Elementi di basi di dati e data mining

Progetto Biblioteca Lazzarotto

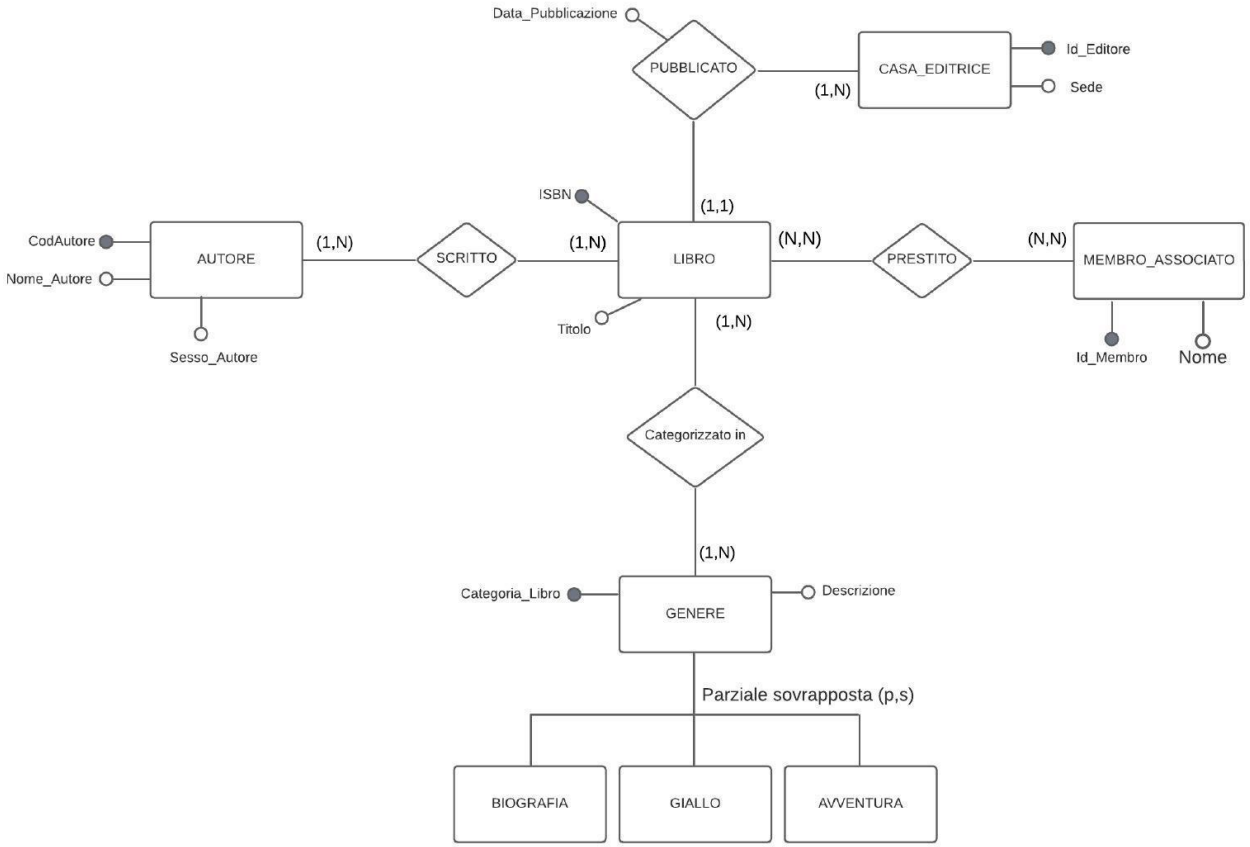
Francesco Lazzarotto - Matricola : 952248

Progetto

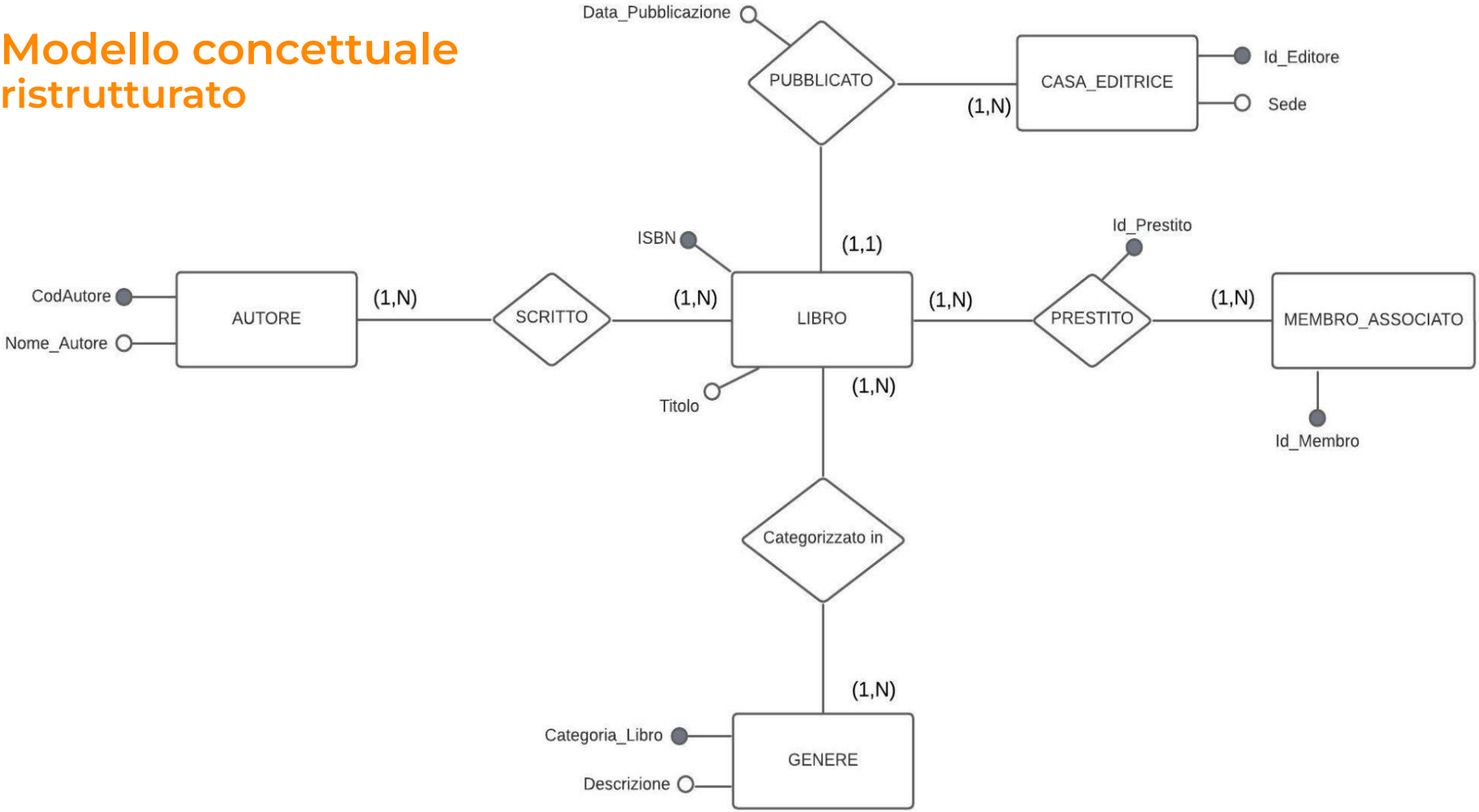
Il progetto consiste in un sistema di gestione dati per una biblioteca locale,integrando un database e un data warehouse. Le entità principali includono "Libro", "Genere", "Autore", "Casa Editrice", "Prestito", e "Membro". La relazione tra "Libro" e "Membro" è gestita attraverso una tabella di collegamento per tracciare i prestiti (Prestito). La relazione tra “Libro” e “Autore” è gestita attraverso una tabella di collegamento “Scritto” e la relazione tra “Libro” e “Genere” è gestita tramite la tabella “Categorizzato”. I Libri sono categorizzati per genere, sono scritti da un autore e vengono pubblicati da una casa editrice.

Il datawarehouse si concentra sulla correlazione fra il genere dei libri scritti dagli autori e la loro fascia d’età-sesso e sulla pubblicazione dei diversi generi in base alla localizzazione della sede delle diverse case editrici e all’anno di pubblicazione.

Modello concettuale



Modello concettuale
ristrutturato



Modello logico database

Autore (CodAutore, Nome_Autore, Cognome_Autore, Sesso_Autore, Anno_Nascita)

Scritto(ISBN, CodAutore)

Libro (ISBN, Titolo, Anno_Pubblicazione, Id_Casa_Editrice)

Categorizzato(ISBN, Genere)

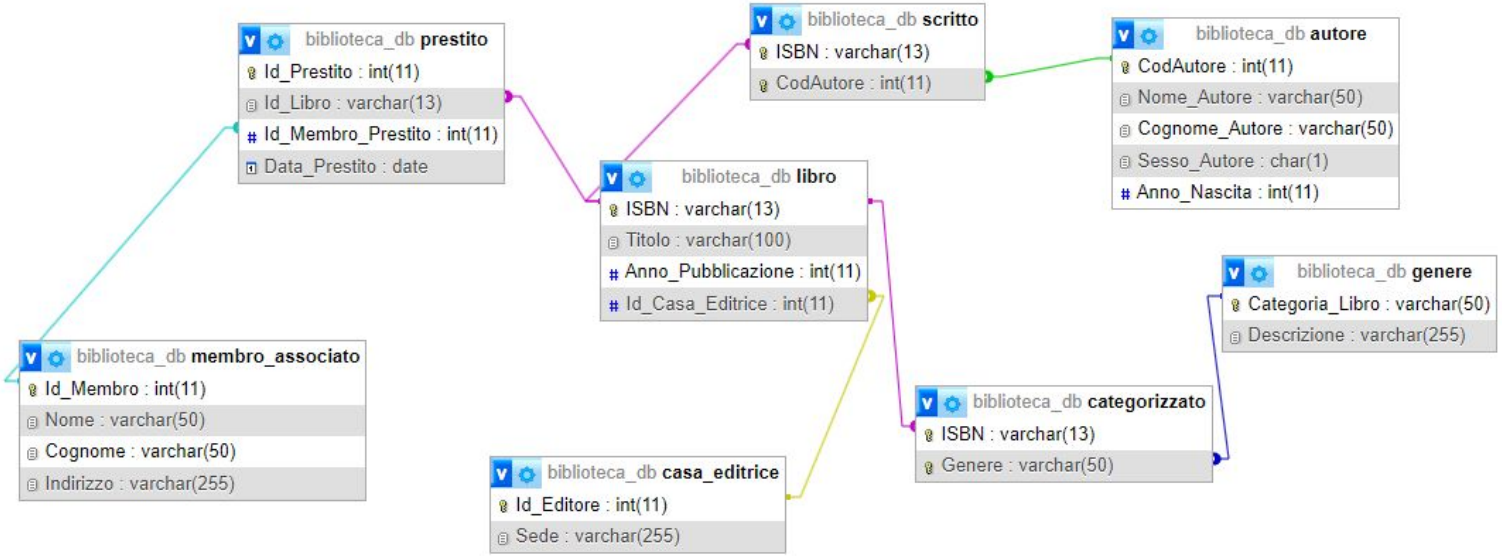
Genere (Categoria_Libro, Descrizione)

Casa_Editrice (Id_Editore, Sede)

Membro_Associato(Id_Membro, Nome, Cognome, Indirizzo)

Prestito(Id_Prestito, Id_Libro, Id_Membro_Prestito, Data_prestito)

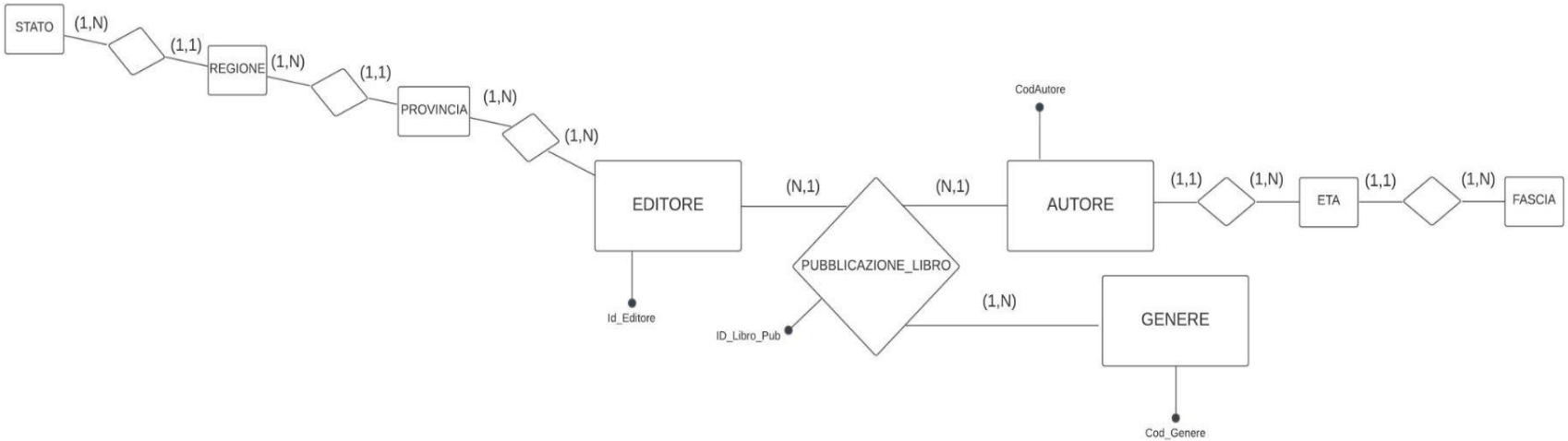
Designer delle relazioni - DB



Datawarehouse

Come anticipato prima il datawarehouse si concentra sulle pubblicazioni dei libri, ha una tabella delle misure Pubblicazione_libro con dei collegamenti diretti alle tabelle dimensionali normalizzate: Autore con sotto tabelle di dimensione età e fasce, la tabella dimensione Editore con sotto tabelle provincia, regione e stato e infine la tabella dimensione Genere con i dettagli del genere del libro pubblicato.

Modello concettuale a fiocco di neve



Modello logico datawarehouse

PUBBLICAZIONE_LIBRO(Id_Libro_Pub, Anno, Titolo, Id_Genere, Id_Autore, Id_Editore)

AUTORE(CodAutore, Id_eta, Sesso)

ETA(Cod_eta, Id_fascia , Eta_Autore)

FASCIA(Cod_Fascia, Fascia_Eta)

EDITORE(Cod_Editore, Id_Provincia, Nome)

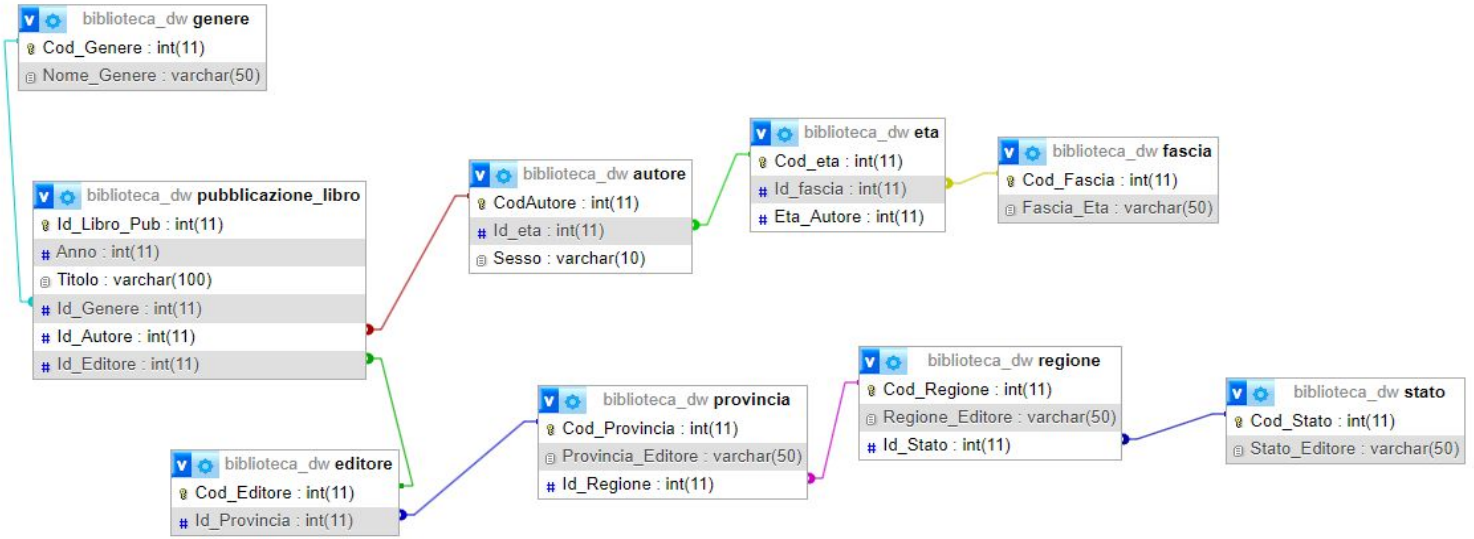
PROVINCIA(Cod_Provincia, Provincia_Editore, Id_Regione)

REGIONE(Cod_Regione, Regione_Editore, Id_Stato)

STATO(Cod_Stato, Stato_Editore)

GENERE(Cod_Genere, Nome_Genere)

Designer delle relazioni - DW



Operazioni OLAP - Tabella dei fatti

id_libro	Sesso	Eta_Autore	fascia_eta	nome_editore	Anno	Titolo	genere	Provincia_Editore	Regione_Editore	Stato_Editore
1	Maschio	25	Giovane	Italiano Editore	2020	Il Segreto del Vento	Romanzo	Teramo	Abruzzo	Italia
2	Femmina	35	Adulto	Francese Editions	2019	La Magia Oscura	Fantasy	Bouches-du-Rhône	Provence-Alpes-Côte 'Azur	Francia
3	Maschio	40	Adulto	Ediciones Espanolas	2021	L'Ultimo Inverno	Thriller	Barcelona	Catalonia	Spagna
4	Femmina	60	Adulto	Italiano Editore	2018	Ombre nella Notte	Romanzo	Teramo	Abruzzo	Italia
5	Maschio	75	Anziano	Francese Editions	2022	Risveglio	Fantasy	Bouches-du-Rhône	Provence-Alpes-Côte 'Azur	Francia
6	Femmina	35	Adulto	Deutsche Verlag	2017	Il Codice Perduto	Mistero	Munich	Bavaria	Germania
7	Maschio	60	Adulto	United Kingdom Press	2023	L'Incanto del Mare	Avventura	London	England	Regno Unito
8	Femmina	25	Giovane	American Publishers	2015	Oltre il Confine	Sci-Fi	Los Angeles	California	Stati Uniti
9	Maschio	40	Adulto	Deutsche Verlag	2016	L'Ultima Frontiera	Mistero	Munich	Bavaria	Germania
10	Femmina	75	Anziano	United Kingdom Press	2022	Segreti del Passato	Avventura	London	England	Regno Unito
11	Femmina	35	Adulto	American Publishers	2019	Notte Infinita	Sci-Fi	Los Angeles	California	Stati Uniti
12	Maschio	60	Adulto	Deutsche Verlag	2020	Cronache del Futuro	Mistero	Munich	Bavaria	Germania
13	Femmina	25	Giovane	United Kingdom Press	2018	Sotto le Stelle	Avventura	London	England	Regno Unito
14	Maschio	40	Adulto	American Publishers	2021	Invisibile	Thriller	Los Angeles	California	Stati Uniti
15	Femmina	75	Anziano	Deutsche Verlag	2014	Risveglio dei Titani	Fantasy	Munich	Bavaria	Germania
16	Femmina	35	Adulto	United Kingdom Press	2013	Ombre del Passato	Romanzo	London	England	Regno Unito
17	Maschio	60	Adulto	American Publishers	2011	Viaggio nel Tempo	Romanzo	Los Angeles	California	Stati Uniti
18	Femmina	25	Giovane	Deutsche Verlag	2012	Nebbia del Mattino	Mistero	Munich	Bavaria	Germania
19	Maschio	40	Adulto	United Kingdom Press	2010	L'Alba del Destino	Fantasy	London	England	Regno Unito
20	Femmina	75	Anziano	American Publishers	2014	L'Ultimo Viaggio	Sci-Fi	Los Angeles	California	Stati Uniti

Slice

Operazioni OLAP

Conteggio di tutti i libri con genere = Fantasy per titolo —-- conteggio di tutti i libri per sesso dell'autore = femmina per titolo

Conteggio di id_libro	Etichette di colonna	
Etichette di riga	Fantasy	Totale complessivo
La Magia Oscura	1	1
L'Alba del Destino	1	1
Risveglio	1	1
Risveglio dei Titani	1	1
Totale complessivo	**4**	**4**

Conteggio di id_libro	Etichette di colonna	
Etichette di riga	Femmina	Totale complessivo
Il Codice Perduto	1	1
La Magia Oscura	1	1
L'Ultimo Viaggio	1	1
Nebbia del Mattino	1	1
Notte Infinita	1	1
Oltre il Confine	1	1
Ombre del Passato	1	1
Ombre nella Notte	1	1
Risveglio dei Titani	1	1
Segreti del Passato	1	1
Sotto le Stelle	1	1
Totale complessivo	**11**	**11**

Dice

Operazioni OLAP

Conteggio di tutti i libri con genere fantasy e età autore compresa fra 35 e 70 —-- conteggio di tutti i libri pubblicati da editori con sede in provincia di Barcellona nell'anno 2021

Conteggio di id_libro	Etichette di colonna			
Etichette di riga	35	40	75	Totale complessivo
Fantasy	1	1	2	4
Thriller		2		2
Totale complessivo	**1**	**3**	**2**	**6**

Conteggio di id_libro	Etichette di colonna	
Etichette di riga	2021	Totale complessivo
Barcelona	1	1
Totale complessivo	**1**	**1**

Roll-up Operazioni OLAP

Mostrare il conteggio di tutti i libri pubblicati per genere nei vari stati (roll-up)

Conteggio di id_libro	Etichette di colonna						
Etichette di riga	Francia	Germania	Italia	Regno Unito	Spagna	Stati Uniti	Totale complessivo
Avventura				3			3
Fantasy	2	1		1			4
Mistero		4					4
Romanzo			2	1		1	4
Sci-Fi						3	3
Thriller					1	1	2
Totale complessivo	**2**	**5**	**2**	**5**	**1**	**5**	**20**

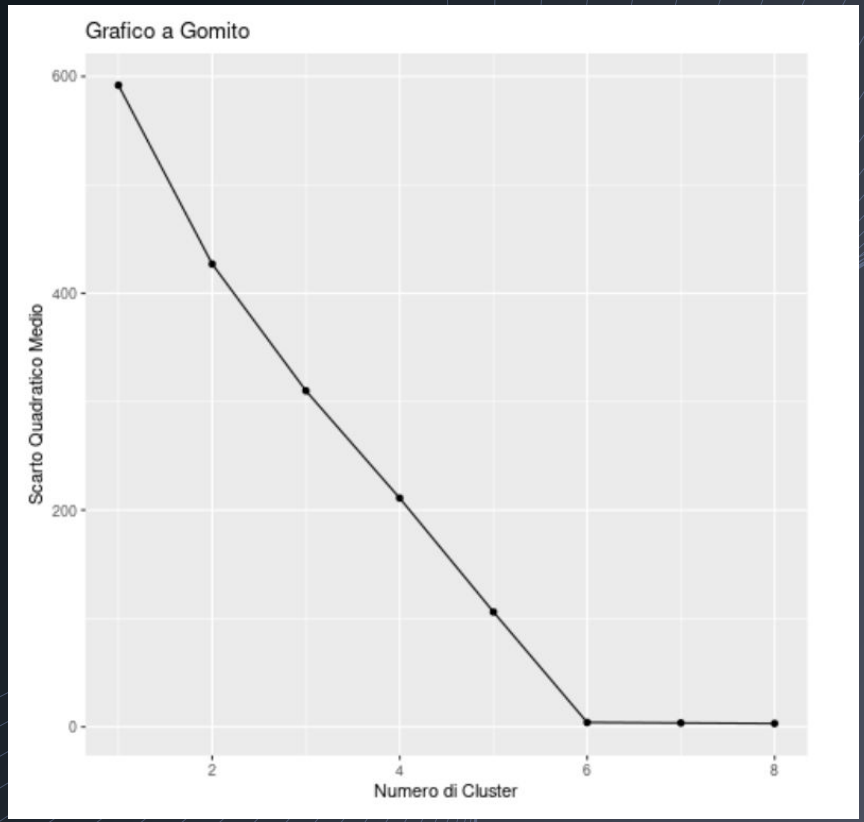
Drill-down Operazioni OLAP

Mostrare il conteggio di tutti i libri pubblicati nelle varie case editrici identificando la provincia nel quale le pubblicazioni sono avvenute (drill-down)

Conteggio di id_libro	Etichette di colonna						
Etichette di riga	**Barcelona**	**Bouches-du-Rhône**	**London**	**Los Angeles**	**Munich**	**Teramo**	**Totale complessivo**
American Publishers				5			5
Deutsche Verlag					5		5
Ediciones Espanolas	1						1
Francese Editions		2					2
Italiano Editore						2	2
United Kingdom Press			5				5
Totale complessivo	**1**	**2**	**5**	**5**	**5**	**2**	**20**

Data mining: Cluster

Attraverso vari tentativi ho costruito il grafico a Gomito che mi ha permesso di trovare il giusto numero di cluster con la quale effettuare l'analisi.
Come si può vedere dal grafico il gomito corrisponde a K=6



Data mining: Cluster

```
Number of iterations: 7
Within cluster sum of squared errors: 4.689104074745843

Initial starting points (random):

Cluster 0: 25,Romanzo
Cluster 1: 67,Mistero
Cluster 2: 55,Avventura
Cluster 3: 46,Fantasy
Cluster 4: 35,Sci-Fi
Cluster 5: 37,Sci-Fi

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5
               (701.0)    (104.0)    (188.0)    (100.0)    (104.0)    (103.0)    (102.0)
=============================================================================================
Eta_Autore      51.8902    26.8173    65.8085      54.45    47.2596    36.8835    69.1667
genere          Mistero    Romanzo    Mistero Avventura    Fantasy     Sci-Fi  Thriller
```

```
=== Model and evaluation on training set ===

Clustered Instances

0       104 ( 15%)
1       188 ( 27%)
2       100 ( 14%)
3       104 ( 15%)
4       103 ( 15%)
5       102 ( 15%)
```

Data mining: Cluster

Data mining: Cluster

-Misura di distanza = Distanza Euclidea
-Algoritmo = SimpleKMeans
-Numero di cluster = 6
-Seed = 10
-Scarto Quadratico = 4.689104074745843
-Descrizione e cluster più significativi:
L'analisi mostra una suddivisione basata su età degli autori e genere dei libri, con cluster che riflettono differenti preferenze e stili di scrittura associati alle diverse fasce d'età degli autori.
I cluster risultano tutti abbastanza significativi, offrendo una visione generale sulle preferenze del genere nella scrittura in relazione alla fascia d'età degli autori.
E quindi sottolineano una correlazione fra l'età di uno/a scrittore/scrittrice e un genere di scrittura favorito.
Es. età > 25 & < 30 genere -> Romanzo
età > 30 & < 40 genere -> Sci-Fi
età > 40 & < 50 genere -> Fantasy
età > 50 & < 60 genere -> Avventura
età > 60 generi -> Thriller e Mistero

Data mining: Albero decisionale

-Variabili di input = Età_Autore, Sesso

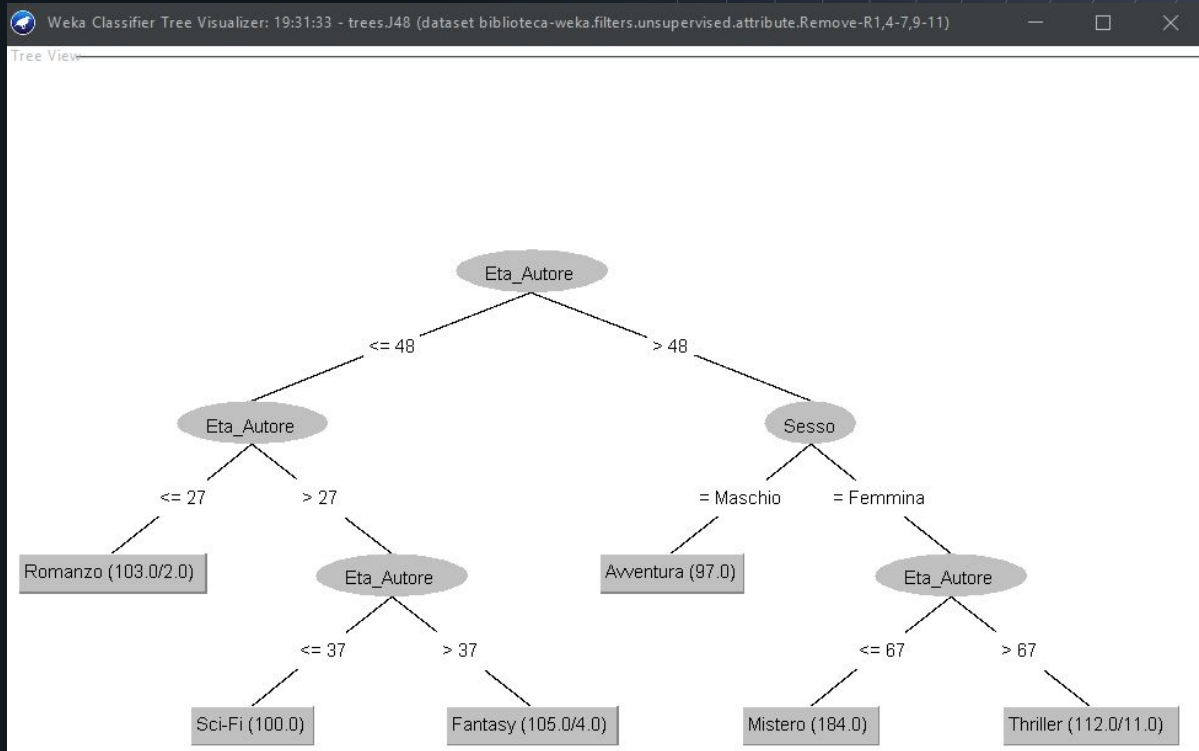
-Variabile di output = Genere

-Tecnica di valutazione = Training Set

-Precision = 0,977

-Recall = 0,976

-F-Measure = 0,976



Data mining: Albero decisionale

```
Correctly Classified Instances         684               97.5749 %
Incorrectly Classified Instances        17                2.4251 %
Kappa statistic                          0.9705
Mean absolute error                      0.0154
Root mean squared error                  0.0878
Relative absolute error                  5.6312 %
Root relative squared error             23.7317 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,971     0,003     0,981       0,971   0,976      0,972    0,994     0,965     Romanzo
                 0,971     0,007     0,962       0,971   0,967      0,961    0,992     0,946     Fantasy
                 0,990     0,018     0,902       0,990   0,944      0,935    0,990     0,898     Thriller
                 0,979     0,000     1,000       0,979   0,989      0,985    0,997     0,991     Mistero
                 0,970     0,000     1,000       0,970   0,985      0,982    0,997     0,984     Avventura
                 0,971     0,000     1,000       0,971   0,985      0,983    0,996     0,983     Sci-Fi
Weighted Avg.    0,976     0,004     0,977       0,976   0,976      0,972    0,995     0,965

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 101   2   1   0   0   0 |   a = Romanzo
   1 101   2   0   0   0 |   b = Fantasy
   1   0 101   0   0   0 |   c = Thriller
   0   1   3 184   0   0 |   d = Mistero
   0   0   3   0  97   0 |   e = Avventura
   0   1   2   0   0 100 |   f = Sci-Fi
```

Descrizione della matrice di confusione

I valori lungo la diagonale principale indicano le predizioni corrette per ciascuna classe, mentre le celle fuori dalla diagonale mostrano gli errori di classificazione. Ad esempio, il modello ha classificato correttamente 101 libri come 'Romanzo' (classe a) - quindi correttamente come veri positivi ma la classe a ha anche delle istanze classificate come falsi negativi (3 totali - 2 nella classe b e 1 nella classe c - Falsi positivi di Romanzo).
Questo è successo per tutte le classi:
-3 Falsi negativi nella classe b (Falsi positivi di Fantasy)
-1 Falso negativo nella classe c (Falso positivo di Thriller)
-4 Falsi negativi nella classe d (Falsi positivi di Mistero)
-3 Falsi negativi nella classe e (Falsi positivi di Avventura)
-3 Falsi negativi nella classe f (Falsi positivi di Fantascienza)

La somma dei falsi negativi è 17 come anche visualizzabile nel campo “Incorrectly Classified Instances”

```
  a   b   c   d   e   f   <-- classified as
101   2   1   0   0   0 |   a = Romanzo
  1 101   2   0   0   0 |   b = Fantasy
  1   0 101   0   0   0 |   c = Thriller
  0   1   3 184   0   0 |   d = Mistero
  0   0   3   0  97   0 |   e = Avventura
  0   1   2   0   0 100 |   f = Sci-Fi
```

```
Correctly Classified Instances         684               97.5749 %
Incorrectly Classified Instances        17                2.4251 %
Kappa statistic                          0.9705
Mean absolute error                      0.0154
Root mean squared error                  0.0878
Relative absolute error                  5.6312 %
Root relative squared error             23.7317 %
Total Number of Instances              701
```

Descrizione albero con regole

-IF Eta_Autore <= 48 THEN
-IF Eta_Autore <= 27 THEN
GENERE = 'Romanzo'
-IF Eta_Autore > 27 THEN
-IF Eta_Autore <= 37 THEN
GENERE = 'Sci-Fi'
-IF Eta_Autore > 37 THEN
GENERE = 'Fantasy'

-IF Eta_Autore > 48 THEN
-IF Sesso = 'Maschio' THEN
GENERE= 'Avventura'
-IF Sesso = 'Femmina' THEN
-IF Eta_Autore <= 67 THEN
GENERE = 'Mistero'
-IF Eta_Autore > 67 THEN
GENERE = 'Thriller'

```
Eta_Autore <= 48
|   Eta_Autore <= 27: Romanzo (103.0/2.0)
|   Eta_Autore > 27
|   |   Eta_Autore <= 37: Sci-Fi (100.0)
|   |   Eta_Autore > 37: Fantasy (105.0/4.0)
Eta_Autore > 48
|   Sesso = Maschio: Avventura (97.0)
|   Sesso = Femmina
|   |   Eta_Autore <= 67: Mistero (184.0)
|   |   Eta_Autore > 67: Thriller (112.0/11.0)

Number of Leaves  :     6

Size of the tree :      11
```

L'albero decisionale implementato ha estratto diverse regole utilizzando le caratteristiche di Età_Autore e Sesso. E' riuscito a predire correttamente il genere del libro scritto basandosi su queste caratteristiche di input.

Confronto con Naive Bayes

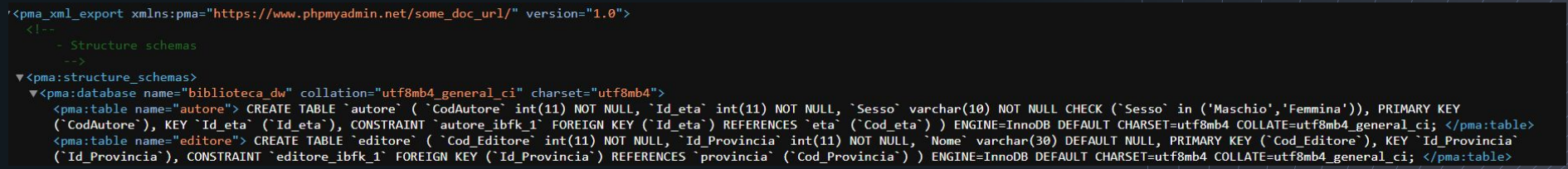
Le differenze principali riguardano l'errore assoluto medio, l'errore quadratico medio, l'errore assoluto relativo e l'errore quadratico relativo. Nel modello con algoritmo Naive Bayes, tutti questi valori sono di poco superiori , indicando una minore precisione e adattamento del modello ai dati di test. Inoltre c'è un leggero aumento di tempo per la preparazione del modello e il test.

Entrambe le tabelle riportano una buona accuratezza di istanze classificate correttamente (97.57% per entrambe). Inoltre la media di precision e recall rimane invariata.

In generale, basandosi sulle performance dei due modelli di classificazione, il modello fatto con J48 risulta di poco più accurato rispetto a quello con Naive Bayes.

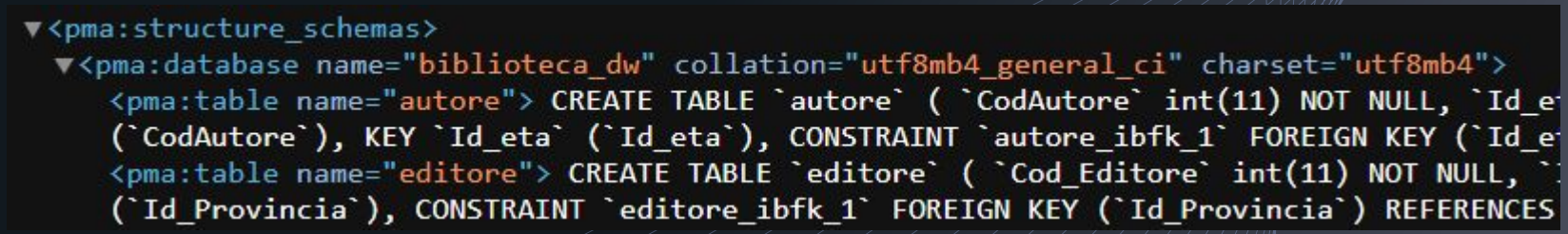
```
Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.04 seconds

=== Summary ===

Correctly Classified Instances         684               97.5749 %
Incorrectly Classified Instances        17                2.4251 %
Kappa statistic                          0.9705
Mean absolute error                      0.0526
Root mean squared error                  0.1353
Relative absolute error                 19.2148 %
Root relative squared error             36.5891 %
Total Number of Instances              701

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,971    0,003    0,981      0,971    0,976      0,972    0,985     0,964     Romanzo
                 0,971    0,007    0,962      0,971    0,967      0,961    0,973     0,956     Fantasy
                 0,990    0,018    0,902      0,990    0,944      0,935    0,982     0,893     Thriller
                 0,979    0,000    1,000      0,979    0,989      0,985    0,996     0,992     Mistero
                 0,970    0,000    1,000      0,970    0,985      0,982    0,983     0,976     Avventura
                 0,971    0,000    1,000      0,971    0,985      0,983    0,982     0,981     Sci-Fi
Weighted Avg.    0,976    0,004    0,977      0,976    0,976      0,972    0,985     0,964

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 101   2   1   0   0   0 |   a = Romanzo
   1 101   2   0   0   0 |   b = Fantasy
   1   0 101   0   0   0 |   c = Thriller
   0   1   3 184   0   0 |   d = Mistero
   0   0   3   0  97   0 |   e = Avventura
   0   1   2   0   0 100 |   f = Sci-Fi
```

XML - PROLOGO E NAMESPACE



`<pma_xml_export xmlns:pma="https://www.phpmyadmin.net/some_doc_url/" version="1.0">`
Questo è il prologo del documento XML del datawarehouse, viene indicata la versione (version="1.0") e il set di caratteri.
Qua viene anche dichiarato il namespace attraverso l'attributo `xmlns:pma` e con identificatore ="https://www.phpmyadmin.net/some_doc_url/".

Namespace viene poi utilizzato per esempio nelle parti del documento relative alla creazione delle tabelle. il namespace (pma:) viene utilizzato nel nome dell'elemento <table>. Quindi, il nome completo dell'elemento è <pma:table>. pma: viene utilizzato come parte del nome dell'elemento per indicare che fa parte del namespace dichiarato nel prologo (xmlns:pma="https://www.phpmyadmin.net/some_doc_url/"). Il namespace dichiarato non viene poi utilizzato successivamente negli elementi interni e negli attributi, e neanche successivamente nella definizione dei dati delle tabelle.



XML - NODO ELEMENTO, DI ATTRIBUTO E DI TESTO

```
▼<table name="autore">
    <column name="CodAutore">1</column>
    <column name="Id_eta">1</column>
    <column name="Sesso">Maschio</column>
  </table>
```

Nodo elemento -> Vi è un primo nodo elemento <table> contenente 3 diversi nodi elemento <column>. Ogni nodo elemento ha la propria chiusura </column> dopo il nodo di testo rispettivo e </table> a fine dei nodi elemento column

Nodo attributo -> L'attributo è associato all'elemento table attraverso name="autore" definendo il nome della tabella e per ogni column il nome del campo (name="CodAutore" - name="Id_eta" - name="Sesso")

Nodo di testo -> all'interno dei nodi di elemento column abbiamo dei nodi di testo.
In particolare nell'elemento <column name="CodAutore"> abbiamo il testo "1" - nell'elemento <column name="Id_eta"> abbiamo il nodo di testo "1" - nell'elemento con attributo <column name="Sesso"> abbiamo l'elemento "Maschio"

Text-mining

ID 1

UE/Renzi una commedia ,sono già d'accordo l'Italia ottiene flessibilità in cambio si tiene i migranti

TEXT
UE/Renzi una commedia ,sono già d'accordo l'Italia ottiene flessibilità in cambio si tiene i migranti
Text_clean
UE/Renzi una commedia ,sono già d'accordo l'Italia ottiene flessibilità in cambio si tiene i migranti
Tokens
'UE' , '/' , 'Renzi' , 'una' , 'commedia' , ',sono' , 'già', 'd'accordo', 'l'italia', 'ottiene', 'flessibilità', 'in', 'cambio', 'si' , 'tiene', 'i', 'migranti'
Tokens_lower
'ue' '/', 'renzi' , 'una' , 'commedia' , ',sono' , 'già', 'd'accordo', 'l'italia', 'ottiene', 'flessibilità', 'in', 'cambio', 'si' , 'tiene', 'i', 'migranti'
Tokens_stopword
'ue' , 'renzi', 'commedia' , 'd'accordo', 'l'italia', 'ottiene', 'flessibilità', 'cambio' , 'tiene', 'migranti'
Lemmas
'ue' , 'renzi' , 'commedia' , 'accordo' , 'italia' , 'ottenere' , 'flessibilità' , 'cambiare' , 'tenere' , 'migrante'

Text_clean -> In questo caso il testo è rimasto invariato non avendo per esempio link o username da decidere se eliminare.

Tokens -> La frase in questione riportava alcune parole che potevano essere interpretate in modo diverso, in particolare ‘UE/Renzi’ , ‘d’accordo’ , l’italia’ . La mia scelta è stata quella di tenere come unità singole ‘d’accordo’ e l’italia’ e invece UE/Renzi come due entità diverse separate dall’unità ‘/’. Questa scelta è stata presa per far si le che le due parole vengano poi analizzate separatamente e indipendentemente l’una dall’altra

Tokens_lower -> il testo viene portato in lowercase dove c’è n’è bisogno. In questo caso UE -> ue e Renzi -> renzi

Tokens_stopword -> il testo viene semplificato attraverso l’eliminazione delle stopwords (prese dall’elenco dato su moodle) e dei segni di punteggiatura. In questo caso ho deciso di eliminare il segno ‘/’ che separava le parole UE/Renzi poiché non ritengo che sia d’importanza e lo considero come un carattere di punteggiatura non informativo e quindi non pertinente alla successiva analisi.

Lemmas -> Le parole derivanti dai precedenti passaggi sono ora portati a una forma base, questo passaggio è stato effettuato in particolare dopo la tokenizzazione e l’eliminazione delle stopwords. Nel caso specifico di 'ue' (Unione Europea) e 'renzi' (un cognome), possono essere considerate eccezioni poiché sono specifici e potrebbero non avere una forma base. Ho deciso quindi di trattarli come termini speciali e mantenere la loro forma originale.

Normalizzazione 2 testo

ID 2

In pratica è sempre colpa dei neri/migranti anche quando vengono uccisi. Perfetto. URL

TEXT
In pratica è sempre colpa dei neri/migranti anche quando vengono uccisi. Perfetto. URL
Text_clean
In pratica è sempre colpa dei neri/migranti anche quando vengono uccisi. Perfetto.
Tokens
'In', 'pratica', 'è' , 'sempre' , 'colpa' , 'dei' , 'neri' , '/' , 'migranti' , 'anche' , 'quando' , 'vengono' , 'uccisi' , '.' , 'Perfetto' , '.'
Tokens_lower
'in', 'pratica', 'è' , 'sempre' , 'colpa' , 'dei' , 'neri' , '/' , 'migranti' , 'anche' , 'quando' , 'vengono' , 'uccisi' , '.' , 'perfetto' , '.'
Tokens_stopword
'pratica', 'colpa' , 'neri' , ' migranti', 'vengono', 'uccisi' , 'perfetto'
Lemmas
'pratica' , 'colpa' , 'nero' , 'migrante' , 'venire' , 'uccidere' , 'perfetto '

Text_clean -> in questo caso ho deciso di togliere URL dal corpo

Tokens -> Nel passaggio di tokenizzazione, ho trovato, come nel caso precedente, due parole separate da un simbolo ‘/’, anche qui ho deciso di separare le unità e prenderle come parole signole. (‘neri’ , ‘/’ , ‘migranti’).

Tokens_lower -> In questo passaggio il testo viene portato in lowercase dove c’è n’è bisogno -> ‘In’ – ‘in’ e

‘Perfetto’ – ‘perfetto’

Tokens_stopword -> il testo viene semplificato attraverso l’eliminazione delle stopwords (prese dall’elenco dato su moodle) e dei segni di punteggiatura. In questo caso ho deciso di eliminare, come precedentemente, il segno ‘/’ che separava le parole neri/migranti poiché non ritengo che sia d’importanza e lo considero come un carattere di punteggiatura non informativo e quindi non pertinente alla successiva analisi.

Lemmas -> Le parole derivanti dai precedenti passaggi sono ora portati a una forma base, questo passaggio è stato effettuato in particolare dopo la tokenizzazione e l’eliminazione delle stopwords.

Per questo caso ho trovato qualche difficoltà in più nel capire se i miei passaggi potessero essere corretti. La frase "In pratica è sempre colpa dei neri/migranti anche quando vengono uccisi. Perfetto. URL" è ambigua a causa dell'espressione potenzialmente controversa e stereotipata. Il processo di normalizzazione, incluso l'eliminazione delle stopwords e la lemmatizzazione, potrebbe alterare il significato originale. Il contesto completo, inclusi eventuali contenuti esterni ("URL"), nonostante l’eliminazione iniziale (eliminato per differenziare il text-clean dall’esercizio precedente), è a mio parere essenziale per non distorcere il significato della frase (che non contiene, secondo me, hate-speech) cioè che “la colpa (secondo chi stereotipizza) è sempre dei neri anche quando vengono uccisi e magari non hanno fatto niente” – la frase potrebbe essere infatti una risposta a un commento e/o all’asserzione di qualcuno d’altro (con probabile hate-speech)

Attributi semplici - ID 1

Text	tokens	Numero_parole	Numero_frasi	!	?	#
UE/Renzi una commedia ,sono già d'accordo l'Italia ottiene flessibilità in cambio si tiene i migranti	'UE' , '/' , 'Renzi' , 'una' , 'commedia' , ',sono' , 'già', 'd'accordo', 'l'italia', 'ottiene', 'flessibilità', 'in', 'cambio', 'si' , 'tiene', 'i', 'migranti'	16	2	0	0	0

Numero_parole -> Al netto della punteggiatura e simbolatura il conteggio delle parole è 16.

Numero_frasi -> Qua le frasi sono due, separate da una virgola (sia se effettuato su testo non ripulito che tokenizzato). UE/Renzi una commedia (1) , (virgola che separa il senso della frase) – sono già d'accordo l'italia ottiene flessibilità in cambio si tiene i migranti (2)

Presenza_! -> Non sono presenti punti esclamativi quindi assume valore 0.

Presenza_? -> Non sono presenti punti interrogativi quindi assume valore 0.

Numero_# (su testo intero) -> Il numero di hashtag presenti è 0.

Attributi semplici - ID 2

Text	tokens	Numero_parole	Numero_frasi	!	?	#
In pratica è sempre colpa dei neri/migranti anche quando vengono uccisi. Perfetto. URL	'In', 'pratica', 'è' , 'sempre' , 'colpa' , 'dei' , 'neri' , '/' , 'migranti' , 'anche' , 'quando' , 'vengono' , 'uccisi' , '.' , 'Perfetto' , '.'	13	2	0	0	0

Numero_parole -> Al netto della punteggiatura il numero di parole è 13.

Numero_frasi -> Ho qua suddiviso il testo in due frasi, distinte dal punto dopo uccisi. È comunque un caso ambiguo e interpretabile, si potrebbe pensare che le frasi siano addirittura 3 contando anche il punto dopo perfetto (se si eseguono gli attributi semplici sul testo non ripulito e non tokenizzato) -> uccisi. (1) - Perfetto. (2) - URL (3)

Presenza_! -> Non sono presenti punti esclamativi quindi assume valore 0.

Presenza_? -> Non sono presenti punti interrogativi quindi assume valore 0.

Numero_# (sul testo intero) -> Il numero di hashtag presenti è 0.

Bag of Words

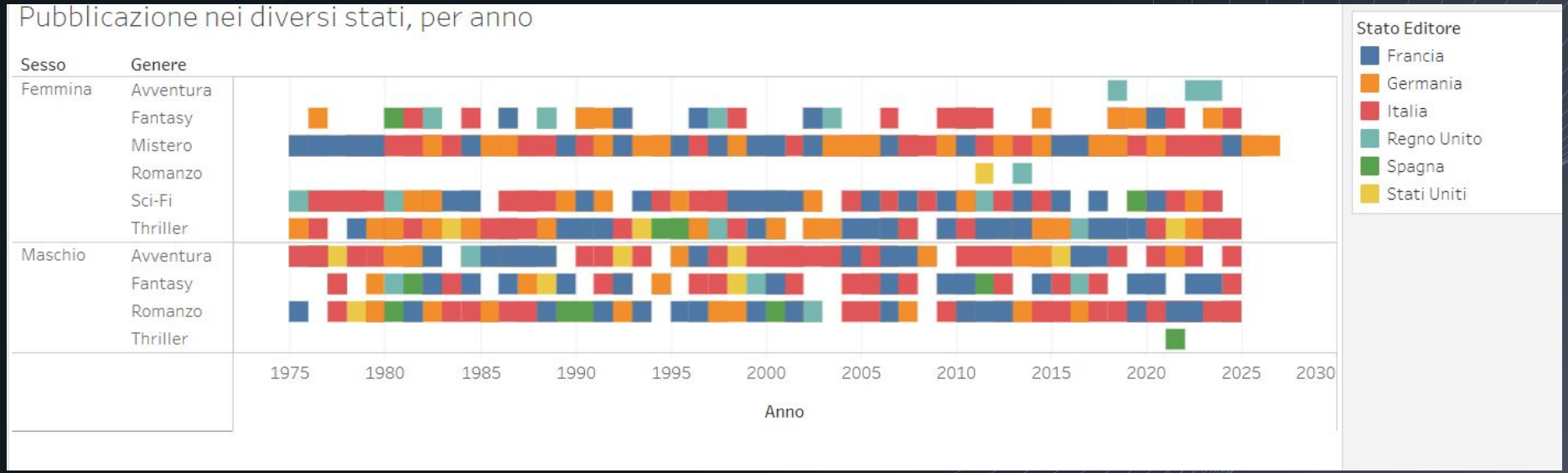
LEMMAS	ottenere	flessibilità	cambiare	tenere
'pratica' , 'colpa' , 'nero' , 'migrante' , 'venire' , 'uccidere' , 'perfetto '	0	0	0	0
'ue' , 'renzi' , 'commedia' , 'accordo' , 'italia' , 'ottenere' , 'flessibilità' , 'cambiare' , 'tenere' , 'migrante'	1	1	1	0

L'unico lemma presente in tutti e due i lemmas è: 'migrante'. Gli altri sono invece tutti differenti.

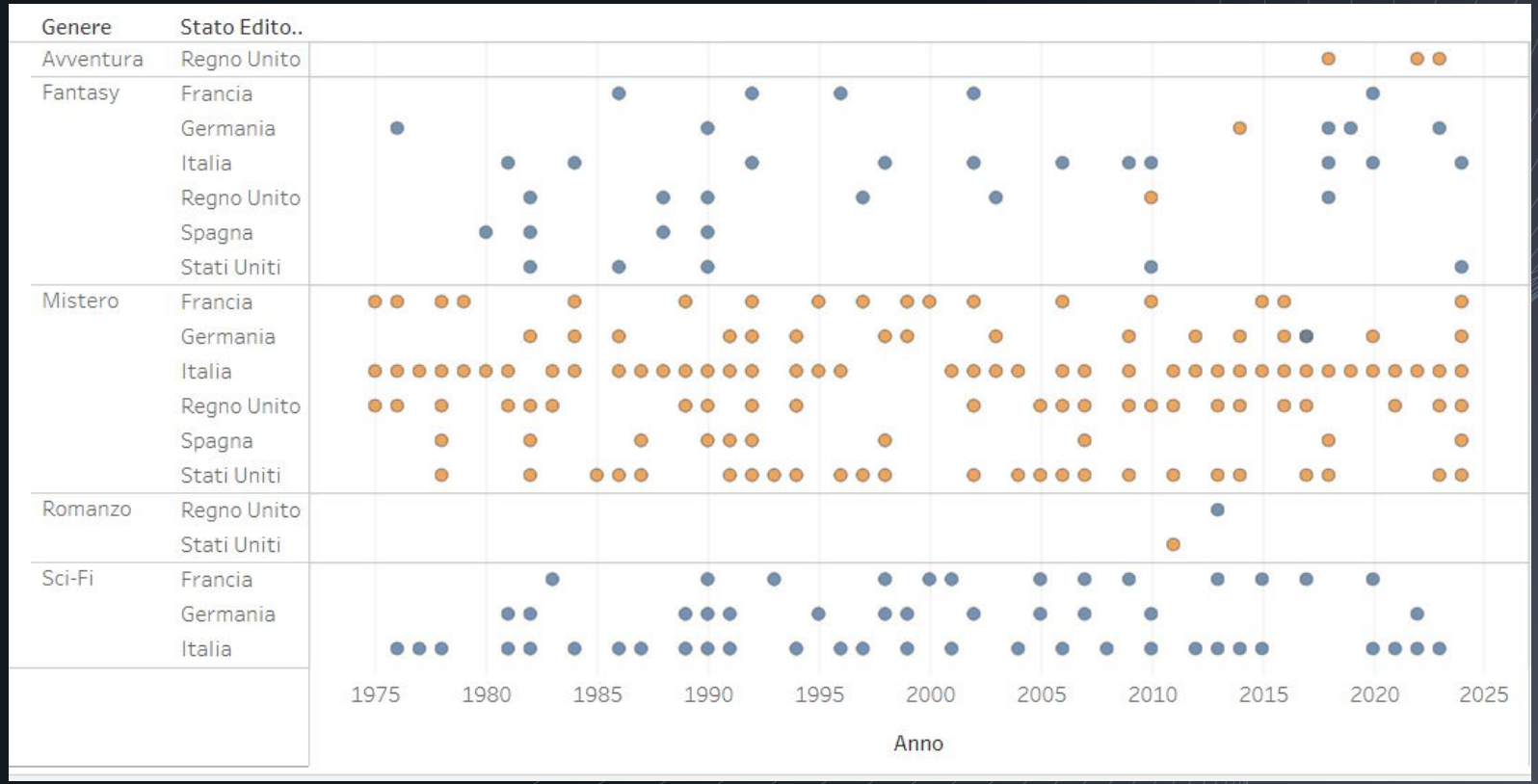
BAG OF WORDS (BOW)

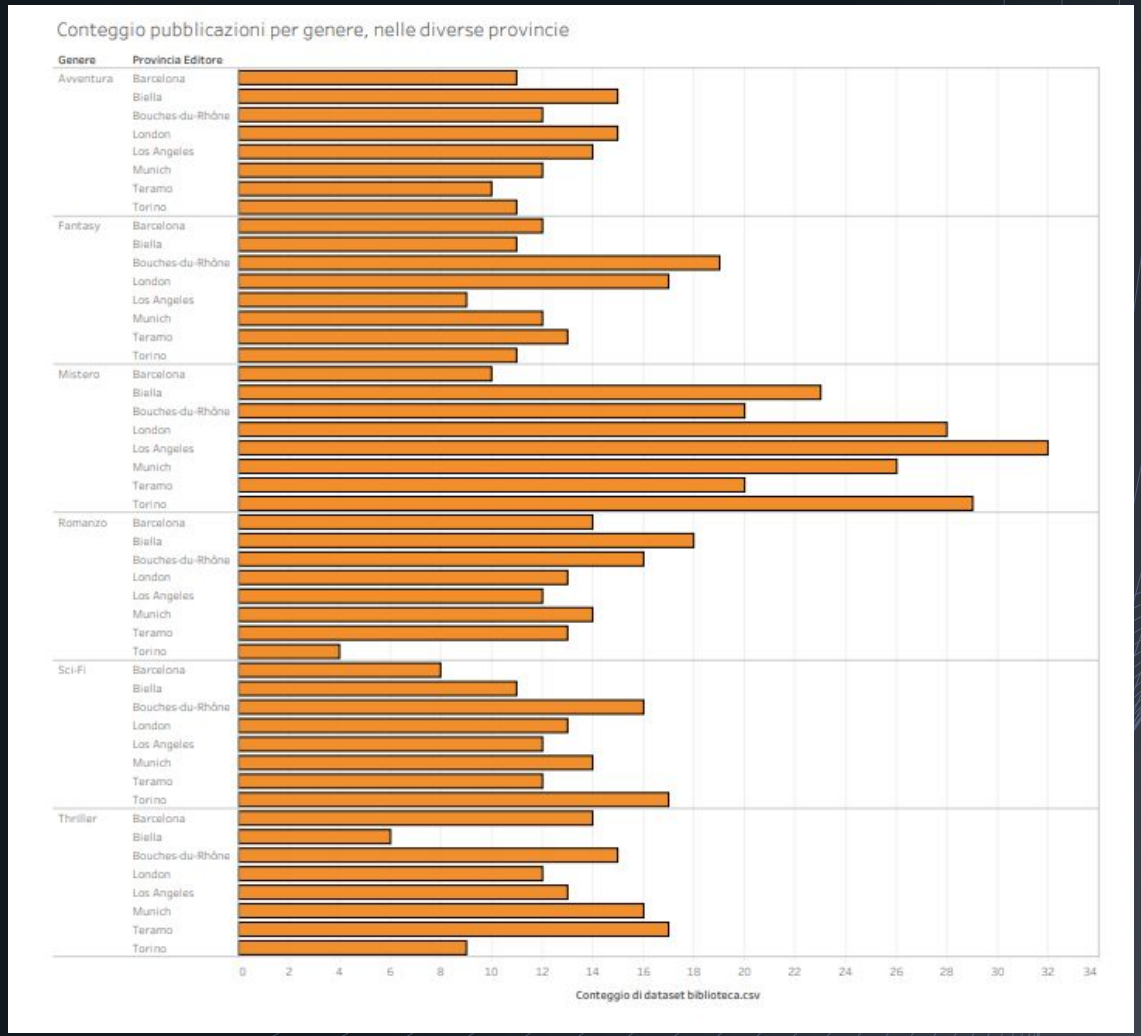
LEMMAS	pratica	colpa	nero	migrante	venire	uccidere	perfetto	ue	renzi	commedia	accordo
'pratica', 'colpa' , 'nero' , 'migrante' , 'venire' , 'uccidere' , 'perfetto '	1	1	1	1	1	1	1	0	0	0	0
'ue' , 'renzi' , 'commedia' , 'accordo' , 'italia' , 'ottenere' , 'flessibilità' , 'cambiare' , 'tenere' , 'migrante'	0	0	0	1	0	0	0	1	1	1	1

Information Visualization
Software Tableau



Information Visualization
Software Tableau




Conteggio pubblicazioni per genere, nelle diverse provincie
Genere
Provincia Editore
Avventura
Barcelona
Biella
Bouches-du-Rhône
London
Los Angeles
Munich
Teramo
Torino
Fantasy
Barcelona
Biella
Bouches-du-Rhône
London
Los Angeles
Munich
Teramo
Torino
Mistero
Barcelona
Biella
Bouches-du-Rhône
London
Los Angeles
Munich
Teramo
Torino
Romanzo
Barcelona
Biella
Bouches-du-Rhône
London
Los Angeles
Munich
Teramo
Torino
Sci-Fi
Barcelona
Biella
Bouches-du-Rhône
London
Los Angeles
Munich
Teramo
Torino
Thriller
Barcelona
Biella
Bouches-du-Rhône
London
Los Angeles
Munich
Teramo
Torino
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
Conteggio di dataset biblioteca.csv