GEOCODING

Metodologia di normalizzazione e georiferimento indirizzi

Consorzio per il Sistema Informativo
www.csipiemonte.it

Sommario

Normalizzazione

Normalizzare un indirizzo significa analizzare e correggere inesattezze o eventuali incongruenze, al fine di ottimizzare le referenze e rendere i record intellegibili ai vari sistemi informativi.
Questa attività tende quindi a "scomporre" una stringa di indirizzo, scritta in modo naturale, e a organizzarla in modo ordinato, andando ulteriormente a intervenire su possibili ambiguità per migliorarne la corretta interpretazione.
Alcune semplici regole permettono di migliorare la qualità dei risultati ottenibili ed hanno un forte impatto su tutte le attività successive: un primo macro requisito per una corretta normalizzazione è che la stringa contenente l'identificativo del comune, la stringa dell'indirizzo e, quando presente, il numero civico, comprensivo di lettere e/o apposizioni, siano separati.
Questi tre attributi (comune, indirizzo e numero civico) vengono analizzati e subiscono una prima normalizzazione di cui queste azioni sono un esempio:

- rimozione degli spazi duplicati;
- rimozione degli spazi iniziali e finali;
- conversione dei caratteri in maiuscolo;
- sostituzione delle parole accentate con il carattere apice;
- rimozione dello spazio prima e dopo l'apice se non necessario;
- sostituzione del carattere back-slash con slash;
- sostituzione o rimozione delle sottostringhe delle abbreviazioni più utilizzate lette da un archivio di riferimento, ad esempio:
    - F.NE{spazio} → FRAZIONE
    - V.LO → VICOLO
    - MONS. → MONSIGNORE

Gli indirizzi che non superano i controlli di validità vengono rigettati e passati alla gestione degli scarti.

Normalizzazione del campo "comune"

La stringa indicante il nominativo del comune viene normalizzata ed eventualmente ricercata nella tavola dei comuni.
La normalizzazione consiste nella:

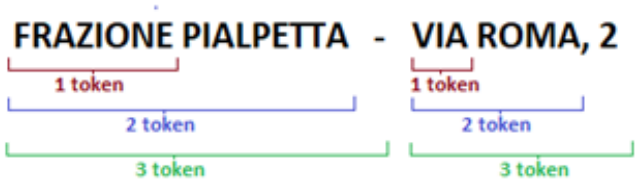
- rimozione degli spazi prima e dopo il carattere trattino (-);
- verifica della presenza dei soli caratteri [A-Z '-];

- esecuzione delle funzioni di “regular expression” per la manipolazione della stringa descritte e definite in un’apposita tavola.

Gli indirizzi normalizzati vengono quindi ricercati nella tavola dei comuni tramite confronto diretto o tramite l’operatore LIKE sostituendo gli spazi e i punti con il carattere '%'.
Anche in questa fase gli indirizzi con i comuni risultanti invalidi vengono scartati.

Estrazione del tipo via

L’estrazione del tipo via o tipi via, se presenti più toponimi nello stesso indirizzo, consiste nel ricercare le possibili voci nel testo scomposto in singole parole (token); si procede da sinistra a destra ricercando nella tavola dei tipi via, prima la voce composta dai primi tre token in sequenza (es. STRADA NUOVA COMUNALE) , poi dai primi due token (es. STRADA PROVINCIALE) ed infine col primo e unico token (es. VIALE); il procedimento viene ripetuto sul token successivo fino ad esaurimento degli stessi:



Il tipo via reperito viene assegnato all’opportuno attributo di normalizzazione e di conseguenza il relativo peso utilizzato in seguito; nel caso sia presente un alias, sempre nella stessa tavola, il tipo via originale viene sostituito con questi (es. LOCALITA’ ha come alias FRAZIONE, quindi nell’attributo di normalizzazione viene riportato FRAZIONE). Il peso associato ad ogni tipo via è un numero intero da 0 a 100 che permette di escludere quei toponimi presenti con altri nello stesso indirizzo e aventi un valore intero inferiore.
Per esempio, nell’immagine precedente sono presenti due toponimi:

- FRAZIONE PIALPETTA
- VIA ROMA

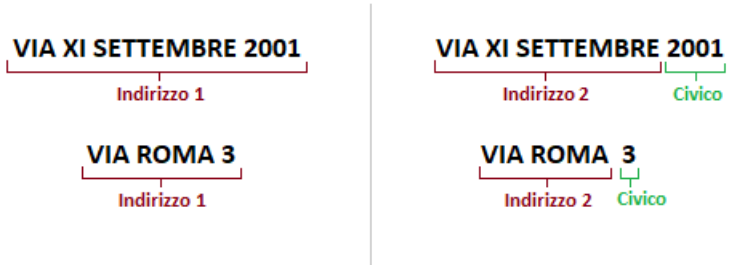
Tutti e due i toponimi vengono estratti per la successiva normalizzazione; nel caso in cui entrambi i sotto indirizzi vengano reperiti nella tavola dei toponimi, è necessario dare una priorità agli stessi: la via, per consuetudine ha un peso maggiore rispetto alla frazione.
Come evidenziato nell’esempio precedente, un singolo indirizzo esaminato può dare origine a più indirizzi simili e correlati tra loro, sia per la presenza di più toponimi, sia per ambiguità sui

tipi via o possibili errori di composizione dell'indirizzo.
Nonostante la proliferazione di indirizzi "derivati" sia un comportamento voluto che permette il reperimento del toponimo con un tasso di successo più elevato, sono state introdotte alcune limitazioni valutando la numerosità dei casi su esperienze passate. Per esempio, il carattere "-" è considerato il separatore tra indirizzi differenti, solitamente indicati per migliorare il posizionamento e per ogni stringa o sottostringa così suddivisa, il numero massimo di token analizzati è limitato a tre. Sarà compito della fase finale della normalizzazione degli indirizzi raggruppare gli stessi, scartando quelli con peso minore e tenendo solo quelli con una qualità di normalizzazione maggiore.

Estrazione civico

Questa sezione si occupa di estrarre l'eventuale numero civico presente nel testo dell'indirizzo; questo dato viene memorizzato in un attributo a parte e utilizzato nel caso in cui non sia stata esplicitata la numerazione civica nel file sorgente. La ricerca e l'estrazione dei numeri civici consiste nella verifica dei singoli token che compongono il testo normalizzato dell'indirizzo; le strategie adottate sono:

- verifica della presenza del calcolo chilometrico (token 'KM') per i tipi via che iniziano con 'STRADA'; viene estratta l'eventuale parte numerica a destra;
- verifica della similitudine di un token ad una notazione numerica; per ogni token validato vengono creati due nuovi indirizzi figli, uno in cui la parola è inclusa nell'indirizzo e un altro in cui il token, e i successivi, vengono estratti come numero civico, non potendo determinare a priori se un numero è effettivamente un civico oppure parte del toponimo:



- creazione di ulteriori indirizzi figli con il token numerico espresso in notazione romana (es. VIA 11 SETTEMBRE diventa VIA XI SETTEMBRE) oppure in cifre arabe (es. VIA XI SETTEMBRE diventa VIA 11 SETTEMBRE);

Come detto in precedenza sarà compito della fase finale della normalizzazione degli indirizzi raggruppare gli stessi, scartando quelli con peso minore ed evidenziando solo quelli con la qualità di normalizzazione maggiore.

Ricerca degli indirizzi

La ricerca degli indirizzi viene eseguita estraendo tutti i record di toponomastica per i soli comuni interessati; viene quindi generata una tabella temporanea che permette la ricerca "a tutto testo". Le ricerche vengono effettuate seguendo un ordine prestabilito successivamente descritto ed ogni passo viene eseguito in sequenza finché si ottiene un risultato.
Viene eseguito un tentativo con diverse combinazioni: preposizione e nome_via, tipo_via e nome_via, tipo_via preposizione e nome_via, ecc. e nei testi da ricercare viene sostituito il carattere punto(.), tipico delle abbreviazioni, con il carattere asterisco (*) per indicare una qualsiasi combinazione per quella singola parola (es. V. diventa V* che combina con VIA, VIALE, ecc.).
Per ogni match viene calcolato un "ranking", che rappresenta il coefficiente di differenza tra il testo da ricercare e il toponimo proveniente dalla banca dati di riferimento; i possibili toponimi vengono quindi ordinati in modo ascendente su tale coefficiente, dando la precedenza a quello più simile al testo normalizzato.

Raggruppamento e bonifica degli indirizzi normalizzati

Fase finale della normalizzazione degli indirizzi necessaria a selezionare per ogni indirizzo di partenza quello normalizzato con la qualità più alta.

Tutti gli indirizzi, validi e non, sono raggruppati in ordine di qualità decrescente.
Gli indirizzi associati a vie non presenti nella banca dati geografica vengono scartati.

Sono altresì scartati tutti i gruppi per quali non è stato possibile reperire un toponimo oppure ne siano stati trovati più uno: in quest'ultimo caso, la procedura determina il motivo dello scarto, verificando che i toponimi differiscano per la sola località oppure siano vie completamente differenti, e viene quindi valorizza l'attributo dedicato "MOTIVO_SCARTO".

Ricerca degli indirizzi suggeriti

Per gli indirizzi scartati si provvede a fornire un suggerimento, memorizzato nel campo

“INDIRIZZO_SUGGERITO” che può essere d’aiuto nella bonifica del set proposto e quindi migliorare la qualità del risultato sperato.
Si tratta di una operazione di similitudine tra testi che può avere un certo “costo“, quindi per accelerare i tempi di esecuzione è stata utilizzata l’estensione ‘spellfix’ di SQLite che permette di eseguire delle ricerche “a tutto testo” anche per nominativi con errori ortografici.
Il toponimo più simile viene quindi ricercato tramite il calcolo del coefficiente di similarità di Jaccard, utilizzando delle classi estratte dalla libreria Python ‘TextDistance’.

Indice_di_Jaccard: https://it.wikipedia.org/wiki/Indice_di_Jaccard

L'indice di Jaccard è un indice statistico utilizzato per confrontare la similarità e la diversità di insiemi campionari. Il coefficiente di Jaccard misura la similarità tra insiemi campionari, ed è definito come la dimensione dell'intersezione divisa per la dimensione dell'unione degli insiemi campionari:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

La procedura provvede ad estrapolare dal database temporaneo SQLlite, i soli toponimi del comune riferito all’indirizzo da ricercare; quindi per ogni toponimo ne calcola il coefficiente di similarità di Jaccard rispetto all’indirizzo da normalizzare, convertendo entrambi i testi in sequenze di due grafemi (o lettere) prima del calcolo. Dalla lista ottenuta si preleva l’indirizzo del dizionario con coefficiente di similarità più elevato, quindi si passa a verificare la congruenza del risultato ottenuto: i toponimi con la similitudine più alta e maggiore di un valore minimo (0.5 su un range da 0 a 1), vengono accettati come suggerimento.
Poiché per ogni indirizzo è necessario scorrere tutti i toponimi del comune in esame, l’operazione può rallentare molto l’operazione della normalizzazione; per questo motivo è stato introdotto l’utilizzo dell’estensione SQLite ‘spellfix’ che permette di accelerare i tempi. Utilizzando l’estensione ‘spellfix’ è possibile effettuare una query “a tutto testo” e calcolare un coefficiente di similarità, restringendo quindi la ricerca su un numero limitato di toponimi; immediatamente dopo viene eseguita la ricerca con Jaccard come descritto in precedenza.
Non viene utilizzato direttamente il risultato dell’interrogazione di ‘spellfix’, poiché l’indice di Jaccard è mirato sulla somiglianza delle singole parole che compongono un indirizzo e fornisce un risultato migliore.

Normalizzazione civico

La normalizzazione consiste nell'applicazione delle espressioni regolari definite puntualmente per la numerazione civica, ad esempio:

- pulizia delle parti non significative (PIANO, SCALA, ecc.);
- formattazione del civico come parte numerica / alfanumerica;
- validazione civico: numero/lettere oppure solo numero.

Gli indirizzi con civici invalidi non vengono scartati, poiché verrà sempre assicurata la georeferenziazione, anche se approssimata.

Georiferimento

Sezione dedicata alla classificazione dei numeri civici in base alla loro valorizzazione e presenza nella tavola della numerazione civica. Vengono importati in memoria tutti i civici dei soli toponimi reperiti dagli indirizzi normalizzati; quindi i civici normalizzati e valorizzati vengono ricercati tra quelli disponibili dalla estrazione della tavola di riferimento e indirizzati su tre differenti logiche di georeferenziazione.

Posizione civici reperiti dalla tavola della numerazione civica

Gli indirizzi il cui civico è stato ritrovato esattamente nella estrazione dalla tavola della numerazione civica, acquisiscono direttamente la posizione del corrispondente elemento puntuale.

Posizione civici calcolati per prossimità

I civici correttamente valorizzati, ma che non sono presenti nella tavola di riferimento, vengono ricercati per prossimità: tra tutti i civici relativi allo stesso indirizzo, si sceglie quello la cui parte numerica è più vicina a quello in esame. Si assegna quindi, all'indirizzo in esame, la posizione dell'elemento puntuale ritrovato.

Posizione civici assenti per il toponimo indirizzo

Gli indirizzi per i quali non è possibile in alcun modo reperire la posizione dalla tavola di riferimento della numerazione civica, vengono georiferiti utilizzando i grafi stradali.
Di tutti gli elementi lineari, corrispondenti all'indirizzo interessato, viene prelevato l'elemento più lungo e calcolato il punto medio. Questa posizione approssimativa viene assegnata al civico di interesse.