UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

TESI DI LAUREA

Odonomastica italiana: creazione di un dataset e analisi antroponimica

Relatori:

Enrica Salvatori

Vittore Casarosa

Candidato:

Sofia Zuffi

ANNO ACCADEMICO 2020/2021

Indice

Introduzione

Il presente studio prende le mosse dallo studio di Camilla Zucchi (Zucchi – 2020[1]) nel quale, partendo dall'analisi di un campione di qualche centinaio di comuni d'Italia, divisi per fasce di popolazione, si indagava sulla percentuale di vie intitolate a figure femminili.

Il duplice obiettivo di questa tesi è stato quello di:

- eseguire nuovamente questo tipo di analisi, concentrandosi sull'aspetto antroponimico (ossia dello studio dei nomi di persona), utilizzando come dato di partenza non un campione ma l'intero dataset delle vie/piazze italiane (7904 comuni);
- creare un sito web che rendesse fruibili sia le analisi effettuate sui dati, sia i dati stessi, accompagnati da ogni passaggio della loro elaborazione.

Nel primo capitolo sono introdotti i concetti di base utili alla comprensione della categoria dei dati presi in analisi. Nel secondo viene illustrato il procedimento di recupero ed estrazione dei dati dal web. Nel terzo capitolo sono affrontate le fasi e le strategie di elaborazione computazionale dei dati attraverso le quali da materiale "grezzo" si è giunti ad una catalogazione e organizzazione che permettesse l'esecuzione dell'analisi. Nel quarto capitolo vengono mostrati i risultati di questa analisi.

Nel quinto capitolo lo sguardo si sposta sull'ideazione del sito web per la fruizione del dataset e sull'organizzazione delle sue due sezioni, la prima di visualizzazione e scaricamento dei dati, la seconda di consultazione (ed eventuale scaricamento) dei risultati dell'analisi.

[1] Zucchi, Camilla. 2020. *La toponomastica femminile in Italia tra retaggi del passato e sfide del presente*

1. Definizioni e terminologia

In questo capitolo vengono introdotte alcune definizioni utili alla comprensione della materia trattata, oltre a sigle in gergo tecnico utilizzate nella fase di elaborazione dei dati.

1.1 Urbanistica

L'urbanistica è la disciplina che si occupa dello studio del territorio e ha come scopo la progettazione dello spazio urbano e la sovraintendenza alle modificazioni del territorio incluso nella città o collegato con essa (Treccani, voce *Urbanistica*).

1.1.1 Odonomastica

L'odonomastica (dal greco *hodós* – "strada", e *onomastikòs* – "atto a denominare") è il ramo dell'urbanistica che ha per oggetto lo studio dei nomi delle strade, sia con riferimento concreto a una determinata zona o località, sia con riguardo alla scelta o al modo della loro formazione (Treccani, voce *Odonomastica*).

1.1.2 Odonimo: DUG e DUF

Odonimo, o "toponimo stradale", è il termine che identifica il nome proprio assegnato a una via. Gli odonimi sono generalmente costituiti da due parti:

- una denominazione generica che specifica la tipologia dello spazio che si vuole identificare;
- un nome proprio dello spazio che si vuole identificare.

La prima parte viene chiamata "appellativo", "qualificatore di toponimo" o DUG, "denominazione urbanistica generica", e la seconda DUF, "denominazione urbanistica ufficiale". Gli esempi più comuni di DUG sono "via" e "piazza", ma numerosi DUG rinviano anche a condizioni regionali o locali, talvolta in veste vicina al dialetto (Marcato, 2005). I DUG possono essere importanti testimoni della storia urbanistica e linguistica di una singola città o di una zona, basti pensare ai veneziani "campo" e "campiello" per indicare piazzette e "calle" per le vie (Treccani – Enciclopedia dell'italiano, voce *Odonimi*).

La possibilità di divisione di un odonimo in DUG e DUF ha un'importanza fondamentale: il separare i nomi delle vie (DUF) dai loro numerosi appellativi (87 DUG secondo la tabella ufficiale dell'Istituto Nazionale di Statistica, diverse

centinaia secondo tabelle non ufficiali) permetterà, durante la fase di elaborazione dei dati, una migliore categorizzazione delle vie stesse (v. 3.2).

1.2 Antroponimia

L'*antroponimia* è uno dei rami dell'onomastica, che ha per oggetto lo studio dei nomi di persona (Treccani, voce *Antroponimia*). Lo studio che sarà effettuato sul dataset elaborato (v. 4) si concentrerà sull'analisi *antroponimica* delle vie, ossia quella relativa a nomi propri di persona.

2. Raccolta dei dati

All'interno di questo capitolo viene illustrata la fase di recupero dei dati necessari alla creazione del dataset di tutte le vie italiane.

2.1 Ricerca di un archivio

Il punto di partenza per la creazione del dataset è stato quello della ricerca di possibili fonti web che contenessero un elenco quanto più aggiornato possibile di tutte le vie italiane, che fossero contemporaneamente di libera consultazione e utilizzo.

Una prima ricerca aveva individuato nell'ANNCSU, "Archivio nazionale dei numeri civici delle strade urbane" (ex. ANSC – "Archivio nazionale degli stradari e dei numeri civici"[1]) un ottimo candidato di archivio dal quale estrarre i dati. Il nuovo archivio ANNCSU avrebbe dovuto basarsi sul precedente ANSC, con un miglioramento portato dagli aggiornamenti annuali (e non più decennali) operati direttamente dalle amministrazioni comunali. Sfortunatamente, non è stato possibile recuperare tali dati perché il collegamento con il sito si è rivelato non affidabile.

Si è presa allora in considerazione la possibilità di utilizzare i dati dell'ultimo censimento della popolazione e delle abitazioni effettuato dall'ISTAT. Questi però, ad una prima analisi sono risultati troppo poco aggiornati, in quanto il censimento è stato attuato nel 2011, e incompleti di tutti i nomi delle aree di circolazione di un centro abitato che risultavano sprovviste di abitazioni (il censimento limitava, infatti, la propria fascia d'interesse alle sole vie, piazze, ecc. fornite di abitazioni e quindi di numeri civici).

La ricerca si è quindi orientata su OpenStreetMap (OSM), un progetto collaborativo finalizzato a creare mappe del mondo a contenuto libero. Il progetto punta ad una raccolta mondiale di dati geografici, con scopo principale la creazione di mappe e

[1] L'archivio nazionale degli stradari e dei numeri civici è un archivio informatizzato e codificato, contenente gli stradari (elenco delle denominazioni delle aree di circolazione) e i numeri civici di tutti i Comuni italiani. Il primo impianto di tale archivio è stato effettuato utilizzando l'infrastruttura tecnologica e i dati già predisposti dall'Agenzia delle Entrate per la costituzione dell'archivio nazionale toponomastica (docs.italia.it, *Anagrafe nazionale numeri civici e strade urbane*, n.d.).

cartografie (Anisa Kuci, *Presentazione ufficiale di OpenStreetMap*, 2021). Il database di OpenStreetMap contiene quindi dati inseriti dagli utenti della piattaforma e aggiornati quasi in tempo reale.

2.2 Download delle vie

Per il download dei dati dal database di OSM è stata utilizzata l'API[2] Overpass: un'API di sola lettura che ritorna i dati della mappa OSM selezionati nella richiesta. L'interfaccia funziona come un database sul Web: il client invia una query all'API e questa restituisce il set di dati corrispondente alla query (OpenStreetMap Wiki, voce *Overpass API*). Per permettere di interrogare i dati di OpenStreetMap seguendo criteri di ricerca specifici è stato sviluppato un particolare linguaggio di ricerca, chiamato Overpass QL (abbreviazione di "Overpass Query Language"), un linguaggio di programmazione procedurale e imperativo scritto con una sintassi in stile C (OpenStreetMap Wiki, voce *Overpass API/Overpass QL*). Attraverso questo linguaggio è stato possibile chiedere al database di estrarre tutte le vie all'interno di una specifica area, delimitandone i confini tramite coordinate geografiche o tramite nome.

Una sola estrazione dell'intero elenco delle vie tramite nome della nazione (Italia) avrebbe comportato la perdita del riferimento del comune di appartenenza di ogni set di vie. È stato perciò scelto di estrarre i nomi delle vie tramite il nome del comune di appartenenza, eseguendo 7.904 chiamate consecutive al server di OpenStreetMap, una per ogni comune. Le chiamate sono state eseguite automaticamente tramite la scrittura di un programma in Python che ad ogni chiamata si è occupato di aggiornare il nome del comune del quale richiedere i dati. Alla fine di ogni chiamata, i dati ricevuti in formato JSON[3] sono stati salvati in file distinti per nome del

[2] Le API (acronimo di "Application Programming Interface", "Interfaccia di programmazione delle applicazioni") sono set di definizioni e protocolli con i quali vengono realizzati e integrati software applicativi; consentono ai prodotti o ai servizi di comunicare con altri prodotti o servizi senza sapere come vengano implementati (redhat.com, *I vantaggi delle interfacce di programmazione delle applicazioni*, 2017).

[3] JSON (acronimo di "JavaScript Object Notation") è un semplice formato per lo scambio di dati. Per le persone è facile da leggere e scrivere, mentre per le macchine risulta facile da generare e da analizzare. JSON è basato su due strutture: un insieme di coppie nome/valore e un elenco ordinato di valori (json.org, n.d.).

comune.

L'elenco dei nomi dei comuni, aggiornato al 2015, è stato estratto dal file 'Elenco dei comuni italiani.csv', scaricato dal sito dell'ISTAT alla pagina "Codici statistici delle unità amministrative territoriali: comuni, città metropolitane, province e regioni". Dallo stesso file, per ogni comune, sono state estratte anche le seguenti informazioni:

- identificativo ISTAT del comune;
- identificativo ISTAT della provincia;
- nome della provincia;
- identificativo ISTAT della regione;
- nome della regione;
- nome del territorio (diviso in Nord-Est, Nord-Ovest, Centro, Sud, Isole).

In ogni file è stato in questo modo possibile mantenere non solo il riferimento al comune di appartenenza, ma anche quello alla provincia, regione e al territorio.

Ad ogni comune è stato infine aggiunto il valore corrispondente al proprio numero di abitanti, estratto dal sito dell'ISTAT (valore aggiornato al 1° gennaio 2020)[4].

[4] Consultabile al link: http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1

3. Fase di elaborazione

In questo capitolo vengono mostrate le diverse fasi di pulizia ed elaborazione dei dati. In particolare è mostrata la procedura con la quale dai dati grezzi ottenuti da OpenStreetMap, mediante tecniche di machine learning[1], sono state estratte le caratteristiche necessarie allo svolgimento dell'analisi. I risultati dell'analisi verranno poi visualizzati e commentati.

3.1 Pulizia dei dati

Completata la fase di raccolta dei dati, l'elenco di vie presenti nei files di ogni comune risulta costituito da una sequenza di sezioni come quella mostrata (in JSON) di seguito:

```
{
  "type": "way",
  "id": 42023679,
  "tags": {
  "highway": "tertiary",
  "lanes": "2",
  "lit": "yes",
  "name": "Strada per Bairo",
  "ref": "SP41",
  "surface": "asphalt"
  }
}
```

Insieme al nome della via, visibile come valore del campo "name", possono quindi essere presenti informazioni accessorie come la categoria della via (in questo esempio "tertiary"), il numero di corsie ("lanes": "2"), l'illuminazione ("lit"), il riferimento ("ref") e la superficie della via ("surface").

È stato quindi necessario operare un passaggio di pulizia dei dati, in modo da separare le informazioni accessorie utili da quelle inutili.

In particolare, il campo "ref" viene utilizzato per numeri o codici di riferimento ed è

[1] Il Machine Learning (ML) è una branca dell'intelligenza artificiale (AI) che si occupa di creare sistemi che apprendono (o migliorano le performance) in base ai dati che utilizzano (Oracle, *Cos'è il machine learning?)*.

comune per strade, uscite autostradali, percorsi, ecc. (OpenStreetMap Wiki, voce *Key:ref*). È stato quindi l'unico campo accessorio a non essere eliminato nella fase di pulizia, in quanto portatore di informazioni utili all'identificazione della via tramite ricerca per nome.

Mediante un programma scritto in Python, da ogni comune è stato creato un ulteriore file in formato JSON, contenente l'elenco delle vie e il riferimento di ogni via (se presente nei dati iniziali). Sono poi state aggiunte le informazioni relative al DUG e DUF di ciascuna via, separando la denominazione urbanistica generica, se presente, dalla rimanente parte del nome della via. Il registro ufficiale delle DUG è stato estratto dal sito dell'ISTAT, e consta di 87 termini[2].

Ricordiamo ora nuovamente, prima di esaminare i seguenti passaggi, lo scopo dell'elaborazione: associare ad ogni via la maggior quantità di informazioni possibili, in modo da creare un dataset che possa essere utilizzato per portare avanti ulteriori studi sull'odonomastica italiana. Per eseguire l'analisi antroponimica, ad ogni via dovrà essere associata una categoria (come "nome di persona", "nome di luogo", "nome di evento", ecc.) e ogni via associata alla categoria "nome di persona" dovrà distinguersi in "nome maschile" o "nome femminile".

3.2 Associazione della categoria

Per l'associazione della categoria si è deciso di utilizzare tecnologie di Natural Language Processing[3], in particolare sono stati utilizzati due processori NER (Named Entity Recognition). Un processore NER si occupa d'identificare e classificare (quasi sempre correttamente) le "Named Entities" presenti in un testo in categorie predefinite come persone, luoghi, oggetti, numeri, espressioni temporali ecc.

[2] Dati consultabili al link http://registry.geodati.gov.it/dug (visitato il 15 ottobre 2021).

[3] Con i termini "elaborazione del linguaggio naturale", o NLP, ci si riferisce alla branca dell'informatica e, più specificamente, alla branca dell'intelligenza artificiale interessata a fornire ai computer la capacità di comprendere testo e parole pronunciate più o meno allo stesso modo degli esseri umani.
L'NLP combina la linguistica computazionale - modellazione del linguaggio umano basata su regole - con modelli statistici, di machine learning e deep learning. Insieme, queste tecnologie consentono ai computer di elaborare il linguaggio umano sotto forma di testo o dati vocali e di "capire" il suo pieno significato, completo dell'intento e del "sentimento" di chi parla o scrive (ibm.com, *Natural language Processing (NLP)*).

(Provino, *Cos'è NER: Named Entity Recognition*). In questo caso sono stati utilizzati i NER delle librerie Python "SpaCy" e "Stanza". Per SpaCy è stata utilizzata la pipeline[4] "it_core_news_lg", per Stanza il pacchetto "FBK". Entrambi i processori categorizzano le named entities in *persona* ("PER), *luogo* ("LOC") o *organizzazione* ("ORG"). Il processore di SpaCy aggiunge anche la categoria miscellanea ("MISC"), che identifica una named entity che non rientra in nessuna delle tre categorie precedenti.

Tramite un programma scritto in Python ogni DUF è stata analizzata da entrambi i processori NER, e ad ogni via sono state aggiunte, se riconosciute dai processori, le seguenti informazioni:

- valore della named entity presente nella DUF;
- categoria dell'entità.

È stato così possibile estrarre 647.718 vie intitolate a persone dal totale di 1.153.513 vie italiane.

3.3 Associazione del genere

Per distinguere le vie intitolate a entità maschili da quelle intitolate a entità femminili sarebbe risultato molto utile l'utilizzo di uno strumento simile a quello utilizzato per l'associazione della categoria. In seguito a ricerche però non è stato possibile individuare nessun software atto a tale scopo. È stato portato avanti un tentativo di elaborazione tramite l'API Genderizo.io, che prende in analisi un nome, lo confronta con un archivio di 13.461.719 nomi (per la lingua italiana) raccolti dal web e restituisce un risultato di questo tipo


```
{
    "name": "Marco",
    "gender": "male",
    "probability": 0.99,
    "count": 165452
}
```


[4] Il termine pipeline in informatica e in elettronica si riferisce a un manufatto composto da più elementi. Ogni elemento provvede a ricevere in ingresso un dato o un segnale, ad elaborarlo e poi a trasmetterlo all'elemento successivo (Wikipedia, voce *Pipeline*).

ma con scarsi risultati. L'API non si comporta bene con dati "non ideali", come quelli presenti nelle DUF delle vie, e non restituisce alcun risultato neanche nei casi in cui come dato in entrata venga presentata l'unione di un nome e di un cognome (es. "Mario Rossi").

È stato perciò deciso di creare uno strumento ad hoc che restituisse un risultato simile a quello dell'API menzionata sopra.

Per l'analisi delle named entities associate alla categoria "persona" (PER) è stato utilizzato il dataset "genderNamesITA"[5] contenente oltre 32.000 nomi propri di persona italiani, creato a partire dalla raccolta dei dati delle anagrafi delle amministrazioni comunali italiane dal 1985 al 2014. Ad ogni nome del dataset sono associati tre valori numerici: il primo rappresenta il numero delle occorrenze totali all'interno dei documenti, il secondo il numero di occorrenze in cui il nome è comparso come nome di persona maschile e il terzo il numero di occorrenze in cui il nome è comparso come nome di persona femminile.

Grazie a questi dati è stato possibile aggiungere ad ogni via, tramite la scrittura di un programma in Python, due valori numerici che rappresentano la probabilità (espressa come numero tra 0.0 e 100.0) che l'entità analizzata contenga al suo interno un nome di persona maschile o femminile. Nei casi in cui l'analisi non è stata in grado di riconoscere alcun nome all'interno dell'entità, entrambi i valori numerici sono stati settati a "0.0".

A questo punto dell'elaborazione le informazioni associate ad ogni via sono quindi riportate di seguito:


```
"Via Alessandro Manzoni": {
      "dug": "via",
       "duf": "Alessandro Manzoni",
       "ner_spacy_ent": "Alessandro Manzoni",
       "ner_spacy_tipo": "PER",
       "ner_stanza_ent": "Alessandro Manzoni",
       "ner_stanza_tipo": "PER",
       "maschio_%": 100.0,
        "femmina_%": 0.0
}
```


[5] Consultabile al seguente link: https://github.com/mrblasco/genderNamesITA.

3.4 Associazione dell'entità

L'ultimo passo dell'elaborazione che è stato portato avanti è stato quello dell'associazione di un'entità fisica a quante più vie possibili. Per fare ciò, è stato prima necessario raggruppare e associare ad un'entità unica tutte quelle named entities corrispondenti formalmente alla stessa entità generica (es. "Giuseppe Garibaldi"), ma differenti per forma (es. "G. Garibaldi", "Garibaldi G.", "Garibaldi").

3.4.1 Raggruppamento delle entità simili

Per il raggruppamento delle entità simili, operato sulle sole named entities associate alla categoria "persona" (PER), è stata usata la libreria Python "textpack"[6], che fornisce una misura di similarità tra due stringhe di caratteri. A questo scopo, costruisce una matrice che ha per righe le named entities e per colonne le stesse named entities divise in n-grammi[7], e ad ogni n-gramma assegna un punteggio TF-IDF[8]. Quindi utilizza la moltiplicazione di matrici per calcolare la somiglianza del coseno (una metrica compresa tra 0 e 1 utilizzata per determinare quanto siano simili tra loro delle stringhe) tra questi valori (Whyte, Luke. 2019. *Group thousands of similar spreadsheet text cells in seconds*).

Modificando, nella funzione utilizzata, i parametri "match_threshold" e "ngram_length", che indicano rispettivamente il valore, compreso tra 0 e 1, della soglia di similarità del coseno usata per determinare se due stringhe devono essere raggruppate e il valore nella lunghezza degli n-grammi, si è raggiunto un raggruppamento ottimale, seppur non perfetto, delle entità. Come è possibile vedere nell'esempio seguente, le named entities che rappresentano l'entità "Giuseppe Garibaldi" sono state raggruppate correttamente. È interessante notare come anche numerosi casi di refusi siano stati raggruppati nel modo corretto ("Giuseppe

[6] Consultabile al seguente link: https://github.com/lukewhyte/textpack.

[7] Un n-gramma è una sottosequenza di n elementi di una data sequenza. In base all'applicazione, gli elementi in questione possono essere fonemi, sillabe, lettere, parole, ecc. (Wikipedia, voce *N-gramma*).

[8] La funzione di peso tf-idf ("term frequency – inverse document frequency") è una funzione utilizzata in information retrieval per misurare l'importanza di un termine rispetto ad un documento o ad una collezione di documenti (Wikipedia, voce *Tf-idf*).

Garibali", "Giseppe Garibaldi", ecc.). Nello stesso gruppo sono state però inserite anche due vie intitolate ad Anita Garibaldi ("A. Garibaldi"):

Named entities	Frequenza	Gruppo
Giuseppe Garibaldi	2812	Giuseppe Garibaldi
Garibaldi	900	Giuseppe Garibaldi
G. Garibaldi	49	Giuseppe Garibaldi
giuseppe Garibaldi	2	Giuseppe Garibaldi
Giuseppe Garibald	2	Giuseppe Garibaldi
Giusepe Garibaldi	2	Giuseppe Garibaldi
A. Garibaldi	2	Giuseppe Garibaldi
Guseppe Garibaldi	1	Giuseppe Garibaldi
Giuseppe Garibali	1	Giuseppe Garibaldi
Giseppe Garibaldi	1	Giuseppe Garibaldi

Le due vie sopra menzionate rappresentano lo 0,05% del totale del gruppo, e non inficiano il risultato dell'analisi che verrà mostrata in seguito.

3.4.2 Collegamento a Wikidata

Da 647.718 named entities di tipo PER sono stati formati 171.003 gruppi unici, ognuno dei quali corrispondente ad un'entità che dovrebbe coincidere un nome di persona. Per identificare univocamente le entità, che si riferiscono a personaggi storici, scienziati, umanisti ecc. si è cercato di collegare ogni nome ad una pagina di Wikidata.

Wikidata è un database libero, collaborativo e multilingue che raccoglie dati strutturati per fornire supporto a Wikipedia, a Wikimedia Commons, ad altri progetti

del movimento Wikimedia e a chiunque nel mondo. L'archivio di Wikidata è costituito principalmente da elementi, ciascuno caratterizzato da una etichetta, una descrizione e diversi alias. Gli elementi sono identificati in modo univoco dalla lettera "Q" seguita da un numero (es. "Q2" corrisponde all'elemento "Terra", terzo pianeta del sistema solare) (Wikidata, voce *Introduction*).

Per eseguire il collegamento è stata usata la funzione di "riconciliazione" di OpenRefine[9], che permette di eseguire un'associazione automatica tra i dati di un determinato dataset e le entità del database di Wikidata. Delle 171.003 entità, 52.869 sono state riconciliate dall'algoritmo. I 52.869 link ottenuti sono stati poi collegati alle named entities contenute nei gruppi: delle 647.718 named entities di tipo PER, 347.266 hanno ottenuto un link a Wikidata.

L'osservazione del file di un comune riporterebbe ora un elenco di vie associate a nomi di persona molto simile a questo esempio:


```
"Via Guglielmo Marconi": {
      "dug": "via",
      "duf": "Guglielmo Marconi",
      "ner_spacy_ent": "Guglielmo Marconi",
      "ner_spacy_tipo": "PER",
      "ner_stanza_ent": "Guglielmo Marconi",
      "ner_stanza_tipo": "PER",
      "maschio_%": 100.0,
      "femmina_%": 0.0,
      "entità": "Guglielmo Marconi",
      "wiki_ent": "Guglielmo Marconi",
      "wiki_id": "Q36488",
      "wiki_link": "https://www.wikidata.org/wiki/Q36488"
      }
```


La possibilità di ripetere l'elaborazione, tentando un raggruppamento e una successiva riconciliazione anche per named entities delle categorie "luogo" e "organizzazione" è stata presa in considerazione e successivamente scartata perché i

[9] OpenRefine (precedentemente Google Refine) è un potente strumento per lavorare con dati disordinati. Permette la pulizia dei dati, la loro trasformazione da un formato all'altro e l'estensione con servizi web e dati esterni (openrefine.org, n.d.).

valori di “ner_spacy_ent” e “ner_stanza_ent” per categorie diverse da “persona” sono risultati troppo “sporchi”, sia per produrre dei gruppi adeguati, sia per essere riconciliati con il database di Wikidata.

4. Analisi quantitativa[1]

Sul dataset così creato, prendendo in considerazione solo le vie intitolate a nomi propri di persona, sono state eseguite due analisi differenti. La prima si è concentrata sul confronto tra vie intitolate a personaggi maschili e vie intitolate a personaggi femminili. La seconda si è concentrata sull'estrazione dei nomi di persona in ordine di frequenza discendente, dal più frequente al meno frequente.

4.1 Analisi per genere

L'analisi è stata eseguita mediante la libreria Pyhton Pandas[2]. Si è proceduto su due binari diversi: le percentuali di vie intitolate a persone maschili e femminili sono state analizzate prima seguendo una divisione geografica del territorio italiano, e poi per divisione in fasce di popolazione. Le divisioni geografiche sono state quelle per:

- territorio: Nord-est, Nord-ovest, Centro, Sud, Isole;
- regione: Valle d'Aosta, Piemonte, Liguria, Lombardia, Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Lazio, Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia e Sardegna;
- provincia: delle 107 province italiane si è scelto di rappresentare tramite grafico a barre le prime 10 province per percentuale di vie intitolate a figure femminili e le ultime 10 (per un totale di 20 province);

Per le divisioni in fasce di popolazione è stato deciso di utilizzare le fasce dell'Istituto Nazionale di Statistica[3]:

[1] Per analisi quantitativa si intende l'esame di un insieme di dati di cui si cerca la ripartizione statistica di fenomeni e categorie di informazioni (html.it, *Analisi qualitativa e quantitativa: cos'è e come si fa*. 2007).

[2] Nella programmazione per computer, Pandas è una libreria software scritta per il linguaggio di programmazione Python per la manipolazione e l'analisi dei dati. In particolare, offre strutture dati e operazioni per manipolare tabelle numeriche e serie temporali (pandas.pydata.org, n.d.).

[3] Consultabile al link: https://finanzalocale.interno.gov.it/docum/studi/varie/200707varclass.html

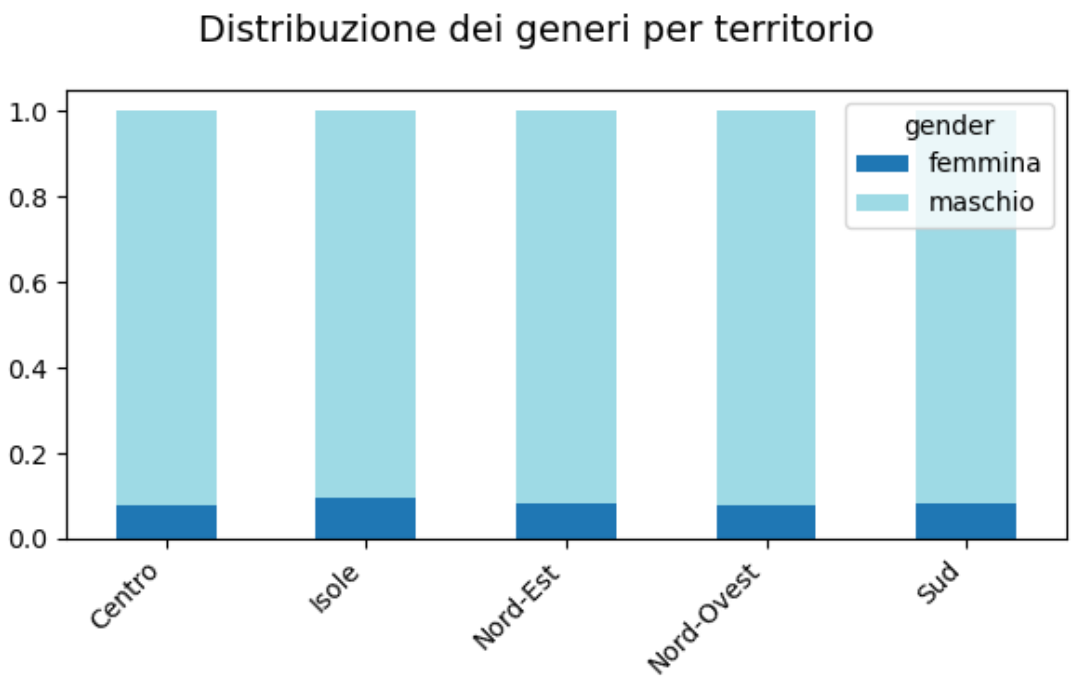
Fascia	Numero di abitanti
I	Meno di 500
II	500 – 999
III	1.000 – 1.999
IV	2.000 – 2.999
V	3.000 – 4.999
VI	5.000 – 9.999

Fascia	Numero di abitanti
VII	10.000 – 19.999
VIII	20.000 – 59.999
IX	60.000 – 99.999
X	100.000 – 249.999
XI	250.000 – 499.999
XII	Più di 500.000

4.1.1 Analisi per suddivisione geografica

I risultati dell'analisi mostrano che, per quanto riguarda la prima divisione, tutti i territori tranne "Isole" si assestano su percentuali simili di vie intitolate a donne e uomini, con la percentuale delle vie intitolate a donne compresa tra l'8% e l'8,2%. Sicilia e Sardegna insieme invece raggiungono una percentuale più alta, quasi del 9,8%.

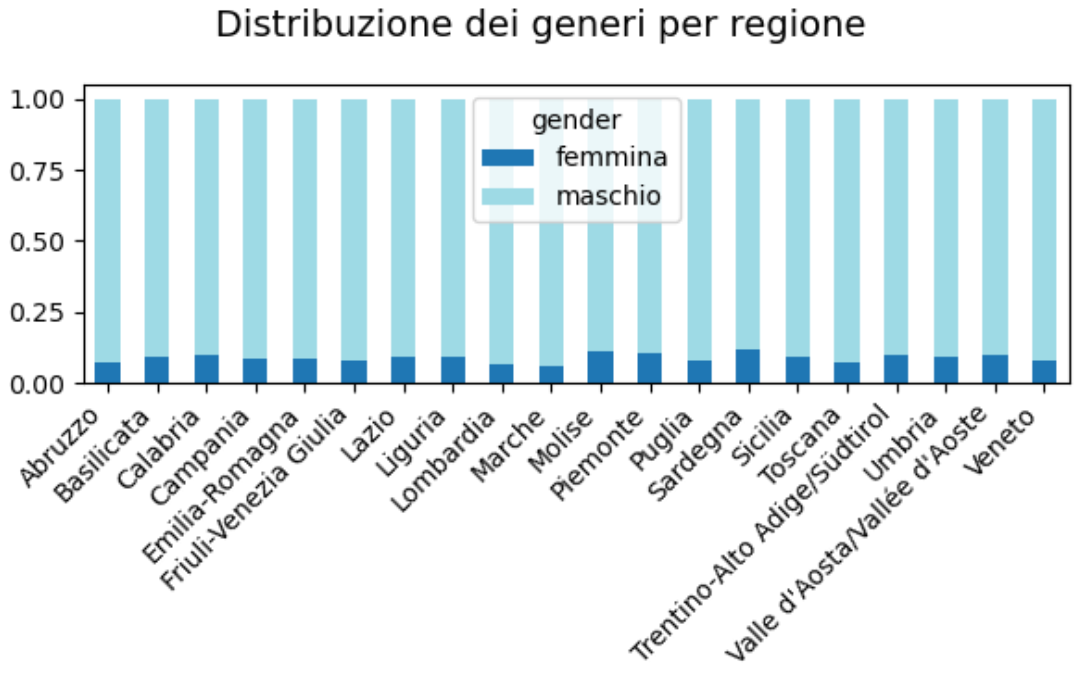
Di seguito viene mostrato il grafico a barre dei territori:



Per una consultazione più precisa dei dati, è possibile fare riferimento alla seguente tabella:

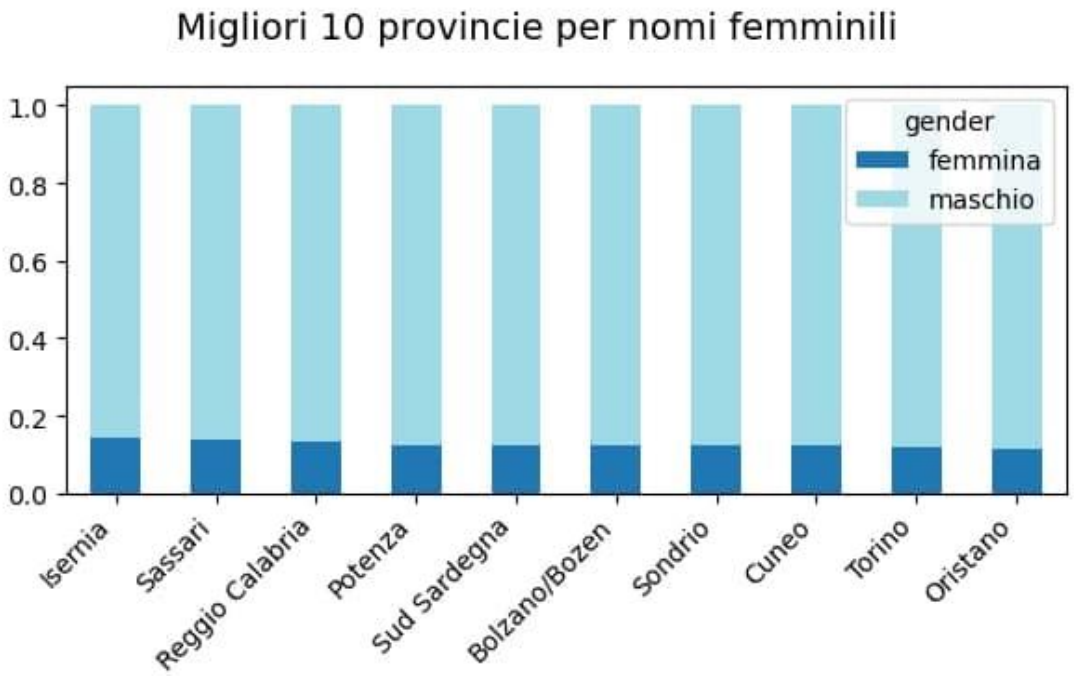
	Maschile	Femminile	Maschile %	Femminile %
Nord-Ovest	88.378	7741	91,95%	8,05%
Nord-Est	62.282	5655	91,68%	8,32%
Centro	50.591	4431	91,95%	8,05%
Sud	67.225	6067	91,72%	8,28%
Isole	43.349	4707	90,21%	9,79%

Per quanto riguarda la suddivisione in regioni, i valori di ogni regione si presentano maggiormente variabili: le regioni con il maggior numero di vie intitolate a figure femminili sono la Sardegna, il Molise e il Piemonte; quelle con il minor numero di vie intitolate a donne sono le Marche e la Lombardia.
Di seguito vengono mostrati grafico a barre e tabella:

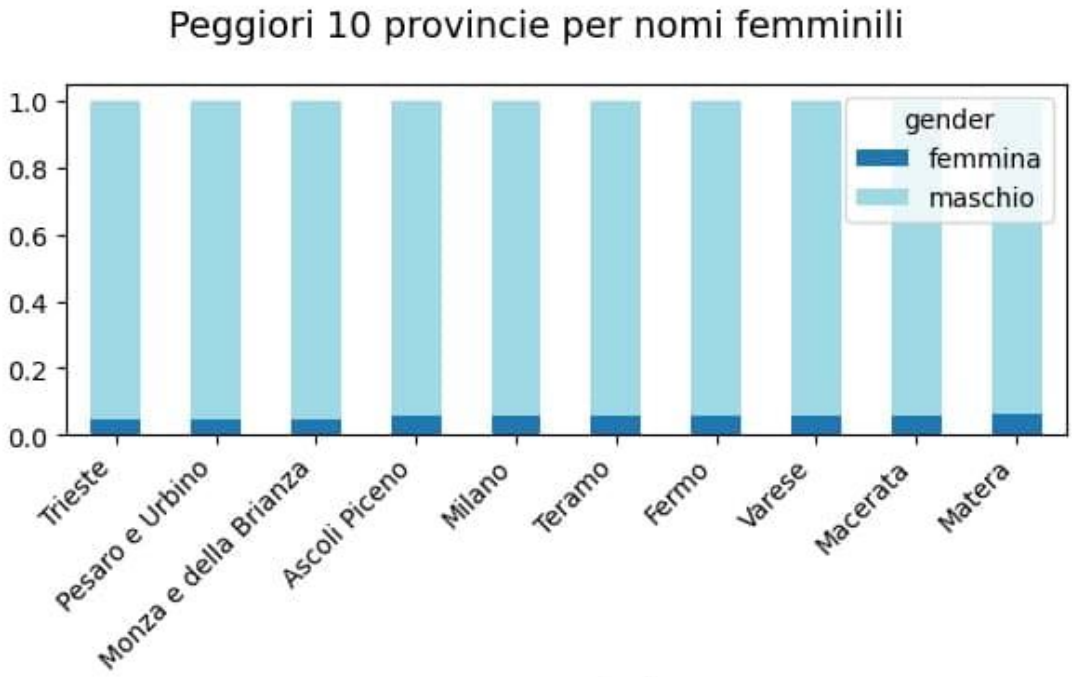


	Maschile	Femminile	Maschile %	Femminile %
Piemonte	24.015	2838	89,43%	10,57%
Valle d'Aosta	1556	169	90,20%	9,80%
Lombardia	56.615	4098	93,25%	6,75%
Liguria	6192	636	90,69%	9,31%
Trentino-Alto Adige	2088	226	90,23%	9,77%
Veneto	24.904	2220	91,82%	8,18%
Friuli-Venezia Giulia	6586	588	91,80%	8,20%
Emilia-Romagna	28.704	2621	91,63%	8,37%
Toscana	18.089	1467	92,50%	7,50%
Umbria	5622	554	91,03%	9,97%
Marche	8532	554	93,90%	6,10%
Lazio	18.348	1856	90,81%	9,19%
Abruzzo	7652	621	92,49%	7,51%
Molise	2244	275	89,08%	10,92%
Campania	16.525	1481	91,78%	8,22%
Puglia	29.015	2428	92,28%	7,72%
Basilicata	3286	320	91,13%	8,87%
Calabria	8503	942	90,03%	9,97%
Sicilia	35.426	3629	90,71%	9,29%
Sardegna	7923	1078	88,02%	11,98%

Infine, per quanto riguarda la suddivisione per province, delle 10 province con percentuale più alta di vie intitolate a figure femminili le prime tre sono risultate essere quelle di Isernia, Sassari e Reggio Calabria. All'altro capo della classifica, tra le 10 province a minor percentuale di figure femminili, le peggiori sono quelle di Trieste, Pesaro e Urbino, Monza e Brianza. Di seguito vengono mostrati grafici e tabelle per entrambe le classifiche:



	Maschile	Femminile	Maschile %	Femminile %
Isernia	543	90	85,78%	14,22%%
Sassari	2594	421	86,04%	13,96%
Reggio Calabria	2521	387	86,69%	13,31%
Potenza	1283	185	87,40%	12,60%
Sud Sardegna	1106	157	87,57%	12,43%
Bolzano	241	34	87,64%	12,36%
Sondrio	681	96	87,64%	12,36%
Cuneo	3826	532	87,79%	12,21%
Torino	8363	1116	88,23%	11,77%
Oristano	970	128	88,34%	11,66%

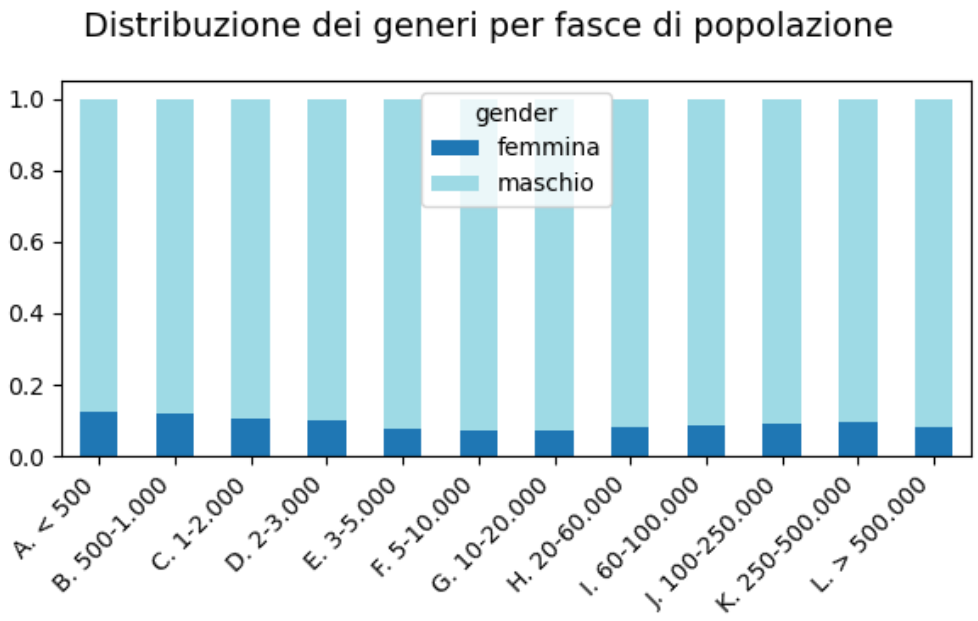


	Maschile	Femminile	Maschile %	Femminile %
Trieste	700	32	95,63%	4,37%
Pesaro e Urbino	2234	115	95,10%	4,90%
Monza e della Brianza	4510	233	95,09%	4,91%
Ascoli Piceno	1102	65	94,43%	5,57%
Milano	12.736	753	94,42%	5,58%
Teramo	1567	93	94,40%	5,60%
Fermo	623	37	94,39%	5,61%
Varese	6781	425	94,10%	5,90%
Macerata	2134	134	94,09%	5,91%
Matera	2003	135	93,69%	6,31%

4.1.2 Analisi per fascia di popolazione

L'analisi delle fasce di popolazione, per cui sono state estratte le percentuali degli stessi valori dell'analisi precedente, sembra riscontrare prima un trend discendente

che parte dalla fascia di comuni con meno abitanti e arriva fino alla fascia 5-10.000 persone, poi un trend ascendente fino ai comuni con 250-500.000 abitanti, e poi una seconda discesa. Di seguito il grafico a barre e la tabella riassuntiva dei valori:



	Maschile	Femminile	Maschile %	Femminile %
< 500	3953	557	87,65%	12,35%
500 – 1.000	6075	833	87,94%	12,06%
1.000 – 2.000	11.721	1408	89,28%	10,72%
2.000 – 3.000	12.120	1380	89,78%	10,22%
3.000 – 5.000	19.606	1608	92,42%	7,58%
5.000 – 10.000	37.920	2900	92,90%	7,10%
10.000 – 20.000	42.456	3283	92,82%	7,18%
20.000 – 60.000	63.049	5466	92,02%	7,98%
60.000 – 100.000	38.977	3732	91,26%	8,74%
100.000 – 250.000	41.339	4109	90,96%	9,04%
250.000 – 500.000	13.066	1358	90,59%	9,41%
> 500.000	21.441	1961	91,62%	8,38%

4.2 Analisi di frequenza

Il primo obiettivo è stato quello di estrarre dal dataset tutte le named entities presenti e di ordinarle dalla più frequente alla meno frequente. I primi 10 nomi sono risultati essere quelli dei seguenti personaggi:

- Giuseppe Garibaldi (3774)
- Guglielmo Marconi (3217)
- Giuseppe Mazzini (2366)
- Dante Alighieri (2315)
- Giacomo Matteotti (2255)
- Aldo Moro (2229)
- Giuseppe Verdi (2168)
- Vittorio Emanuele (2120)
- Antonio Gramsci (1899)
- Alcide De Gasperi (1887)

Tra parentesi è stato riportato il valore della frequenza.

Vale la pena notare che in quinta posizione il risultato dell'analisi riportava il nome di luogo "Vittorio Veneto" (2305), estratto erroneamente perché inserito dai processori NER nella categoria "persona", e processato come nome proprio di persona maschile nella fase di elaborazione del genere in quanto contenente al suo interno il nome maschile "Vittorio".

La lista dei primi 100 nomi è consultabile in appendice (v. Appendice – Liste di frequenza). È interessante notare che la somma delle frequenze dei primi 100 nomi ricopre circa il 16% della frequenza totale.

Anche per questa analisi è stata operata poi una distinzione per aree geografiche (questa volta solo per territorio) e fasce di popolazione. Vengono di seguito elencati i primi 10 risultati per territorio:

	Nord-Est	Nord-Ovest	Centro	Sud	Isole
1°	Guglielmo Marconi	Guglielmo Marconi	Giuseppe Garibaldi	Vittorio Emanuele	Giuseppe Garibaldi
2°	Giuseppe Garibaldi	Giuseppe Garibaldi	Guglielmo Marconi	Giuseppe Garibaldi	Vittorio Emanuele
3°	Dante Alighieri	Giuseppe Mazzini	Giacomo Matteotti	Aldo Moro	Giuseppe Mazzini
4°	Giuseppe Verdi	Giuseppe Verdi	Antonio Gramsci	Umberto I	Guglielmo Marconi
5°	Aldo Moro	Dante Alighieri	Giuseppe Mazzini	Guglielmo Marconi	Alcide De Gasperi
6°	Giacomo Matteotti	Giacomo Matteotti	Aldo Moro	Giuseppe Mazzini	Aldo Moro
7°	Giuseppe Mazzini	Alessandro Manzoni	Dante Alighieri	Giovanni XXIII	Giacomo Matteotti
8°	Cesare Battisti	Cesare Battisti	Vittorio Emanuele	Dante Alighieri	Dante Alighieri
9°	Alcide De Gasperi	Aldo Moro	Giovanni XXIII	Alcide De Gasperi	Antonio Gramsci
10°	Antonio Gramsci	Alessandro Volta	Giuseppe Verdi	Cesare Battisti	Cristoforo Colombo

I risultati non presentano grandi differenze tra di loro e nei confronti della lista precedente. Anche in questo caso è stato eliminato il luogo “Vittorio Veneto”, che si trovava rispettivamente in sesta, quarta, ottava, sesta e decima posizione.

È interessante notare come gli unici pochissimi nomi femminili a comparire nelle liste siano nomi di sante (in particolare Santa Lucia, Sant’Anna e Madre Teresa di Calcutta), nomi di regine (Elena del Montenegro, Margherita di Savoia) e la scrittrice italiana vincitrice del Premio Nobel per la letteratura nel 1926 Grazia Deledda.

In appendice (v. Appendice – Liste di frequenza) è possibile consultare la lista dei primi 100 nomi femminili presenti nell’elenco completo delle named entities. Tra

questi nomi compaiono diversi nomi maschili (come “Vittorini” – probabilmente lo scrittore Elio Vittorini – in seconda posizione), erroneamente interpretati come nomi femminili durante il passaggio di associazione del genere.

I risultati per fasce di popolazione non apportavano sostanziali differenze all’analisi e non sono stati qui riportati. Tutti i dati di questa analisi sono stati inseriti in una cartella compressa, disponibile con licenza CC BY 4.0[4] sull’archivio Zenodo[5], al link: https://zenodo.org/record/5894837.

[4] Permette che altri copino, distribuiscano, mostrino ed eseguano copie dell’opera e dei lavori derivati da questa a patto che venga indicato l’autore dell’opera, con le modalità da questi specificate (Wikipedia, voce *Licenze Creative Commons*).

[5] Zenodo è un archivio open access per le pubblicazioni e i dati da parte dei ricercatori. È gestito dal CERN e rende possibile l’autoarchiviazione anche ai ricercatori il cui ente fosse privo di un deposito istituzionale o non ammettesse l’archiviazione di certi formati, come codice sorgente e open data (Wikipedia, voce *Zenodo*).

5. Sito web per l'accesso ai dati

In questo capitolo verranno spiegate le fasi di costruzione e le funzionalità del sito web "Odonomastica italiana", creato per permette una libera consultazione del dataset e dell'analisi del capitolo precedente.

5.1 Creazione del sito

Per poter rendere fruibile il dataset, si è deciso di trasferire i dati su un database e di collegare quest'ultimo ad un sito web che ne permetta una facile consultazione.

Per la creazione del database è stato usato il dbms[1] MySQL, un software supportato da molti sistemi e linguaggi di programmazione (Wikipedia, voce *MySQL*). Lo schema logico è stato creato mediante il sito Dbdiagram[2] ed è riportato di seguito:

[1] Un Database Management System ("DBMS" o "Sistema di gestione di basi di dati") è un sistema software progettato per consentire la creazione, manipolazione e interrogazione efficiente di database. È ospitato su architettura hardware dedicata oppure su semplice computer (Wikipedia, voce *Database management system*).

[2] Dbdiagram è uno strumento gratuito di semplice utilizzo che permette di creare uno schema logico mediante scrittura di codice (dbdiagram.io, *Home*, n.d.).

Sono state create cinque tabelle, relative a territori, regioni, province, comuni e vie. Il riempimento delle tabelle è stato eseguito mediante un programma scritto in Python.

Le informazioni relative alla posizione geografica delle vie sono state inserite nelle corrispondenti tabelle. In particolare, alla tabella "comune" sono stati aggiunti anche i dati relativi alla popolazione di ogni comune, e nella tabella via sono state riportate tutte le informazioni ricavate attraverso i vari passaggi dell'elaborazione del dataset.

Per la creazione del sito sono stati usati i linguaggi PHP[3], CSS[4], HTML[5] e JavaScript[6] (in particolare la libreria jQuery[7]).

5.2 Navigazione

Il sito è composto da una pagina di home e dalla pagina in cui vengono mostrati i risultati dell'analisi. Nella home, partendo dall'alto e procedendo verso il basso, si trovano:

- la definizione di odonomastica e una breve introduzione al sito;
- un bottone tramite il quale è possibile spostarsi alla pagina dell'analisi;
- una sezione in cui viene spiegato come visualizzare o scaricare i dati;
- due menù a tendina (il secondo del quale compare solo dopo aver compilato il primo) e due bottoni attraverso i quali è possibile scaricare o visualizzare parti del dataset, selezionando un comune, una provincia, una regione o un territorio a scelta;

[3] PHP è un popolare linguaggio di scripting generico particolarmente adatto allo sviluppo web, è veloce, flessibile e pragmatico (php.net, n.d.).

[4] CSS è il linguaggio utilizzato per stilare (descrivere come deve essere visualizzato) un documento HTML (W3Schools, *CSS Tutorial*).

[5] HTML (acronimo di "HyperText Markup Language" – "Linguaggio di contrassegno per gli Ipertesti") è un linguaggio di markup che permette di indicare come disporre gli elementi all'interno di una pagina (html.it, *Introduzione all'HTML*).

[6] JavaScript è un linguaggio di programmazione orientato agli oggetti e agli eventi, comunemente utilizzato nella programmazione Web lato client per la creazione, in siti web e applicazioni web, di effetti dinamici interattivi (Wikipedia, voce *JavaScript*).

[7] jQuery è una libreria JavaScript veloce e ricca di funzionalità, che rende cose come la manipolazione di documenti HTML, la gestione degli eventi, l'animazione e Ajax molto più semplici mediante un'API facile da usare (jquery.com, n.d.).

- un riquadro in cui è presente un elenco di termini e spiegazioni, utile a chiarire il significato dei nomi assegnati ad ogni colonna della tabella di visualizzazione/scaricamento dei dati.

Analizzando la pagina dedicata all'analisi si trovano:

- la definizione di antroponimia;
- una breve spiegazione delle analisi effettuate sul dataset;
- i risultati dell'analisi per genere (eseguita per suddivisioni geografiche e per fasce di popolazione), completi di grafici a barre e tabelle dei valori;
- i risultati dell'analisi di frequenza (eseguita per tutta Italia, per suddivisioni geografiche e per fasce di popolazione), completi di tabelle dei valori;
- una sezione finale in cui è possibile scaricare i files con gli elenchi completi delle named entities, in formato TXT[8], ordinate per frequenza discendente. Nel caso delle divisioni per territorio e fasce di popolazione i files riportano solo i primi 100 nomi. Per ulteriori analisi sull'intero dataset di nomi propri di persona è presente, infine, l'elenco completo delle named entities.

5.3 Funzionalità

Le principali funzionalità del sito sono quelle di scaricamento e visualizzazione, utilizzabili dalla home.

L'apertura del primo menù a tendina dà la possibilità di selezionare i dati per comune, provincia, regione o territorio. Dopo la prima scelta, tramite un secondo menù a tendina, si può scegliere un pacchetto di dati specifico. Premendo il bottone "Scarica" sul terminale del richiedente verrà scaricato un file CSV[9], premendo invece il bottone "Visualizza" verrà caricata dinamicamente una nuova pagina, attraverso la quale è possibile visualizzare i dati selezionati mediante una tabella

[8] Un file TXT è un documento di testo standard che contiene testo non formattato. È riconosciuto da qualsiasi programma di elaborazione testi e può essere elaborato dalla maggior parte degli altri programmi software (Wikipedia, voce *File di Testo*).

[9] Un file CSV (acronimo di "Comma-Separated Values" – "Valori separati da virgola è un file di testo che utilizza le virgole per separare i dati contenuti all'interno delle singole celle di una tabella. Il formato CSV costituisce uno dei modi più semplici per rappresentare dati in forma tabellare (html.it, *File CSV: cosa sono, come si aprono e come crearli*, 2019).

Excel. Ogni colonna della tabella è selezionabile e ordinabile in ordine alfabetico (A-Z) e ordine alfabetico inverso (Z-A), ogni cella può essere modificata dall'utente e l'ultima colonna di ogni tabella riporta l'elenco dei link di Wikidata associati alle entità delle vie. Cliccando su di un link si viene reindirizzati alla corrispondente pagina di Wikidata.

Considerate tutte le scelte di dati possibili, il sito web può generare in tutto circa 8.000 pagine di visualizzazione diverse.

Sia in caso di visualizzazione che in caso di scaricamento il processo attraverso il quale i dati vengono raccolti dal database è il medesimo, e consiste nell'esecuzione di una query che prende in entrata il valore dell'area geografica scelta e restituisce all'utente tutti i valori della tabella "via" delle vie appartenenti al comune, provincia, regione o territorio richiesto.

Conclusioni

Il primo obiettivo che questo studio si poneva era quello della creazione di un dataset che contenesse abbastanza informazioni da poter essere utilizzato per l'esecuzione di uno studio focalizzato sulle vie contenenti nomi propri di persona.

Data la massiccia mole di dati iniziali, ogni passaggio dell'arricchimento del dataset richiedeva obbligatoriamente un'elaborazione automatica. I passaggi in cui sono state impiegate tecniche di machine learning sono soggetti a possibilità di errore. È bene perciò tenere conto dell'accuratezza (o F1 score[1]) dei singoli processori NER: 88% quella di SpaCy e 87,92% quella di Stanza[2]. Per l'algoritmo di raggruppamento delle named entities non è stato possibile stabilire un valore di accuratezza.

Alla fine dell'elaborazione è stato comunque possibile non solo associare ai dati iniziali tutte le informazioni necessarie ad eseguire l'analisi, ma anche creare un dataset completo di tutte le vie d'Italia (relativo alle vie presenti su OpenSteetMap), che potrà continuare ad essere arricchito di nuove informazioni e utilizzato per ulteriori analisi in campo umanistico.

Proprio per questo scopo, alla creazione del dataset è stata affiancata la creazione di un sito web, che permettesse una consultazione dei risultati ottenuti e una libera fruizione del dataset mediante un semplice processo di visualizzazione e scaricamento dei dati.

[1] L'F1 score (nota anche come F-score o F-measure) è una misura dell'accuratezza di un test. La misura tiene in considerazione i valori di "precisione" e "recupero" del test (Wikipedia, voce *F1 Score*).

[2] Dati consultabili ai seguenti link:

- https://spacy.io/models/it
- https://stanfordnlp.github.io/stanza/performance.html

Appendice – Liste di frequenza

Italia (posizione. nome, frequenza)

1.Giuseppe Garibaldi, 3774
2.Guglielmo Marconi, 3217
3.Giuseppe Mazzini, 2366
4.Dante Alighieri, 2315
5.Vittorio Veneto, 2305
6.Giacomo Matteotti, 2255
7.Aldo Moro, 2229
8.Giuseppe Verdi, 2168
9.Vittorio Emanuele, 2120
10.Antonio Gramsci, 1899
11.Alcide De Gasperi, 1887
12.Cesare Battisti, 1878
13.Alessandro Manzoni, 1772
14.Giovanni XXIII, 1691
15.Leonardo da Vinci, 1465
16.Alessandro Volta, 1396
17.Enrico Fermi, 1353
18.Giacomo Leopardi, 1321
19.Giovanni Pascoli, 1310
20.Galileo Galilei, 1305
21.San Rocco, 1291
22.Cristoforo Colombo, 1276
23.Umberto I, 1252
24.Ugo Foscolo, 1158
25.San Francesco, 1096
26.Giacomo Puccini, 1089
27.Sandro Pertini, 1062
28.Gioacchino Rossini, 1018
29.Silvio Pellico, 954
30.Armando Diaz, 939
31.Nazario Sauro, 921
32.Salvo D'Acquisto, 888
33.Camillo Benso Conte di Cavour, 880
34.Francesco Petrarca, 865
35.Giotto, 831
36.Vincenzo Bellini, 831
37.Don Luigi Sturzo, 823
38.Giuseppe Di Vittorio, 817
39.San Giuseppe, 775
40.Filippo Turati, 752
41.Michelangelo Buonarroti, 733
42.Gaetano Donizetti, 731
43.Pietro Nenni, 731
44.Giovanni Ventitreesimo, 712
45.Pietro Mascagni, 701
46.Giovanni Paolo II, 680
47.San Lorenzo, 680
48.Mazzini, 676
49.Giovanni Falcone, 676
50.Don Giovanni Minzoni, 676
51.Santa Lucia, 663
52.Vittorio Alfieri, 657
53.Amerigo Vespucci, 657
54.Palmiro Togliatti, 647
55.Paolo Borsellino, 643
56.Marco Polo, 643
57.Nino Bixio, 641
58.Carlo Alberto Dalla Chiesa, 641
59.John Fitzgerald Kennedy, 641
60.Vittorini, 637

61.Rose, 637
62.Dante, 636
63.Goffredo Mameli, 633
64.Enrico Berlinguer, 628
65.San Giorgio, 607
66.Luigi Einaudi, 600
67.Bruno Buozzi, 594
68.Raffaello Sanzio, 590
69.Edmondo De Amicis, 579
70.Grazia Deledda, 576
71.Regina Margherita, 575
72.Croce, 574
73.Luigi Pirandello, 569
74.Torquato Tasso, 560
75.San Marco, 556
76.Sant'Anna, 554
77.Gabriele D'Annunzio, 554
78.Enrico Toti, 550
79.Benedetto Croce, 547
80.Guglielmo Oberdan, 544
81.Giovanni Verga, 542
82.Ludovico Ariosto, 525
83.Antonio Vivaldi, 514
84.Umberto, 512
85.Giuseppe Parini, 505
86.Carlo Alberto, 500
87.Fabio Filzi, 497
88.Arturo Toscanini, 496
89.Achille Grandi, 491
90.Francesco Crispi, 480
91.Antonio Meucci, 476
92.San Giacomo, 450
93.San Nicola, 442
94.San Sebastiano, 440
95.Eugenio Montale, 437
96.Giovanni Boccaccio, 437
97.Luigi Cadorna, 430
98.Giovanni Amendola, 429
99.Santo Stefano, 429
100.Salvatore Quasimodo, 426

Nomi femminili (posizione. nome, frequenza)

51.Santa Lucia, 663
60.Vittorini, 637
70.Grazia Deledda, 576
71.Regina Margherita, 575
76.Sant'Anna, 554
155.Maria Montessori, 393
161.Madre Teresa di Calcutta, 373
179.Santa Chiara, 344
196.Annunziata, 322
198.Regina Elena, 320
199.Margherita, 320
222.Adua, 284
236.Maddalena, 272
256.Ada Negri, 245
271.Anna Frank, 232
277.Marina, 230
283.Crocetta, 222
292.Colombaro, 218
293.Rosa, 218
304.Andrea Doria, 211
307.Martina, 211
311.Santa Barbara, 206
327.Stellina, 199
385.Silva, 164
437.Flavio Gioia, 144
438.Santa Rita, 144
462.Valeria, 138
474.Umberto Saba, 134
516.Matilde Serao, 124
517.Viola, 124
522.Santa Teresa, 123
535.Marinella, 122
537.Novella, 121
547.Ilaria Alpi, 119
555.Elsa Morante, 117
584.Nilde Iotti, 113
619.Edera, 106
622.Palma, 106
626.Santa Cecilia, 105
660.Piero della Francesca, 99
668.Eleonora Duse, 97
695.Giulia, 93
717.Rosa Luxemburg, 90
739.Maria, 87
740.Addolorata, 87
758.Olivarella, 85
779.Anita Garibaldi, 83
784.Santa Marta, 82
785.Diana, 82
786.Iris, 82
801.Santa Cristina, 80
811.Sibilla Aleramo, 79
824.Anna Magnani, 78
844.Julia, 76
857.Regina, 74
890.Alberto Mora, 72
893.Costanza, 71
894.Oriana Fallaci, 71
921.Claudia, 69
945.Doria, 67
1017.Serena, 62
1037.Salvator Rosa, 61
1060.Mafalda di Savoia, 59
1072.Francesca, 59

1116.Madonna di Fatima, 56
1120.Deledda, 56
1171.Speranza, 53
1209.Matilde di Canossa, 52
1239.Mora, 50
1261.Cà Bianca, 49
1267.Margherita di Savoia, 49
1302.Eleonora d'Arborea, 48
1312.Caterina Percoto, 48
1329.Maria Callas, 47
1350.Dora, 46
1360.Elena, 46
1373.Emanuela Loi, 45
1385.Irma Bandiera, 45
1420.Francesca Morvillo, 44
1472.Santa Elisabetta, 42
1493.Flora, 41
1511.Sibilla, 41
1537.Erica, 40
1550.Vittoria Nenni, 40
1630.Bianca, 38
1661.Atleti Azzurri d'Italia, 37
1668.Maria Ausiliatrice, 37
1676.Vittoria Colonna, 37
1736.Melissa, 36
1787.Emanuela Setti, 35
1792.Sirena, 35
1797.Eleonora D'Arborea, 35
1811.Pia, 34
1829.Artemisia Gentileschi, 34
1865.Alda Merini, 33
1878.Sofia, 33
1879.Adelaide Ristori, 33
1887.Maria Gaetana Agnesi, 33
1929.Principessa Mafalda, 32
1991.Anna Kuliscioff, 31

Bibliografia

Blasco, Andrea. 2015. *Gender of Italian Names* https://github.com/mrblasco/genderNamesITA (visitato il 15 gennaio 2022).

Dbdiagram, *Home* https://dbdiagram.io/home (visitato il 18 gennaio 2022).

Docs Italia, *Anagrafe nazionale numeri civici e strade urbane* https://docs.italia.it/italia/daf/pianotri-schede-bdin/it/stabile/anncsu.html (visitato il 15 gennaio 2022).

Honnibal, M., & Montani, I. 2017. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

html.it, *Analisi qualitativa e quantitativa: cos'è e come si fa* https://www.html.it/articoli/analisi-quantitativa-e-qualitativa/ (visitato il 16 gennaio 2022).

html.it, *Introduzione all'HTML* https://www.html.it/pag/16026/introduzione22/ (visitato il 18 gennaio 2022).

IBM, *Natural language Processing (NLP)* https://www.ibm.com/cloud/learn/natural-language-processing (visitato il 15 gennaio 2022).

INSPIRE Italia Registry, *DUG* http://registry.geodati.gov.it/dug (visitato il 15 gennaio 2022).

ISTAT, *Codici statistici delle unità amministrative territoriali: comuni, città metropolitane, province e regioni* https://www.istat.it/it/archivio/6789 (visitato il 15 gennaio 2022).

ISTAT *Popolazione residente al 1° gennaio* http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1 (visitato il 15 gennaio 2022).

jQuery, *Homepage* https://jquery.com/ (visitato il 18 gennaio 2022).

JSON, *Introducing JSON* https://www.json.org/json-en.html (visitato il 15 gennaio 2022).

Kuci, Anisa. *Presentazione ufficiale di OpenStreetMap · MERGE-it 2021* https://video.linux.it/videos/watch/4182dcea-1d3c-4adc-bf41-6ea490ab51df (visitato il 15 gennaio 2022).

Marcato, Carla. 2005. *Il lessico delle «aree di circolazione»*, in Mastrelli. 2005. pp. 63-75.

Marcato, Carla. 2011. *Enciclopedia dell'italiano*. Treccani. https://www.treccani.it/enciclopedia/odonimi_%28Enciclopedia-dell%27Italiano%29/ (visitato il 14 gennaio 2022).

OpenRefine, *Welcome* page https://openrefine.org/ (visitato il 15 gennaio 2022).

OpenStreetMap Wiki, voce *Key:ref* https://wiki.openstreetmap.org/wiki/Key:ref (visitato il 15 gennaio 2022).

OpenStreetMap Wiki, voce *Overpass API* https://wiki.openstreetmap.org/wiki/Overpass_API (visitato il 15 gennaio 2022).

OpenStreetMap Wiki, voce *Overpass API/Overpass QL* https://wiki.openstreetmap.org/wiki/Overpass_API/Overpass_QL (visitato il 15 gennaio 2022).

Oracle, *Cos'è il machine learning?* https://www.oracle.com/it/data-science/machine-learning/what-is-machine-learning/ (visitato il 19 gennaio 2022).

Pandas, *Blog* https://pandas.pydata.org/community/blog/ (visitato il 16 gennaio 2022).

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages.* In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

PHP, *Homepage* https://www.php.net/ (visitato il 18 gennaio 2022).

Provino, Andrea. 2019. *Cos'è NER: Named Entity Recognition* https://andreaprovino.it/ner-named-entity-recognition/ (visitato il 15 gennaio 2022).

Redhat, *I vantaggi delle interfacce di programmazione delle applicazioni* https://www.redhat.com/it/topics/api/what-are-application-programming-interfaces (visitato il 15 gennaio 2022).

Sarnataro, Raffaele e Contaldo, Michele. 2007. *Un approfondimento sulla variabilità delle classi e della popolazione residente dei comuni italiani* https://finanzalocale.interno.gov.it/docum/studi/varie/200707varclass.html (visitato il 16 gennaio 2022).

Treccani, voce *Antroponimia*
https://www.treccani.it/vocabolario/antroponimia/ (visitato il 14 gennaio 2022).
Treccani, voce *Odonomastica*
https://www.treccani.it/vocabolario/odonomastica/ (visitato il 14 gennaio 2022).
Treccani, voce *Urbanistica*
https://www.treccani.it/vocabolario/urbanistica/ (visitato il 14 gennaio 2022).
W3Schools, *CSS Tutorial*
https://www.w3schools.com/css/ (visitato il 18 gennaio 2022).
Whyte, Luke. 2019. *Group thousands of similar spreadsheet text cells in seconds*
https://towardsdatascience.com/group-thousands-of-similar-spreadsheet-text-cells-in-seconds-2493b3ce6d8d (visitato il 15 gennaio 2022).
Whyte, Luke. 2019. *TextPack*
https://github.com/lukewhyte/textpack (visitato il 15 gennaio 2022).
Wikidata, voce *Introduction*
https://www.wikidata.org/wiki/Wikidata:Introduction (visitato il 15 gennaio 2022).
Wikipedia, voce *Database management system*
https://it.wikipedia.org/wiki/Database_management_system (visitato il 18 gennaio 2022).
Wikipedia, voce *F1 Score*
https://it.wikipedia.org/wiki/F1_score (visitato il 21 gennaio 2022).
Wikipedia, voce *File di testo*
https://it.wikipedia.org/wiki/File_di_testo (visitato il 18 gennaio 2022).
Wikipedia, voce *JavaScript*
https://it.wikipedia.org/wiki/JavaScript (visitato il 18 gennaio 2022).
Wikipedia, voce *Licenze Creative Commons*
https://it.wikipedia.org/wiki/Licenze_Creative_Commons (visitato il 21 gennaio 2022).
Wikipedia, voce *MySQL*
https://it.wikipedia.org/wiki/MySQL (visitato il 18 gennaio 2022).
Wikipedia, voce *N-gramma*
https://it.wikipedia.org/wiki/N-gramma (visitato il 15 gennaio 2022).
Wikipedia, voce *Odonomastica*
https://it.wikipedia.org/wiki/Odonomastica (visitato il 14 gennaio 2022).
Wikipedia, voce *Pipeline*

https://it.wikipedia.org/wiki/Pipeline (visitato il 15 gennaio 2022).

Wikipedia, voce *Tf-idf*
https://it.wikipedia.org/wiki/Tf-idf (visitato il 15 gennaio 2022).

Wikipedia, voce *Zenodo*
https://it.wikipedia.org/wiki/Zenodo (visitato il 18 gennaio 2022).

Zucchi, Camilla. 2020. *La toponomastica femminile in Italia tra retaggi del passato e sfide del presente*.