# Text mining per l'analisi delle fiabe più famose dei fratelli Grimm

Fabiana Rapicavoli

Luglio 2021

## 1 Introduzione

In Internet possiamo leggere molte teorie sulla crudeltà e l'oscurità dietro tutti i racconti dei fratelli Grimm.
Ho svolto una semplice analisi in R per verificare questa tesi. In particolare, ho preso in considerazione sei storie: "Biancaneve"; "Cenerentola", "I musicanti di Brema", "Cappuccetto rosso", "Raperonzolo" e "Hansel e Gretel".
Quindi, i racconti dei fratelli Grimm sono così oscuri come sostiene Internet?

## 2 Analisi delle parole più frequenti

In un'analisi preliminare, ho trovato le parole più ricorrenti in tutti i racconti:

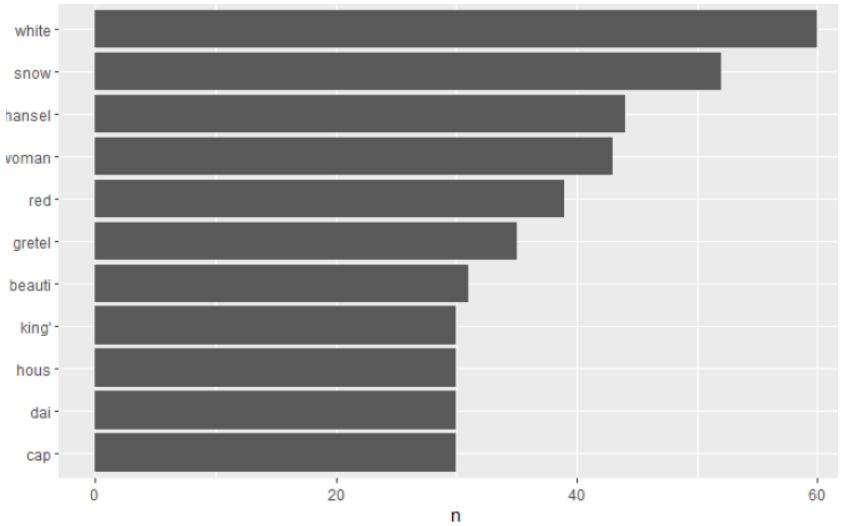


Figure 1: parole più frequenti

Molte parole in questo grafico (tra cui bianco/a, neve, Gretel, cappuccetto ecc) sono riconoscibili e si può facilmente intuire da quale racconto provengano. Tra queste parole possiamo notare "re" e "bellezza", riconducibili ai temi delle fiabe, era un risultato prevedibile.

È interessante notare che alcune parole (come hous o dai) sono scritte in modo errato, probabilmente perché allora quella era la forma corretta e hanno subito una mutazione successiva.

# 3 Analisi dei sentimenti

Tramite il text mining, è anche curioso affrontare l'analisi dei sentimenti.

Il linguaggio di programmazione R prevede l'utilizzo di tre dizionari che associano a ogni parola un sentimento:

- **bing**, che associa alla parola un'emozione positiva o negativa.
- **AFINN**, che associa alla parola numero intero compreso tra -5 a 5, che classifica anche l'intensità dell'emozione stessa, dove se il valore è -5, la parola è estremamente negativa, se il valore è la parola è estremamente positiva e se il valore è 0 la parola è neutra.
- **nrc**, dove ogni parola viene presa in considerazione due volte: una per il polo positivo o negativo e una per la classificazione delle emozioni (rabbia, paura, anticipation, fiducia, sorpresa, tristezza, gioia, e disgusto).

## 3.1 Parole più frequenti positive e negative

La prima classificazione che ho svolto, anche in funzione della precedente, è la divisione delle parole più frequenti tra l'influenza positiva e negativa associata, attraverso il dizionario bing:

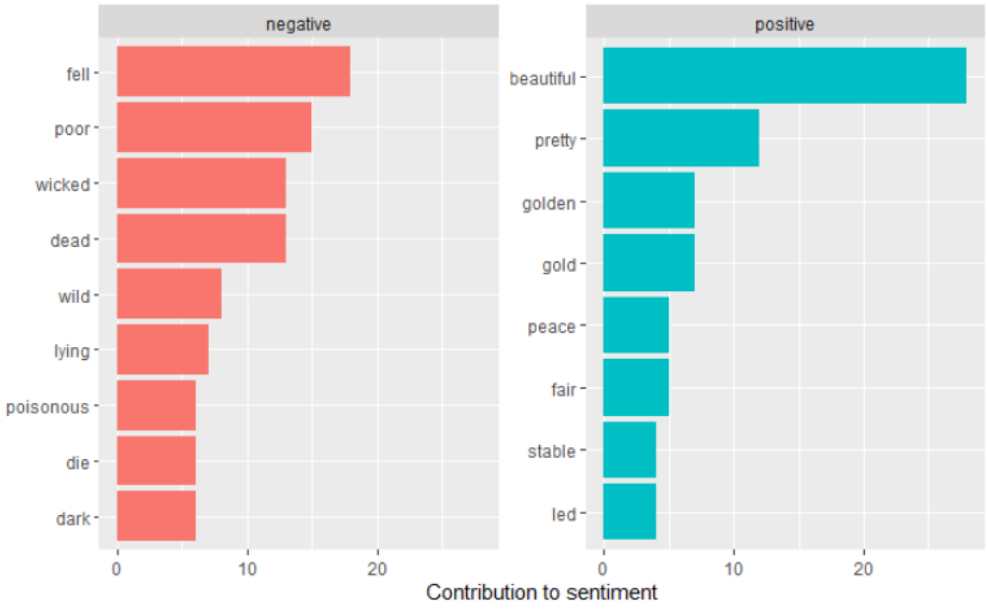


Figure 2: parole più frequenti positive e negative

Le parole positive sembrano avere un'occurrenza maggiore di quelle negative.

Singolarmente, le parole non suggeriscono la fiaba di provenienza, contrariamente a prima. Questo significa che le parole caratterizzanti le storie sono di entità neutra.

Le parole positive più frequenti rigurdano la bellezza, la ricchezza e la pace, mentre le parole negative più frequenti rigurdano la povertà, la morte, l'oscurità e il veleno.

## 3.2 Le parole più frequenti per ogni sentimento

In questa fase, ho sfruttato il dizionario nrc per analizzare le parole più frequenti per ogni tipo di emozione:

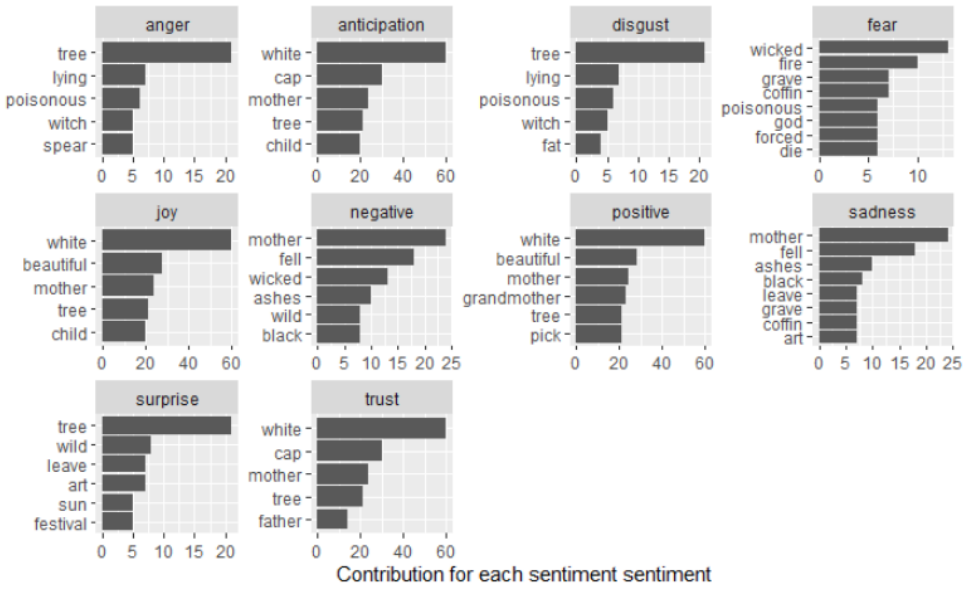


Figure 3: parole più frequenti per emozione

Dall'immagine si può subito notare con piacere che le parole positive superano in frequenza del più del doppio le parole negative (come predetto precedentemente); anche questo risultato è facilmente riconducibile alle fiabe ed è un fattore estrememanete positivo, dato che il range d'età del pubblico medio è giovane.

Le parole più frequenti si dividono tra i sentimenti di anticipazione, gioia e fiducia, tutte e tre emozioni positive, ma le parole negative, seppur meno frequenti, sembrano in quantità maggiore, soprattutto tra le emozioni di paura e tristezza.

È curioso notare come la parola "mother" (mamma) è presenza sia nell'emozione di gioia, fiducia, anticipazione e, in generale, di positività, sia nell'emozione di tristezza e di negatività.
Nel nostro caso, in un contesto fiabesco, spesso la figura genitoriale è più una figura di un antagonista, che di amico (ammesso che i protagonisti non siano orfani), ma è curiosa questa generale doppia associazione per la parola "mamma".

## 3.3 Andamento dei sentimenti all'interno della storia

Durante un'analisi dei sentimenti, una delle operazioni più interessanti da svolgere è verificare come le emozioni si evolvono all'interno della storia, utilizzando, invece, il dizionario bing:

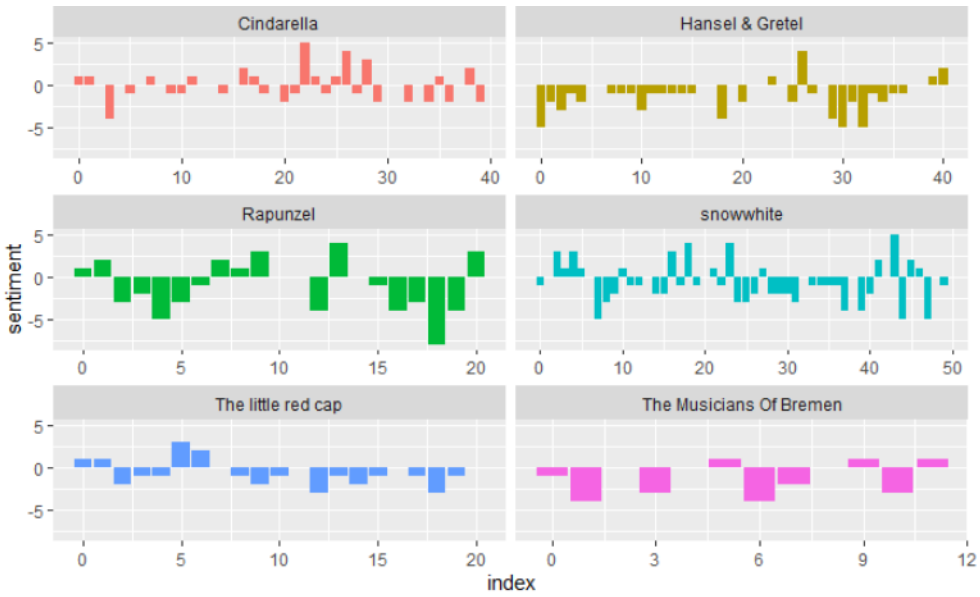


Figure 4: Andamento dei sentimenti

Dalla premesse, mi aspettavo che i libri fossero estremamente negativi. Infatti ho voluto svelgere questa ricerca per confermare la mia teoria.
Nella maggior parte delle storie, le barre positive si concentrano nella parte finale e in una/poche parti verso il centro.

"Rapunzel" è la fiaba con le barre negative più lunghe, quindi più denso di negatività.
Non mi aspettavo questo risultato, anche perché da piccola ho conosciuto "Raperonzolo" grazie al film della Barbie, pieno di magia, e quello della Disney, pieno di avventura.

"Snowwhite" ha la maggior parte delle barre negative, il che è un risultato prevedibile, anche se mi aspettavo che "Hansel  Gretel" lo superasse.

Strano che non ci sia un'happy ending in tutte le storie ed è presente solo nelle storie di "Rapunzel", "Hansel  Gretel" e "i musicanti di Brema"; probabilmente i fratelli Grimm hanno usato più spesso parole per lo più neutre, causando la mancanza totale della barra.

# 4 Legge di Zipf

La legge di zipf sostiene che in un qualsiasi testo di qualsiasi argomento ci sono poche parole usate tante volte e tante usate poche. Questi testi confermano la teoria:

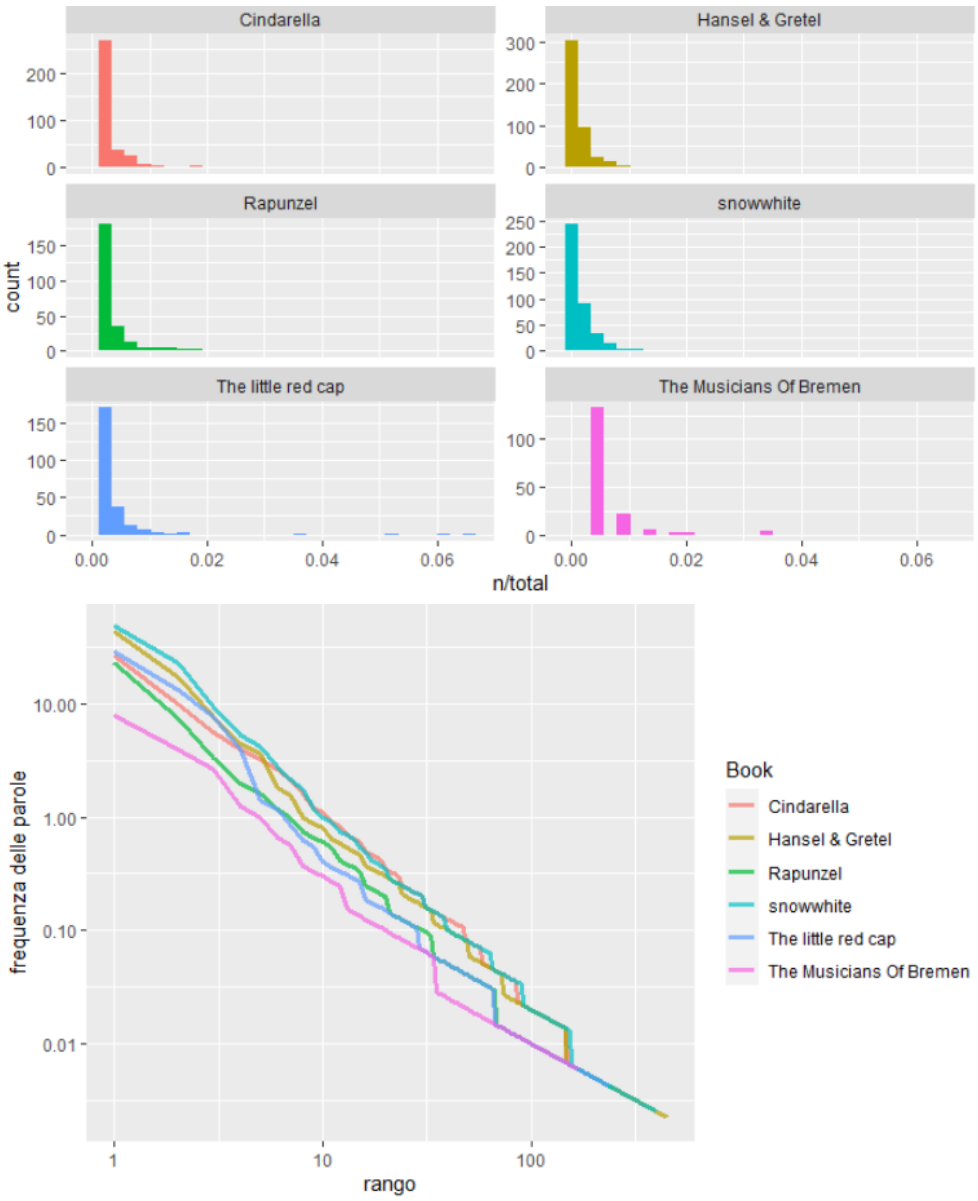


Figure 5: Legge di Zipf

# 5 TF-IDF

Una cosa importante del text mining è associare un certo grado di importanza alle parole, anche per capire l'argomento principale di un testo, relazione a più

documenti.

La term frequency è la frequenza di una parola in un solo documento.
Il document frequency è la frequenza di una parola in tutti i documenti presi in esame.
Una parola è molto importante in un solo specifico documento quando ha un'alta term frequency e bassa document frequency, quindi il rapporto tra queste due misure determina l'importanza della parola stessa.
L'importanza di una parola in un documento di una collezione o corpus di documenti viene misurata dalla statistica tf-idf:

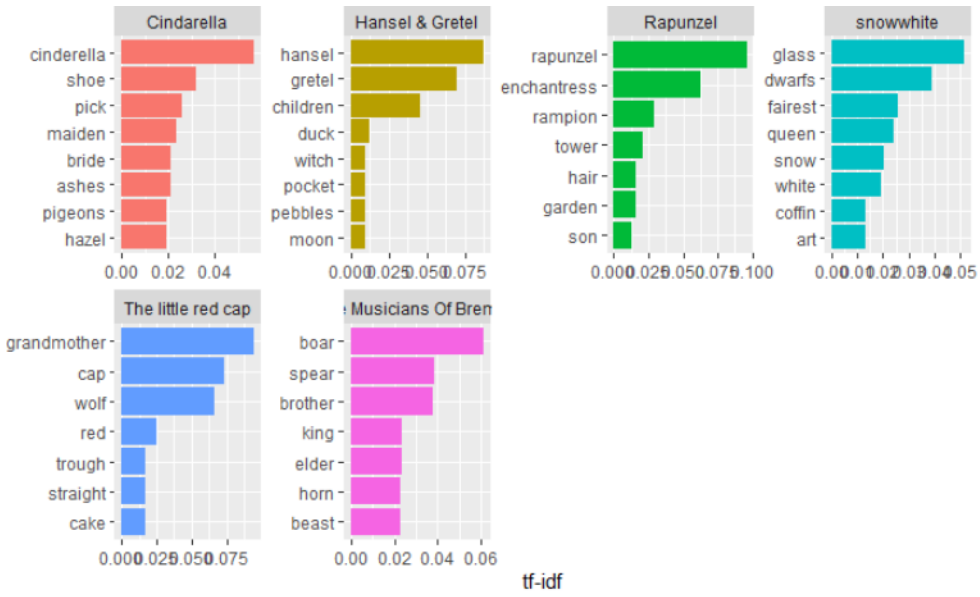


Figure 6: TF-IDF

Da questa analisi, si può notare che le prime parole più frequenti che caratterizzano i libri sono proprio i nomi dei personaggi, gli oggetti e altre personaggi caratteristici delle fiabe, come:

- La scarpetta, la cenere o la moglie in "Cenerentola";
- I bambini o la strega di "Hansel  Gretel";
- L'incantesimo, la torre e i capelli di "Raperonzolo";
- I nani, la regina o la bara in "Biancaneve";
- La nonnina, il lupo e il cappuccetto rosso di "Cappuccetto Rosso";
- Il re e le bestie de "i musicanti di Brema".

# 6 Suddivisione in argomenti

Tra le ultime interrogazioni poste al dataset delle fiabe, ho sfruttato una tecnica di machine learning di unsupervised learning per suddividere le sei storie in argomenti e analizzare le associazioni che una macchina obiettiva poteva svolgere:

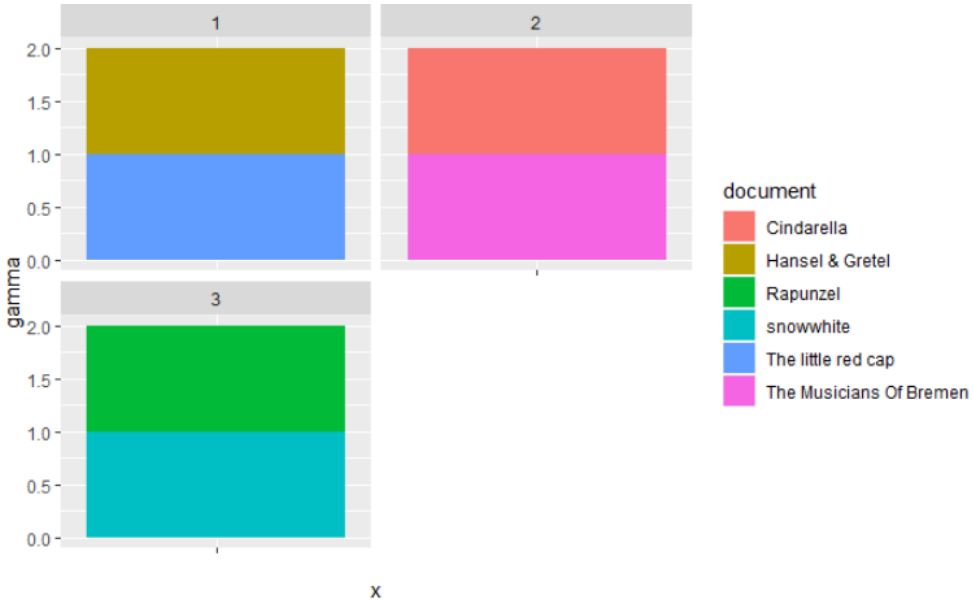


Figure 7: Suddivisione in topics

Essendo questa una tecnica di unsupervised learning, non c'è una precisa classificazione per la suddivisione in argomenti delle fiabe, potrebbero essere molteplici, anche perché non è una formula matematica.

Proviamo a interpretare la classificazione data dal nostro modello obiettivo:

- "Hansel  Gretel" e "Cappuccetto Rosso" sono state unite insieme nello stesso argomento.  Effettivamente entrambe le storie sono accumunate dalla paura e dalla sottrazione dei proprio cari per i protagonisti.
- "Cenerentola" è stato unito nello stesso argomento de "i musicanti di Brema" ed entrambe le storie parlano della tristezza dei protagonisti, tutti maltratti dal propria matrigna per Cenerentola e dai propri padroni per gli animali e tutti sono in cerca della felicità e di una miglior condizione di vita.
- Nel terzo argomento sono stati uniti "Raperonzolo" e "Biancaneve", dove entrambe le principesse sono isolate dalla propria famiglia o in una torre o

in una casa in mezzo al bosco e avranno bisogno dell'aiuto di un principe per salvarsi.

Trovo la classificazione del modello piuttosto sensata.

# 7 Quindi, le versioni originali delle storie sono più negative di quelle distribuite ai bambini?

Ho analizzato questo confronto in due parti, simili a quelle precedenti, ma prendendo in esame sia la versione originale delle storie, scritte dai fratelli Grimm, sia le versioni distribuite ai bambini di altri autori e i risultati sono i seguenti:

## 7.1 Andamento delle emozioni durante la storia

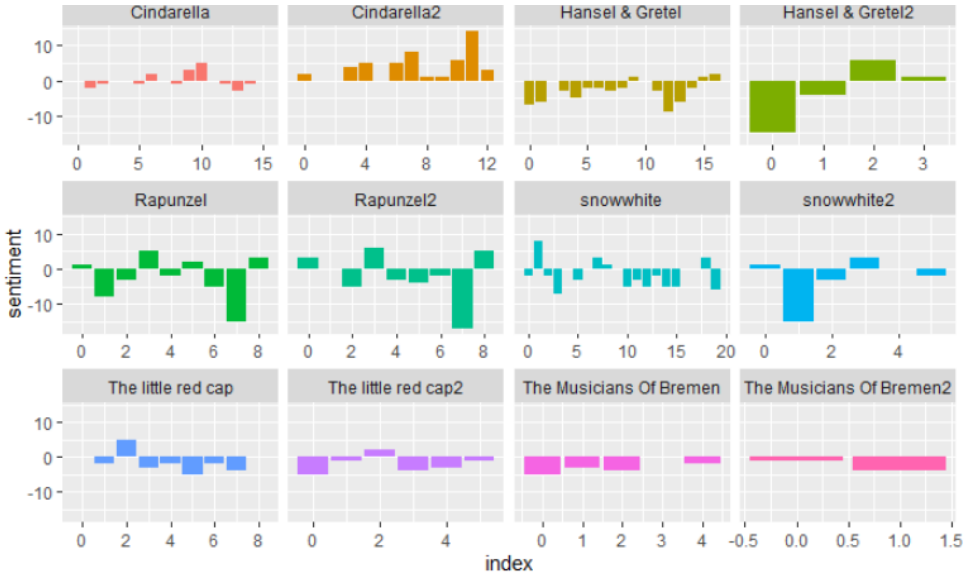


Figure 8: Andamento dei sentimenti nelle due versioni della storia

Già in una prima occhiata, questo risultato conferma la mia teoria che le versione delle fiabe distribuite ai bambini sono tipicamente più positive di quelle originali dei fratelli Grimm, soprattutto per le storie di "Cenerentola", ma mi aspettavo ci fosse maggior differenza.
Continua a stupirmi il risultato di "Raperonzolo".

### 7.2 Suddivisione in topics

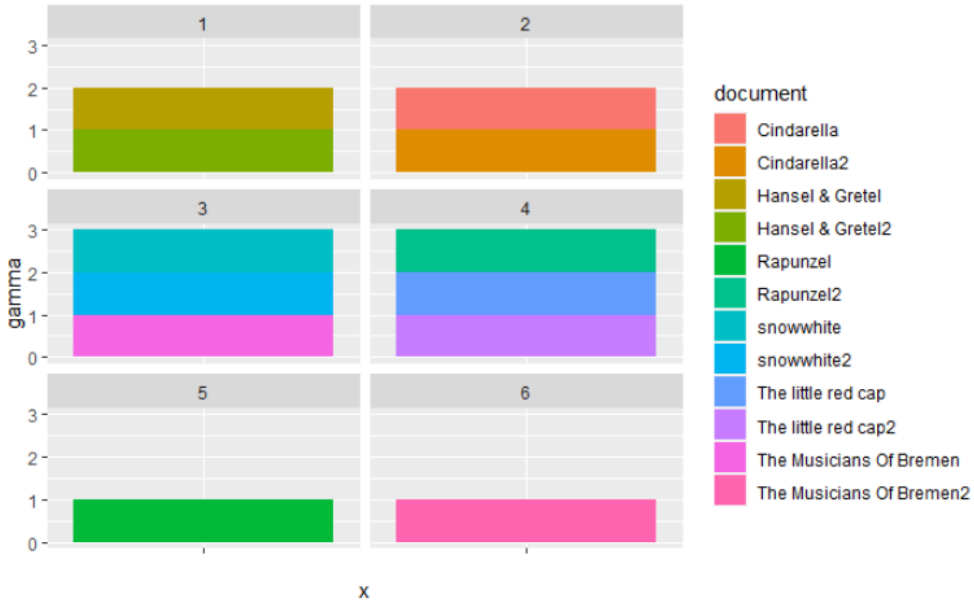


Figure 9: Suddivisione in topics

L'intento di questa operazione era quello di verificare se le due versioni delle stesse storie venissero riunite nello stesso argomento.
Questo succede per "Cenerentola", "Hansel e Gretel", "Biancaneve" e "Cappuccetto Rosso", mentre "I musicanti di Brema" e "Raperonzolo" vengono divisi. Il primo unito con "Biancaneve" (cone nella precedente unione) e il secondo con "Cappuccetto Rosso", probabilmente perché entrambe le storie parlano di animali come protagonisti principali.
Queste differenze possono intedere un possibile cambiamento di trama all'interno delle due versione, che ostacola l'associazione tra loro, sebbene l'andamento dei sentimenti sembra piuttosto simile.

## 8 Conclusioni

Questa analisi dimostra che le versioni delle fiabe distribuite ai bambini sono tipicamente più positive, rispetto alle originali, ma in misura molto minore rispetto alle dicerie che si possono trovare in internet, dove le storie vengono arricchite di particolari inventati, per far risultare la storia più macabra e negativa, il che è un risultato molto positivo, visto il pubblico a cui sono rivolte.

In ogni caso, una dose di negatività nelle fiabe è normale, anzi fondamentale per la creazione di una storia interessante e di imbattibili eroi.