in spazi vettoriali, è opportuno distinguere due tipologie di corpora testuali:

- **Il corpus di istruzione:** si tratta del corpus sulla base del quale si definisce il vocabolario funzionale alla trasformazione di tutte le risorse testuali in vettori.
- **Il corpus dei soggetti delle query:** si tratta dei documenti testuali che caratterizzano le categorie d6 di interesse.

Indipendentemente dalla loro funzione e dal contesto di utilizzo, tutti i corpora utilizzati per le elaborazioni hanno delle **caratteristiche comuni**, elencate nel seguito.

- Ogni corpus è assimilabile ad una lista di documenti.
- Ogni documento del corpus è assimilabile ad una lista di termini.
- I termini all'interno di un documento sono riportati nell'ordine in cui si possono leggere nella risorsa testuale di origine.
- Sono esclusi da ogni documento sia le stopwords sia la punteggiatura.
- I termini contenuti nei documenti sono temi (*stemming* effettuato con algoritmo definito in [5]).

- Ciascuna categoria ICF rappresentata nel corpus, è descritta da un singolo documento che contiene la lista dei termini estratti dal relativo titolo, dalla definizione e dalla eventuale inclusione.

Indipendentemente dalla loro funzione e dal contesto di elaborazione, i corpora utilizzati **si differenziano tra loro in base ad alcune caratteristiche**, elencate nel seguito.

- Le risorse testuali di origine di ciascun documento che possono essere:
  - titolo, definizione ed eventuale inclusione di una categoria ICF;
  - definizione di un lemma attinto da una risorsa esterna (lemmatizzazione con wordnet[25], risorsa esterna costituita da DBpedia[17]). A tale proposito è opportuno notare che:
    - la definizione di un lemma può non essere presente nel database esterno consultato ed in tal caso la definizione viene trascurata;
    - la definizione esterna di un lemma viene considerata una volta per ogni frase in cui compare, anche se la radice che ha originato il lemma compare più volte nella frase stessa.
- Le modalità di organizzazione dei documenti nel corpus.

Nel seguito, con riferimento al precedente elenco, si elencano i diversi corpora utilizzati sottolineandone le caratteristiche particolari. Nel seguito ci si riferirà a ciascuno di essi con la sigla sottolineata e tra parentesi presente nel titolo del relativo paragrafo.

4.5.1 - Corpus soggetti: titoli e definizioni d6 (d6)

- Corpus con 26 documenti, ciascuno dei quali corrisponde ad una categoria d6.
- Ciascuna categoria è descritta dal suo titolo e dalla sua definizione.
- Corpus con 386 termini in totale.
- Dizionario con 236 termini.

4.5.2 - Corpus istruzione: titoli e definizioni b (bg_noinc)

- Corpus con 19 documenti, ciascuno dei quali corrisponde ad una macro categoria b***-b***.
- Ciascuna categoria è descritta dal suo titolo, dai titoli delle categorie b afferenti e dalle relative definizioni.
- Corpus con 522 termini in totale.
- Dizionario con 379 termini.

4.5.3 - Corpus istruzione: titoli, definizioni e inclusioni b (bg_inc)

- Corpus con 19 documenti, ciascuno dei quali corrisponde ad una macro categoria b***-b***.
- Ciascuna categoria è descritta dal suo titolo, dai titoli delle categorie b afferenti e dalle relative definizioni ed inclusioni.
- Corpus con 1128 termini in totale.
- Dizionario con 852 termini.

4.5.4 - Corpus istruzione: titoli, definizioni ed inclusioni b con lemmi DBpedia (solo titoli) in documenti indipendenti (bgdbptits_inc)

- Corpus con 98 documenti:
  - 19 dei quali corrispondono alle macro categorie ICF b***-b***. Ciascuna categoria è descritta dal suo titolo, dai titoli delle categorie b afferenti e dalle relative definizioni ed inclusioni.
  - I restanti 79 corrispondono alle definizioni DBpedia dei diversi lemmi contenuti nei titoli delle macro categorie ICF ed in quelli delle categorie afferenti.
- Corpus con 6551 termini in totale.
- Dizionario con 2469 termini.

[25] http://www.nltk.org/api/nltk.stem.html#module-nltk.stem.wordnet

4.5.5 - Corpus istruzione: titoli, definizioni ed inclusioni b con lemmi DBpedia in documenti indipendenti (bgdbpstrs_inc)

- Corpus con 98 documenti:
  - 19 dei quali corrispondono alle macro categorie ICF b***-b***. Ciascuna categoria è descritta dal suo titolo, dai titoli delle categorie b afferenti e dalle relative definizioni ed inclusioni.
  - I restanti 79 corrispondono alle definizioni DBpedia dei diversi lemmi contenuti nei titoli, delle macro categorie ICF e nei titoli, nelle definizioni e nelle inclusioni delle categorie afferenti.
- Corpus con 27750 termini in totale.
- Dizionario con 6337 termini.

4.6 - I modelli utilizzati per le elaborazioni

Nelle elaborazioni sono stati considerati diversi modelli di rappresentazione vettoriale[26] dei testi che sono brevemente presentati nei paragrafi che seguono.

4.6.1 - TF-IDF

Il modello TF-IDF[27](Term Frequency- Inverse Document Frequency) [6] rappresenta ciascun termine presente in un documento di un corpus con un numero proporzionale al numero di volte che una parola compare nel documento stesso e inversamente proporzionale alla frequenza della parola nell'intero corpus. Ciò tiene conto del fatto che, da un lato una parola che si ripete in un documento tende a caratterizzarlo, ma che, d'altra parte, le parole più comuni in tutti i documenti del corpus sono quelle meno significative per stabilire le differenze tra i diversi documenti. Le dimensioni dello spazio vettoriale sono analoghe a quelle del modello BOW. Il modello TF-IDF, anche in questo caso come quello BOW, è costruito a partire da un corpus testuale senza che sia necessario specificare grandezze aggiuntive.

4.6.2 - LSI

Il modello di rappresentazione LSI[28](Latent Semantic Indexing) [7] è un modello di rappresentazione che mira a ridurre le dimensioni dello spazio vettoriale rispetto a quelle che caratterizzano i modelli BOW e TF-IDF. La nuova rappresentazione si ottiene per via matematica a partire dalle rappresentazioni vettoriali TF-IDF o BOW, utilizzando la tecnica nota come decomposizione ai valori singolari, detta anche SVD[29](dall'acronimo inglese Singular Value Decomposition). La riduzione delle dimensioni dello spazio vettoriale si basa sulla possibilità di identificare dei *pattern* (che possono essere associati a concetti, contesti, argomenti, temi descritti nel corpus) nei rapporti tra i termini e dei concetti da essi rappresentati nel corpus. In altre parole LSI si basa sul principio che le parole che vengono utilizzate in contesti simili tendono anche ad avere significati simili. L'aggettivo latente è legato quindi alla sua capacità di porre in relazione termini semanticamente correlati anche nel caso in cui alcuni termini siano latenti in un insieme di testi. LSI mira a superare due degli aspetti più problematici nella determinazione della somiglianza tra testi in spazi vettoriali cioè quelli legati a più parole che hanno significati simili (sinonimi) e a parole che hanno più di un significato (polisemia). Per costruire la rappresentazione vettoriale di un corpus testuale secondo il modello LSI è necessario specificare il numero di pattern da identificare. Per corpus di grandi dimensioni è accettato in modo euristico un numero di pattern dell'ordine dei 100-200.

---

[26] http://radimrehurek.com/gensim/tut2.html

[27] http://en.wikipedia.org/wiki/Tf%E2%80%93idf

[28] http://en.wikipedia.org/wiki/Latent_semantic_indexing

[29] http://en.wikipedia.org/wiki/Singular_value_decomposition

4.6.3 - LDA

Il modello di rappresentazione LDA[30](Latent Dirichlet Allocation) [8] si basa sul postulato che ogni documento rappresenta un numero di *argomenti/temi* molto più piccolo del numero di parole in esso presenti e che la comparsa di ogni parola è attribuibile ad uno dei temi del documento. In tal senso ha molte somiglianze con il modello LSI, precedentemente presentato. A differenza dell'LSI però, in questo caso si fanno delle ipotesi sulla distribuzione probabilistica dei temi all'interno del documento. Per costruire la rappresentazione vettoriale di un corpus testuale secondo il modello LDA è necessario specificare il numero di temi da identificare.

4.7 - Post-processing

Durante la fase di post-processing si devono gestire i risultati delle elaborazioni che variano in funzione dei seguenti parametri[31]:

(1) la categoria d6 di interesse;

(2) il corpus di istruzione utilizzato per effettuare la query di somiglianza;

(3) il modello di rappresentazione vettoriale utilizzato per effettuare la query di somiglianza.

In particolare per ogni combinazione di (1), (2) e (3) il risultato dell'elaborazione consiste nell'elenco delle macro categorie b***-b***, ordinato in base al punteggio ottenuto nella query di somiglianza.

Le query restituiscono un numero reale per ogni associazione d6/b***-b*** dove tale "punteggio" (che può essere o non essere normalizzato all'unità) rappresenta un indice di somiglianza.

Il valore di tale indice dipende dal modello di rappresentazione adottato e l'algoritmo secondo il quale esso viene attribuito può essere descritto in modo relativamente semplice solo per i modelli di rappresentazione più semplici. Ad esempio nel caso del modello TF-IDF, la somiglianza tra documenti si basa sulla effettiva compresenza di uno o più termini nei documenti stessi, per cui la query di somiglianza restituirà un punteggio nullo per tutte le combinazioni tra documenti che non condividono termini, mentre sarà maggiore di zero se i documenti condividono anche solo un termine, pure se molto diffuso nel corpus e che porta di conseguenza ad un punteggio relativamente basso.

Al contrario, gli altri modelli di rappresentazione considerati non restituiscono punteggi che permettono una netta separazione tra documenti somiglianti e non somiglianti in quanto gli indici di somiglianza assumono un continuo di valori tra il valore massimo ed il minimo, generalmente non nullo.

In considerazione di ciò, e tenendo conto che si desidera in prima istanza determinare, nel modo più semplice possibile, una soglia di indice di somiglianza oltre la quale due documenti potessero essere considerati simili, si è scelto di fissare tale soglia ad una frazione del valore massimo determinato. Per le elaborazioni fino ad ora effettuate sono state utilizzate frazioni pari a 1/10 e 1/100del massimo indice di somiglianza ottenuto (le tabelle riportate in appendice si riferiscono al secondo dei due valori).

5 - Risultati ed approfondimenti

5.1 - La rappresentazione della attività ICF d6300

Per presentare alcuni esempi di risultati e discuterli si è scelto di considerare una attività specifica, cioè la**d6300–preparing simple meals** perché si è giudicato che in tal caso vi fossero delle associazioni con funzioni corporee molto evidenti ed altre da escludere in modo abbastanza oggettivo.

---

[30]http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[31] In teoria si dovrebbe considerare anche la possibilità di rappresentare le categorie d6 con diverse tipologie di documenti (considerando, in altre parole, diversi corpora dei soggetti delle query) ma tale possibilità non è stata considerata in questo lavoro.

La stringa che la rappresenta durante le elaborazioni è la seguente.

Organizing, cooking and serving meals with a small number of ingredients that require easy methods of preparation and serving, such as making a snack or small meal, and transforming food ingredients by cutting and stirring, boiling and heating food such as rice or potatoes.

Una volta eliminate le stopwords e la punteggiatura e applicato lo stemming, tale stringa è ricondotta al seguente insieme ditemi (28 in tutto):

prepar, simpl, organ, cook, serv, meal, small, number, ingredi, requir, easi, method, prepar, serv, make, snack, small, meal, transform, food, ingredi, cut, stir, boil, heat, food, rice, potato

Avendo considerato diversi corpora di istruzione, la rappresentazione di questi temi in forma vettoriale sarà caratterizzata da tanti più elementi quanto più ricco è il vocabolario del corpus considerato. Nei seguenti paragrafi si descrivono le rappresentazioni vettoriali in formato BOW e TF_IDF della categoria d6300 secondo i vocabolari definiti dai diversi corpora testuali considerati. Per facilitare la comprensione si è sostituito il tema per esteso all'indice che lo rappresenta per cui, in formato BOW, la notazione `(food,2)` significa che il tema 'food' compare due volte mentre in formato TF-IDF, la notazione `(food,0.88)` significa che al tema 'food' è stato assegnato il punteggio 0.88 in base alla sua molteplicità nel documento in cui compare e alla sua frequenza nell'intero corpus considerato. Non si riporta invece la rappresentazione secondo i modelli LSA e LDA, sia perché sono in genere rappresentazioni meno compatte, in quanto molti più elementi del vettore assumono punteggi non nulli, sia perché sono rappresentazioni di non immediata interpretazione.

5.1.1 - bg_noinc

d63100-BOW: [(make, 1),(requir, 1),(food, 2)]

d63100-TFIDF:[(make,0.44),(requir,0.20),(food,0.88)]

5.1.2 - bginc

d63100-BOW: [(make, 1),(requir, 1),(organ, 1),(simpl, 1),(food, 2),
(small, 2)]

d63100-TFIDF:[(make,0.35),(requir,0.16),(organ,0.35),(simpl,0.26),(food,0.69),
(small,0.43)]

5.1.3 - bg_dbptits_inc

d63100-BOW: [(make, 1),(requir, 1),(organ, 1),(simpl, 1),(food, 2),
(small, 2),(serv, 2),(heat, 1),(transform, 1),(number, 1),(cut, 1)]

d63100-TFIDF:[(make,0.23),(requir,0.17),(organ,0.11),(simpl,0.26),(food,0.39),
(small,0.37),(serv,0.39),(heat,0.32),(transform,0.32),(number,0.24),(cut,0.37)]

5.1.4 - bg_dbpstrs_inc

d63100-BOW: [(make, 1),(requir, 1),(organ, 1),(simpl, 1),(food, 2),
(small, 2),(number, 1),(serv, 2),(cut, 1),(prepar, 2),(heat, 1),
(transform, 1),(method, 1),(meal, 2),(easi, 1),(rice, 1)]

d63100-TFIDF:[(make,0.08),(requir,0.06),(organ,0.04),(simpl,0.12),(food,0.17),
(small,0.16),(number,0.06),(serv,0.22),(cut,0.22),(prepar,0.50),(heat,0.15),
(transform,0.20),(method,0.12),(meal,0.59),(easi,0.25),(rice,0.29)]

5.2 - I risultati delle elaborazioni

I risultati delle elaborazioni sono visualizzati in tabelle in cui, per ogni combinazione di modello e corpus di istruzione considerati, si fornisce un elenco di macro categorie b***-b***, ordinato in base al punteggio ottenuto nella query di somiglianza. Poiché il codice ICF di ciascuna macro categoria è composto da 9 caratteri, per una maggiore chiarezza di lettura che richiede una compattazione delle tabelle, il codice viene sostituito da un ID numerico che va da 1 a 19. L'associazione tra codice ICF e ID numerico è riportata in Tab. 1.

Tab. 1: codifica delle macro categorie delle funzioni corporee ICF.

ID	CODICE ICF	TITOLO
1	b110-b139	Global mentalfunctions
2	b140-b189	Specificmentalfunctions
3	b210-b229	Seeing and related functions
4	b230-b249	Hearing and vestibular functions
5	b250-b279	Additional sensory functions
6	b280-b289	Pain
7	b410-b429	Functions of the cardiovascular system
8	b430-b439	Functions of the haematological and immunological systems
9	b440-b449	Functions of the respiratory system
10	b450-b469	Additional functions and sensations of the cardiovascular .....
11	b510-b539	Functions related to the digestive system
12	b540-b559	Functions related to metabolism and the endocrine system
13	b610-b639	Urinary functions
14	b640-b679	Genital and reproductive functions
15	b710-b729	Functions of the joints and bones
16	b730-b749	Muscle functions
17	b750-b789	Movement functions
18	b810-b849	Functions of the skin
19	b850-b869	Functions of the hair and nails

Nelle successiva Tab. 2 si riportano i risultati delle elaborazioni ottenuti considerando una soglia (relativa) di punteggio restituito dalla query di somiglianza oltre le quali considerare 'simili' due documenti pari ad 1/100 del punteggio relativo alla macro-categoria più simile. In particolare, la prima riga di Tab. 2 (e di quelle riportate in appendice) descrive la *soluzione di riferimento*, costituita dagli ID (rif. Tab. 1) che individuano le funzioni corporee associate alla attività d6300, nell'ambito di un precedente lavoro [2], mediante processo logico deduttivo. Nelle righe successive (una per ogni combinazione di corpus di istruzione/modello di rappresentazione) sono riportati gli ID delle funzioni corporee associate dalle applicazioni alla attività d6300. Per facilitare la lettura sono evidenziate tutte le celle in cui compare l'ID di una categoria che fa parte della soluzione di riferimento. In tabella sono riportate tutte le funzioni corporee, ordinate da destra a sinistra secondo il punteggio restituito dalla query di somiglianza.

Tab. 2: risultati delle elaborazioni per la categoria d6300-preparing simple meals..

d6300																						
sogg-istr	**mod.**	**max**																				
selezionate manualmente			*1*	*2*	*3*	*4*	*8*	*11*	*12*	*15*	*16*	*17*										
d6-bg_noinc	tf-idf	3.50E-01	*11*	*1*	7	10	*16*	*2*			*3*	*4*	5	6	*8*	9	*12*	13	14	*15*	*17*	18	19
d6-bg_inc		3.80E-01	*11*	*1*	*2*	*16*	*15*	14	*17*	10	7			*3*	*4*	5	6	*8*	9	*12*	13	18	19
d6-bgdbptits_inc		1.30E-01	*11*	*1*	*16*	*15*	*2*	14	10	*17*	7			*3*	*4*	5	6	*8*	9	*12*	13	18	19
d6-bgdbpstrs_inc		4.70E-02	*11*	*16*	*1*	*2*	*15*	*17*	14	10	7			*3*	*4*	5	6	*8*	9	*12*	13	18	19
d6-bg_noinc	lsi	7.20E-01	*11*	*1*	10	7	*16*	*2*		*12*	*8*	18	19	9	14	13		*3*	*4*	5	6	*15*	*17*
d6-bg_inc		8.40E-01	*11*	*15*	14	*1*	10	*16*	*2*	7	*17*		*3*	*8*	5	19	*12*	6		*4*	9	13	18
d6-bgdbptits_inc		6.40E-01	*11*	6	*15*	14	10	*1*	*16*	*2*	7	*17*		*8*	5	*12*	18	9	*3*		*4*	13	19
d6-bgdbpstrs_inc		5.20E-01	*11*	10	*1*	5	*12*	19	18	14	*8*	*2*	*3*	*17*	*4*	6	13	7	*16*		9		13
d6-bg_noinc	lda	7.10E-01	*1*	5	*11*	*15*			*2*	*3*	*4*	6	7	*8*	9	10	*12*	13	14	*16*	*17*	18	19
d6-bg_inc		6.10E-01	*16*	*2*	*11*	*1*			*3*	*4*	5	6	7	*8*	9	10	*12*	13	14	*15*	*17*	18	19
d6-bgdbptits_inc		4.00E-01	*8*			*1*	*2*	*3*	*4*	5	6	7	9	10	*11*	*12*	13	14	*15*	*16*	*17*	18	19
d6-bgdbpstrs_inc		4.60E-01	6	*16*	*17*	*12*	18	*11*			*1*	*2*	*3*	*4*	5	7	*8*	9	10	13	14	*15*	19

Su ciascuna riga di Tab. 2, la cella bianca e vuota delimita le funzioni corporee per le quali la query ha restituito un punteggio superiore alla soglia pari ad 1/100 del massimo punteggio ottenuto (a sinistra della cella bianca) da quelle per le quali la query ha restituito un punteggio minore di tale livello (a destra della cella bianca). Su ciascuna riga, la cella nera e vuota delimita le funzioni corporee per le quali la query ha restituito un punteggio diverso da 0 (a sinistra della cella nera) da quelle per le quali la query ha restituito un punteggio nullo (a destra della cella nera).

In Tab. 3 si riporta il numero di volte che ciascuna funzione corporea è stata associata dalle applicazioni all'attività d6300, limitandosi a quelle associate ad un punteggio maggiore della soglia scelta. In base a tale tabella si osserva quanto segue.

- Le categorie con ID 1, 2, 11, 16 ed in misura minore 16 e 17 possono essere definite **veri positivi**, cioè sono categorie che fanno parte della soluzione di riferimento che sono state associate in modo corretto dalle applicazioni nella maggior parte dei casi.
- Le categorie con ID 3, 4, 8, 12 possono essere definite **falsi negativi** perché, pur facendo parte della soluzione di riferimento, nella maggior parte dei casi non sono state associate dalle applicazioni alla attività d6300.
- Le categorie con ID 7, 10 ed in misura minore 14 possono essere definite come **falsi positivi** perché pur non facendo parte della soluzione di riferimento sono state associate alla attività d6300 dalle applicazioni in un numero relativamente alto di casi.
- Le categorie con ID 5, 6, 9, 13, 18, 19 possono essere infine considerate come **veri negativi**.

Tab. 3: conteggio delle categorie b***-b*** risultati delle elaborazioni per la categoria d6300 Preparing simple meals.

ID	CODICE ICF	TITOLO	N (1/100)
1	b110-b139	Global mental functions	10
2	b140-b189	Specific mental functions	9
3	b210-b229	Seeing and related functions	1
4	b230-b249	Hearing and vestibular functions	1
5	b250-b279	Additional sensory functions	2
6	b280-b289	Pain	3
7	b410-b429	Functions of the cardiovascular system	8
8	b430-b439	Functions of the haematological and immunological systems	2
9	b440-b449	Functions of the respiratory system	0
10	b450-b469	Additional functions and sensations of the cardiovascular and	8
11	b510-b539	Functions related to the digestive system	11
12	b540-b559	Functions related to metabolism and the endocrine system	2
13	b610-b639	Urinary functions	1
14	b640-b679	Genital and reproductive functions	6
15	b710-b729	Functions of the joints and bones	6
16	b730-b749	Muscle functions	10
17	b750-b789	Movement functions	7
18	b810-b849	Functions of the skin	2
19	b850-b869	Functions of the hair and nails	1

5.3 - Analisi dei risultati ottenuti con il modello TF-IDF

Nei seguenti paragrafi si forniscono alcuni approfondimenti relativi alle associazioni frutto delle elaborazioni. In ciascuno di essi ci si riferisce alla soluzione della query di somiglianza che si basa sul corpus di istruzione citato nel titolo, dove, per semplicità e brevità, si considererà solo il modello TF-IDF. In Tab.4 si riporta

un estratto di Tab. 2, in cui ci si limita al modello TF-IDF e alle funzioni corporee effettivamente individuate dalla query di somiglianza.

Tab. 4: risultati delle elaborazioni per la categoria d6300, modello tf-idf

d6300 tf-idf											
sogg-istr	**max**										
selezionate manualmente		*1*	*2*	*3*	*4*	*8*	*11*	*12*	*15*	*16*	*17*
d6-bg_noinc	3.50E-01	*11*	*1*	7	10	*16*	*2*				
d6-bg_inc	3.80E-01	*11*	*1*	*2*	*16*	*15*	14	*17*	10	7	
d6-bgdbptits_inc	1.30E-01	*11*	*1*	*16*	*15*	*2*	14	10	*17*	7	
d6-bgdbpstrs_inc	4.70E-02	*11*	*16*	*1*	*2*	*15*	*17*	14	10	7	

In ciascuno dei successivi paragrafi si riporta inizialmente la rappresentazione TF-IDF associata alla attività d6300 e in seguito, per ogni funzione corporea ICF, si riporteranno i soli termini in comune con la descrizione della attività d6300 di cui sopra. Infine, al termine del paragrafo si riportano alcune note esplicative.

5.3.1 - Vocabolario bg_noinc

d6300-TFIDF: [(make,0.44),(requir,0.20),(food,0.88)]

ID=1 - b110-b139-TFIDF : [(make,0.74),(requir,0.67)]

ID=2 - b140-b189-TFIDF : [(requir,1.00)]

ID=3 - b210-b229-TFIDF : []

ID=4 - b230-b249-TFIDF : []

ID=5 - b250-b279-TFIDF : []

ID=6 - b280-b289-TFIDF : []

ID=7 - b410-b429-TFIDF : [(requir,1.00)]

ID=8 - b430-b439-TFIDF : []

ID=9 - b440-b449-TFIDF : []

ID=10 - b450-b469-TFIDF : [(requir,1.00)]

ID=11 - b510-b539-TFIDF : [(food,1.00)]

ID=12 - b540-b559-TFIDF : []

ID=13 - b610-b639-TFIDF : []

ID=14 - b640-b679-TFIDF : []

ID=15 - b710-b729-TFIDF : []

ID=16 - b730-b749-TFIDF : [(requir,1.00)]

ID=17 - b750-b789-TFIDF : []

ID=18 - b810-b849-TFIDF : []

ID=19 - b850-b869-TFIDF : []

- I termini che si ritrovano nella descrizione di d300 ed in almeno una descrizione delle funzioni corporee del corpus di istruzione considerato sono `food`, `make` e `requir`.
- I veri positivi sono stati associati in base al termine specifico `food` oppure alla presenza di `make` e `requir` che sono temi meno specifici.
- I falsi positivi si manifestano a causa del tema `requir`.

5.3.2 - Vocabolario bg_inc

d6300-TFIDF: [(make,0.35),(requir,0.16),(organ,0.35),(simpl,0.26),(food,0.69),(small,0.43)]

ID=1 - b110-b139-TFIDF : [(make,0.74),(requir,0.67)]

ID=2 - b140-b189-TFIDF : [(requir,0.34),(organ,0.75),(simpl,0.57)]

ID=3 - b210-b229-TFIDF : []

ID=4 - b230-b249-TFIDF : []

ID=5 - b250-b279-TFIDF : []

ID=6 - b280-b289-TFIDF : []

ID=7 - b410-b429-TFIDF : [(requir,1.00)]

ID=8 - b430-b439-TFIDF : []

ID=9 - b440-b449-TFIDF : []

ID=10 - b450-b469-TFIDF : [(requir,1.00)]

ID=11 - b510-b539-TFIDF : [(food,1.00)]

ID=12 - b540-b559-TFIDF : []

ID=13 - b610-b639-TFIDF : []

ID=14 - b640-b679-TFIDF : [(small,1.00)]

ID=15 - b710-b729-TFIDF : [(small,1.00)]

ID=16 - b730-b749-TFIDF : [(requir,0.59),(small,0.81)]

ID=17 - b750-b789-TFIDF : [(simpl,1.00)]

ID=18 - b810-b849-TFIDF : []

ID=19 - b850-b869-TFIDF : []

- I nuovi termini che si ritrovano nella descrizione di d300 ed in almeno una descrizione delle funzioni corporee del corpus di istruzione considerato (termini sovrapposti nel seguito) sono `organ, simpl` e `small`.
- I nuovi termini di cui soprahanno portato a nuove associazioni rispetto al caso precedente ed in particolare ad un nuovo falso positivo (ID=14) e a due nuovi veri positivi (ID=15 e ID=17)

5.3.3 - Vocabolario bgdbptits_inc

d6300-TFIDF:
[(make,0.23),(requir,0.17),(organ,0.11),(simpl,0.26),(food,0.39),(small,0.37),(serv,0.39),(heat,0.32),(transform,0.32),(number,0.24),(cut,0.37)]

ID=1 - b110-b139-TFIDF : [(make,0.55),(requir,0.83)]

ID=2 - b140-b189-TFIDF : [(requir,0.52),(organ,0.34),(simpl,0.79)]

ID=3 - b210-b229-TFIDF : []

ID=4 - b230-b249-TFIDF : []

ID=5 - b250-b279-TFIDF : []

ID=6 - b280-b289-TFIDF : []

ID=7 - b410-b429-TFIDF : [(requir,1.00)]

ID=8 - b430-b439-TFIDF : []

ID=9 - b440-b449-TFIDF : []

ID=10 - b450-b469-TFIDF : [(requir,1.00)]

ID=11 - b510-b539-TFIDF : [(food,1.00)]

ID=12 - b540-b559-TFIDF : []

ID=13 - b610-b639-TFIDF : []

ID=14 - b640-b679-TFIDF : [(small,1.00)]

ID=15 - b710-b729-TFIDF : [(small,1.00)]

ID=16 - b730-b749-TFIDF : [(requir,0.68),(small,0.74)]

ID=17 - b750-b789-TFIDF : [(simpl,1.00)]

ID=18 - b810-b849-TFIDF : []

ID=19 - b850-b869-TFIDF : []

- Da notare che l'incremento di termini nel vocabolario non ha portato a nuovi termini sovrapposti rispetto al caso precedente.
- Da notare che il numero di termini della descrizione d6300 che compaiono nel corpus di istruzione è salito a 11. I 5 termini non sovrapposti sono contenuti nei documenti del corpus di istruzione diversi dalle descrizioni delle funzioni corporee. In questo particolare caso si tratta dei 79 documenti che contengono le definizioni DBpedia dei diversi lemmi contenuti nei titoli delle macro categorie ICF che descrivono le funzioni corporee ed in quelli delle categorie afferenti.

5.3.4 - Vocabolario bg_dbpstrs_inc

d6300-TFIDF:
[(make,0.08),(requir,0.06),(organ,0.04),(simpl,0.12),(food,0.17),(small,0.16),(number,0.06),(serv,0.22),(cut,0.22),(prepar,0.50),(heat,0.15),(transform,0.20),(method,0.12),(meal,0.59),(easi,0.25),(rice,0.29)]

ID=1 - b110-b139-TFIDF : [(make,0.55),(requir,0.83)]

ID=2 - b140-b189-TFIDF : [(requir,0.44),(organ,0.27),(simpl,0.86)]

ID=3 - b210-b229-TFIDF : []

ID=4 - b230-b249-TFIDF : []

ID=5 - b250-b279-TFIDF : []

ID=6 - b280-b289-TFIDF : []

ID=7 - b410-b429-TFIDF : [(requir,1.00)]

ID=8 - b430-b439-TFIDF : []

ID=9 - b440-b449-TFIDF : []

ID=10 - b450-b469-TFIDF : [(requir,1.00)]

ID=11 - b510-b539-TFIDF : [(food,1.00)]

ID=12 - b540-b559-TFIDF : []

ID=13 - b610-b639-TFIDF : []

ID=14 - b640-b679-TFIDF : [(small,1.00)]

ID=15 - b710-b729-TFIDF : [(small,1.00)]

ID=16 - b730-b749-TFIDF : [(requir,0.60),(small,0.80)]

ID=17 - b750-b789-TFIDF : [(simpl,1.00)]

ID=18 - b810-b849-TFIDF : []

ID=19 - b850-b869-TFIDF : []

- Da notare che l'incremento di termini nel vocabolario non ha portato a nuovi termini sovrapposti rispetto al caso precedente.
- Da notare che il numero di termini della descrizione d6300 che compaiono nel corpus di istruzione è salito a 16. I 10 termini non sovrapposti sono contenuti nei documenti del corpus di istruzione diversi dalle descrizioni delle funzioni corporee. In questo particolare caso si tratta dei 79 documenti che contengono le definizioni DBpedia dei diversi lemmi contenuti nei titoli delle macro categorie ICF che descrivono le funzioni corporee ed in quelli delle categorie afferenti.

6 - Commenti ai risultati e sviluppi futuri

Gli approfondimenti del paragrafo 5.3 -, pur se riferiti al modello di rappresentazione più basilare, permettono alcune osservazioni che forniscono preziose indicazioni per i possibili futuri sviluppi.

Innanzi tutto, anche considerando il corpus di istruzione più ampio tra quelli utilizzati, il numero totale di termini sovrapposti è pari a 16, sui 28 totali di cui è composta la descrizione della attività d6300. Di conseguenza, la prima indicazione che si raccoglie è quella di utilizzare corpus di istruzione ancora (molto) più ampi.

In secondo luogo si osserva che dei 6 termini sovrapposti, che sono `food`, `make`, `organ`, `requir`, `simple`, `small`, solo alcuni possono essere considerati specifici dell'attività d6300. Ad esempio il termine `small` è molto generico ed infatti è responsabile di numerosi falsi positivi. La indicazione che si raccoglie è quindi di filtrare il corpus dei soggetti in modo che sia composto da soli termini specifici.

Le due precedenti osservazioni sono coerenti se si pensa al paragone del sistema di applicazioni creato con un generico motore di ricerca che una volta fissato l'algoritmo di elaborazione, fornirà risultati tanto migliori quanti più documenti sono indicizzati e quanto più il testo che istruisce la query è specifico.

In particolare alcune ipotesi concrete di ulteriori sviluppi riguardano gli aspetti elencati nel seguito.

- Restringimento corpus dei soggetti ai soli termini specifici (in prima istanza quelli contenuti nei soli titoli della categoria).
- Arricchimento corpus di istruzione da altre fonti rispetto a DBpedia, sia generiche (wordnet) sia specifiche (ad esempio database usati in ambiente medico/sanitario).
- Aggiunta al corpus di istruzione esclusivamente di lemmi relativi a specifiche parti del discorso (verbi o sostantivi, pos_tagging[32]).

[32] http://www.nltk.org/book/ch05.html

- Utilizzo di corpus di istruzione molto grandi (anche il database completo di Wikipedia[33]).
- Approfondimento degli aspetti legati al punteggio restituito dalle query di somiglianza ed in particolare alla scelta di una soglia basata su criteri più specifici rispetto a quelli utilizzati. Ad esempio lo studio della distribuzione dei punteggi restituiti dalle query e il tentativo di individuare delle discontinuità o più in generale trarne elementi utili.
- Utilizzo di un diverso livello di aggregazione dei documenti che rappresentano le categorie ICF, basandosi ad esempio:
  - sul livello ICF con maggior dettaglio (singole categorie ICF b****)
  - sul livello ICF con minor dettaglio (ad esempio b1***-mentalfunctions, b2***sensorialfunction).
- Studio approfondito del funzionamento dei modelli di rappresentazione in grado di evidenziare associazioni latenti (LSA e LDA) basandosi in particolare sull'associazione con la funzione corporea ICF legata alla vista (ID=3, b210-b229).

Bibliografia

[1] International Classification of Functioning, Disability and Health (ICF)http://www.who.int/classifications/icf/en/

[2] L.Burzagli, P.L.Emiliani, N.Zoppetti, "Formalizzazione dei dati, delle informazione e della conoscenza relativa al design for all", Deliverable 1.24 del Progetto D4All Sw integration and advanced Human Machine Interfaces in design for Ambient Assisted Living, Dicembre 2014

[3] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python" O'Reilly Media, ISBN 978-0-596-51649-9

[4] R. Řehůřek and P.Sojka "Software Framework for Topic Modelling with Large Corpora" Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp45-50, Valletta, Malta May 2010 http://is.muni.cz/publication/884893/en

[5] M.F.Porter, "An Algorithm for Suffix Stripping", Program, **14**(3): 130–137 (1980)

[6] S.Robertson "Understanding inverse document frequency: On theoretical arguments for IDF". Journal of Documentation **60** (5): 503–520 (2004). doi:10.1108/00220410410560582

[7] S.Deerwester et al, "Improving Information Retrieval with Latent Semantic Indexing", Proceedings of the 51st Annual Meeting of the American Society for Information Science, (1988), pp. 36–40.

[8] D.Blei, et al. "Latent Dirichlet allocation". Journal of Machine Learning Research **3**, pp. 993–1022. (2003) doi:10.1162/jmlr.2003.3.4-5.993

[33] http://radimrehurek.com/gensim/wiki.html

7 - *Appendice*
Tabelle dei risultati

Di seguito si riportano le tabelle dei risultati relative a tutte le attività ICF considerate (per l'interpretazione si faccia riferimento alla Tab. 2 riportata nel testo ed alle relative spiegazioni; tale tabella è ripetuta per convenienza in questa appendice). Tutte le tabelle si riferiscono alle associazioni effettivamente restituite dalle query di somiglianza (cioè alle funzioni corporee associate a punteggi superiori ad 1/100 del massimo punteggio ottenuto).

d6102																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*8*	*9*	*11*	*12*	*15*	*16*	*17*							
d6-bg_noinc	tf-idf	1.30E-01	1.30E-03	*11*																	
d6-bg_inc		8.20E-02	8.20E-04	*1*	*11*																
d6-bgdbptits_inc		2.20E-02	2.20E-04	*1*	*11*																
d6-bgdbpstrs_inc		1.10E-02	1.10E-04	*1*	*11*																
d6-bg_noinc	lsi	7.40E-01	7.40E-03	*11*																	
d6-bg_inc		6.40E-01	6.40E-03	*1*	*11*																
d6-bgdbptits_inc		2.00E-01	2.00E-03	*1*	*11*	10	*8*														
d6-bgdbpstrs_inc		2.40E-01	2.40E-03	*12*	10	*1*	*4*	7	*16*	14	5	*11*	*8*	*9*	*2*	18	*3*	6	19		
d6-bg_noinc	lda	1.00E+00	1.00E-02	*11*	*15*																
d6-bg_inc		7.10E-01	7.10E-03	*1*	*11*																
d6-bgdbptits_inc		0.00E+00	0.00E+00																		
d6-bgdbpstrs_inc		2.80E-01	2.80E-03	6	*17*	*16*	*12*	18													

d6300																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*8*	*11*	*12*	*15*	*16*	*17*								
d6-bg_noinc	tf-idf	3.50E-01	3.50E-03	*11*	*1*	7	10	*16*	*2*												
d6-bg_inc		3.80E-01	3.80E-03	*11*	*1*	*2*	*16*	*15*	14	*17*	10	7									
d6-bgdbptits_inc		1.30E-01	1.30E-03	*11*	*1*	*16*	*15*	*2*	14	10	*17*	7									
d6-bgdbpstrs_inc		4.70E-02	4.70E-04	*11*	*16*	*1*	*2*	*15*	*17*	14	10	7									
d6-bg_noinc	lsi	7.20E-01	7.20E-03	*11*	*1*	10	7	*16*	*2*												
d6-bg_inc		8.40E-01	8.40E-03	*11*	*15*	14	*1*	10	*16*	*2*	7	*17*									
d6-bgdbptits_inc		6.40E-01	6.40E-03	*11*	6	*15*	14	10	*1*	*16*	*2*	7	*17*								
d6-bgdbpstrs_inc		5.20E-01	5.20E-03	*11*	10	*1*	5	*12*	19	18	14	*8*	*2*	*3*	*17*	*4*	6	13	7	*16*	
d6-bg_noinc	lda	7.10E-01	7.10E-03	*1*	5	*11*	*15*														
d6-bg_inc		6.10E-01	6.10E-03	*16*	*2*	*11*	*1*														
d6-bgdbptits_inc		4.00E-01	4.00E-03	*8*																	
d6-bgdbpstrs_inc		4.60E-01	4.60E-03	6	*16*	*17*	*12*	18	*11*												

d6301																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*11*	*12*	*15*	*16*	*17*						
d6-bg_noinc	tf-idf	2.20E-01	2.20E-03	*11*	*2*	*16*	*1*	*17*	7	10											
d6-bg_inc		2.90E-01	2.90E-03	*11*	*2*	*17*	*15*	*16*	10	*1*	7										
d6-bgdbptits_inc		1.10E-01	1.10E-03	*11*	*2*	*17*	*15*	*16*	10	*1*	7										
d6-bgdbpstrs_inc		5.30E-02	5.30E-04	*11*	*2*	*17*	*16*	*15*	*1*	10	7										
d6-bg_noinc	lsi	6.30E-01	6.30E-03	*11*	*2*	*16*	*1*	10	7	*17*											
d6-bg_inc		7.80E-01	7.80E-03	*11*	*2*	*17*	*15*	10	*1*	*16*	7										
d6-bgdbptits_inc		6.20E-01	6.20E-03	*11*	*2*	*17*	*15*	10	*1*	*16*	7	*6*	9								
d6-bgdbpstrs_inc		4.50E-01	4.50E-03	*11*	*2*	10	*3*	*1*	*17*	19	18	*8*	*5*	14	7	13	*12*	9	*4*	*15*	
d6-bg_noinc	lda	8.40E-01	8.40E-03	*2*	9	*11*	*15*	*16*	*1*	*5*	*17*										
d6-bg_inc		8.50E-01	8.50E-03	*2*	*11*	*16*	*17*	13	*15*												
d6-bgdbptits_inc		7.90E-01	7.90E-03	*2*	*5*	*11*	*4*	*8*	*17*	*6*											
d6-bgdbpstrs_inc		3.00E-01	3.00E-03	*11*	*12*	*6*															

d6400																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*15*	*16*	*17*							
d6-bg_noinc	tf-idf	3.10E-01	3.10E-03	*9*	*3*																
d6-bg_inc		7.90E-02	7.90E-04	*9*	*3*	10	*15*	*16*	*17*												
d6-bgdbptits_inc		7.70E-02	7.70E-04	*3*	*9*	*15*	*16*	10	*17*												
d6-bgdbpstrs_inc		6.30E-02	6.30E-04	*3*	*9*	*16*	*15*	10	*17*												
d6-bg_noinc	lsi	7.70E-01	7.70E-03	*9*	*3*																
d6-bg_inc		6.90E-01	6.90E-03	*9*	*3*	10	*15*	*17*	*16*												
d6-bgdbptits_inc		5.50E-01	5.50E-03	10	*9*	*3*	*15*	*16*	*17*	19	*6*	*8*	*5*	7	12						
d6-bgdbpstrs_inc		2.50E-01	2.50E-03	*3*	*15*	*17*	11	*9*	12	*5*	*2*	*8*	13								
d6-bg_noinc	lda	8.90E-01	8.90E-03	*3*	13	*2*	*9*														
d6-bg_inc		8.10E-01	8.10E-03	*3*	*9*	*16*															
d6-bgdbptits_inc		6.70E-01	6.70E-03	*16*	*3*																
d6-bgdbpstrs_inc		4.10E-02	4.10E-04	*17*	*16*																

d6401																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*10*	*15*	*16*	*17*						
d6-bg_noinc	tf-idf	1.20E-01	1.20E-03	*3*	11																
d6-bg_inc		5.30E-02	5.30E-04	*3*	11																
d6-bgdbptits_inc		3.00E-02	3.00E-04	11	*3*																
d6-bgdbpstrs_inc		1.60E-02	1.60E-04	11	*3*																
d6-bg_noinc	lsi	6.50E-01	6.50E-03	*3*	11																
d6-bg_inc		6.70E-01	6.70E-03	*3*	11																
d6-bgdbptits_inc		1.10E-01	1.10E-03	19	*3*	11	*9*	*4*	13												
d6-bgdbpstrs_inc		7.40E-02	7.40E-04	11	*9*	18	*2*	14	*10*	*5*	*4*	*15*	7								
d6-bg_noinc	lda	7.10E-01	7.10E-03	*3*	11	13	*15*														
d6-bg_inc		7.10E-01	7.10E-03	*3*	11																
d6-bgdbptits_inc		0.00E+00	0.00E+00																		
d6-bgdbpstrs_inc		2.30E-01	2.30E-03	*9*	18	12	11														

d6403																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*10*	*11*	*12*	*15*	*16*	*17*				
d6-bg_noinc	tf-idf	6.80E-02	6.80E-04	*2*																	
d6-bg_inc		7.90E-02	7.90E-04	*2*	*12*																
d6-bgdbptits_inc		9.90E-02	9.90E-04	*12*	*2*																
d6-bgdbpstrs_inc		5.70E-02	5.70E-04	*12*																	
d6-bg_noinc	lsi	7.10E-01	7.10E-03	*2*																	
d6-bg_inc		6.30E-01	6.30E-03	*12*	*2*																
d6-bgdbptits_inc		1.10E-01	1.10E-03	*12*	*2*	*6*	19	*17*	13	*5*											
d6-bgdbpstrs_inc		1.10E-01	1.10E-03	*16*	*3*	*2*	*6*	*12*	7	*17*	*9*	*1*	*8*								
d6-bg_noinc	lda	1.00E+00	1.00E-02	*2*	*9*																
d6-bg_inc		8.90E-01	8.90E-03	*2*	*12*																
d6-bgdbptits_inc		7.90E-01	7.90E-03	*4*																	
d6-bgdbpstrs_inc		1.60E-01	1.60E-03	*11*																	

d6404																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*10*	*11*	*12*	*15*	*16*	*17*				
d6-bg_noinc	tf-idf	3.70E-01	3.70E-03	*11*	*2*	7	*10*	*16*	*1*												
d6-bg_inc		4.50E-01	4.50E-03	*11*	*2*	*10*	*1*	7													
d6-bgdbptits_inc		1.50E-01	1.50E-03	*11*	*2*	*10*	*1*	7	*16*												
d6-bgdbpstrs_inc		1.00E-01	1.00E-03	*11*	*2*	*1*	*10*	7	*16*												
d6-bg_noinc	lsi	7.50E-01	7.50E-03	*11*	*10*	7	*2*	*16*	*1*												
d6-bg_inc		8.50E-01	8.50E-03	*11*	*10*	*1*	*2*	7	*16*												
d6-bgdbptits_inc		7.70E-01	7.70E-03	*11*	*10*	*1*	*2*	*6*	7	*16*											
d6-bgdbpstrs_inc		6.40E-01	6.40E-03	*11*	*10*	*12*	*1*	7	*5*	*3*	19	14	*8*	18	*17*	*2*	*9*	13	*16*		
d6-bg_noinc	lda	8.60E-01	8.60E-03	*11*	*15*	*2*	*9*														
d6-bg_inc		8.60E-01	8.60E-03	*11*	*2*																
d6-bgdbptits_inc		3.40E-01	3.40E-03	*8*	*2*	*5*	*11*	*4*													
d6-bgdbpstrs_inc		2.50E-01	2.50E-03	*12*	*11*																

d6405																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*10*	*11*	*12*	*15*	*16*	*17*				
d6-bg_noinc	tf-idf	8.90E-02	8.90E-04	13	*3*	18	*2*														
d6-bg_inc		8.70E-02	8.70E-04	13	*3*	*2*	*6*	*5*	18												
d6-bgdbptits_inc		8.90E-02	8.90E-04	13	*6*	*3*	*5*	18	*2*												
d6-bgdbpstrs_inc		2.40E-02	2.40E-04	*6*	*5*	13	18	*3*													
d6-bg_noinc	lsi	5.80E-01	5.80E-03	13	*3*	18	*2*														
d6-bg_inc		5.80E-01	5.80E-03	13	*6*	*3*	*5*	18	*2*												
d6-bgdbptits_inc		2.60E-01	2.60E-03	13	*6*	*3*	*5*	18	*2*	19	*17*										
d6-bgdbpstrs_inc		1.60E-01	1.60E-03	*3*	*6*	*16*	*2*	*12*	7	*9*	*10*	*17*	*8*	*1*	13	19	*5*	18			
d6-bg_noinc	lda	8.10E-01	8.10E-03	*3*	13	*2*	*9*	18													
d6-bg_inc		8.10E-01	8.10E-03	*3*	*2*	13	*15*														
d6-bgdbptits_inc		8.10E-01	8.10E-03	*3*	13																
d6-bgdbpstrs_inc		4.20E-01	4.20E-03	*11*	13	*12*	*6*														

d6502																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*10*	*11*	*12*	*15*	*16*	*17*				
d6-bg_noinc	tf-idf	3.60E-01	3.60E-03	18	*11*	7															
d6-bg_inc		2.20E-01	2.20E-03	18	*11*	7															
d6-bgdbptits_inc		2.60E-01	2.60E-03	18	*11*	7															
d6-bgdbpstrs_inc		2.40E-01	2.40E-03	18	*11*	7															
d6-bg_noinc	lsi	7.20E-01	7.20E-03	18	*11*	7															
d6-bg_inc		7.30E-01	7.30E-03	18	*11*	7															
d6-bgdbptits_inc		5.30E-01	5.30E-03	18	19	*11*	*6*	7	*10*												
d6-bgdbpstrs_inc		3.70E-01	3.70E-03	*11*	*12*	7	*1*	14	*16*	*10*	19	18	*9*	*5*	*8*	*2*	*17*	*15*	*6*	*3*	*4*
d6-bg_noinc	lda	8.10E-01	8.10E-03	18	*11*	*15*	7														
d6-bg_inc		7.10E-01	7.10E-03	*11*	18																
d6-bgdbptits_inc		5.70E-01	5.70E-03	18																	
d6-bgdbpstrs_inc		4.20E-01	4.20E-03	*9*	*12*	18	19	*11*													

d6600																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*10*	*11*	*12*	*15*	*16*	*17*				
d6-bg_noinc	tf-idf	1.10E-01	1.10E-03	*1*	18	*11*	14	*8*	*12*	*2*											
d6-bg_inc		8.10E-02	8.10E-04	*1*	18	*11*	14	*2*	*8*	*12*	*3*										
d6-bgdbptits_inc		3.60E-02	3.60E-04	*1*	18	*11*	*2*	14	*8*	*12*	*3*										
d6-bgdbpstrs_inc		1.90E-02	1.90E-04	*2*	*1*	*11*	18	14	*8*												
d6-bg_noinc	lsi	6.00E-01	6.00E-03	*1*	18	*11*	14	*8*	*12*	*2*											
d6-bg_inc		6.80E-01	6.80E-03	*1*	18	14	*11*	*8*	*12*	*3*	*2*										
d6-bgdbptits_inc		1.30E-01	1.30E-03	*1*	18	14	*11*	19	*8*	*6*	*12*	*3*	*2*								
d6-bgdbpstrs_inc		3.00E-01	3.00E-03	*11*	18	14	*12*	*1*	19	*10*	*8*	*4*	*9*	*5*	*16*	*2*	7	*15*	*6*	*17*	
d6-bg_noinc	lda	8.00E-01	8.00E-03	18	*11*	*15*															
d6-bg_inc		8.80E-01	8.80E-03	*1*	*11*	14	*2*														
d6-bgdbptits_inc		2.10E-01	2.10E-03	*2*	*5*	*11*	*8*	*4*													
d6-bgdbpstrs_inc		3.50E-01	3.50E-03	*6*	*16*	*17*	*11*	*10*	*2*	18	*8*										

d6604																					
sogg-istr	**mod.**	**max**	**soglia**																		
selezionate manualmente				*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	*10*	*11*	*12*	*15*	*16*	*17*				
d6-bg_noinc	tf-idf	9.50E-02	9.50E-04	*1*	18	*11*															
d6-bg_inc		7.40E-02	7.40E-04	*1*	18	*11*															
d6-bgdbptits_inc		2.30E-02	2.30E-04	*1*	18	*11*															
d6-bgdbpstrs_inc		1.50E-02	1.50E-04	*11*	*1*	18															
d6-bg_noinc	lsi	5.20E-01	5.20E-03	*1*	18	*11*															
d6-bg_inc		5.90E-01	5.90E-03	*1*	18	*11*															
d6-bgdbptits_inc		9.00E-02	9.00E-04	*1*	18	*11*	*6*	19	*10*												
d6-bgdbpstrs_inc		3.70E-01	3.70E-03	18	*11*	19	14	*10*	*12*	*9*	*1*	*16*	*8*	*5*	*2*	*4*					
d6-bg_noinc	lda	8.90E-01	8.90E-03	18	*11*	*15*															
d6-bg_inc		6.30E-01	6.30E-03	18	*11*	*1*															
d6-bgdbptits_inc		0.00E+00	0.00E+00																		
d6-bgdbpstrs_inc		4.30E-01	4.30E-03	*11*	*6*																