Text mining for the comprehension and the analysis of the Grimm brothers' most famous tales

Fabiana Rapicavoli

July 2021

1 Introduction

In Internet we can read many theories about the cruelty and the darkness behind all the Grimm brothers' tales.
I've made an simple analysis in R to verify this thesis. In particular, I take into consideration six tales: "Cinderella", "snow-white", "the musician of Breme", "Rapunzel", "The little red cap" and "Hansel and Gretel".
So, do the Grimm's brothers' tales are so dark as Internet supports?

2 words' occurrence analysis

In a preliminary analysis, I've found the most occurrence words in all the tales:

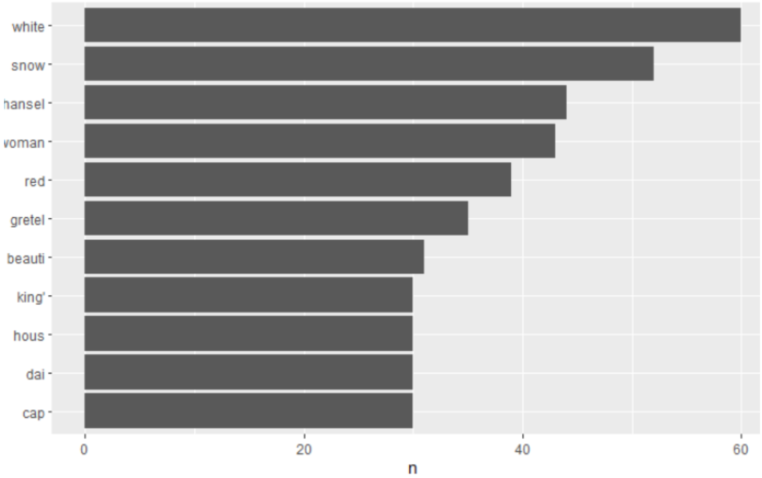

Figure 1: words' occurrence

Many words in this chart, including snow, white, Gretel, cap etc, are recognizable and you can easily guess from which tales came from.
Between these words we can notice "king" and "beauty", easily referable to the themes of fairy tales, it was predictable.

It's interesting to notice that some words (such as hous or dai) are misspelled, probably because that was the correct form at the time and they have had a later mutation.

3 sentiment's analysis

Through text mining, it's curious to analyse sentiment trend, during the story.

The programming language R provides three types of dictionary, which associate a sentiment for each word:

- **bing**, in which associate a positive or negative pole to each word.
- **AFINN**, in which associate an integer number between -5 and 5, which classify the intense of the emotion itself.
- **nrc**, in which each word is take into consideration twice, once for the positive or negative pole and the second time for the second time for the classification of the emotion (anger, fear, anticipation, trust, surprise, sadness, joy or disgust).

3.1 positive and negative words' occurrence analysis

The first classification I've made is the division of the most occurrence words between their positive or negative influence, through the bing dictionary:

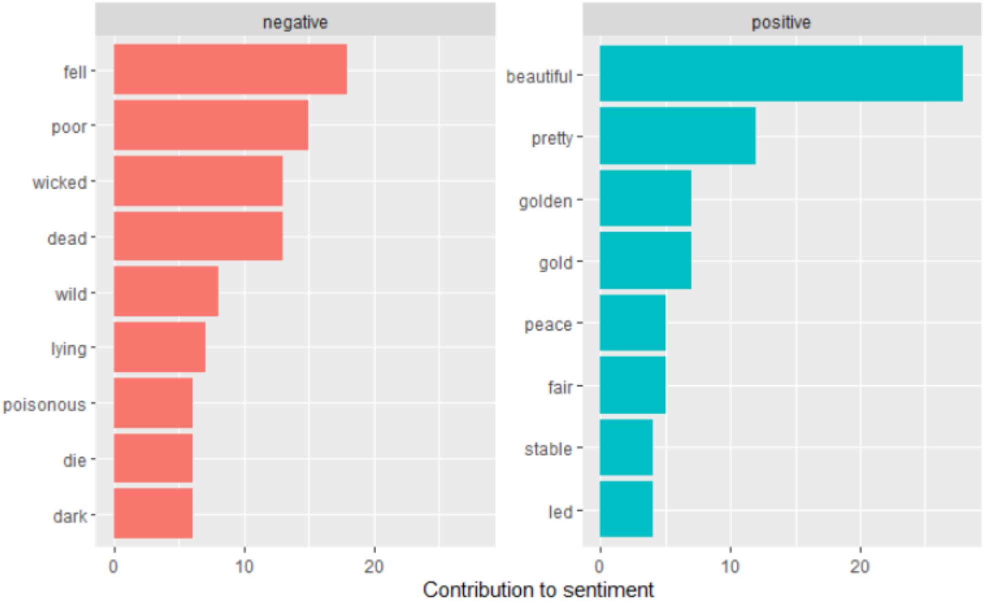

Figure 2: positive and negative words' occurrence

Positive words seems have a major occurrence, than the negative ones.

Individually, words don't suggest the provenience fairy tale, contrary to before. This means the most story's characterizing words are of neutral entity.

The most positive words relate beauty, richness and peace, while the most negative words relate poorness, death, darkness and poison.

3.2 words' occurrence for each sentiment

In this phase, I've used the nrc dictionary to detect the most frequent words for each sentiment:

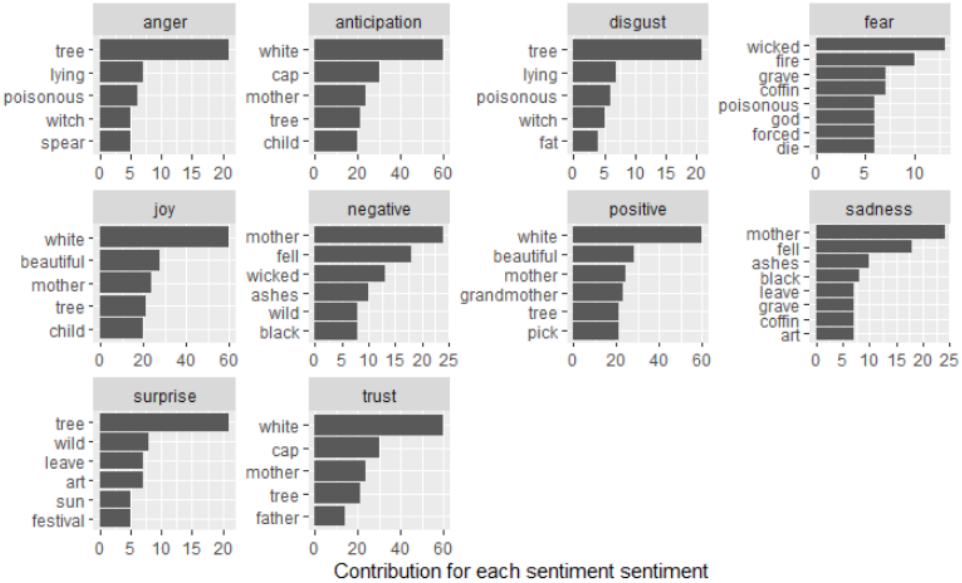

Figure 3: words' occurrence for each sentiment

From the chart, we can notice that positive words are nearly twice as many the negative words; this result is also easily attributable to fairy tales and is an extremely positive factor, given that the age range of the average audience is young.

The most frequent words are divided between the sentiments of anticipation, joy and trust - all of them are positive emotions -, but, although negative words are less frequent, they are more of the positive ones, especially in the sentiments of fear and sadness.

It's curious to notice the word "mother" is presence both in emotions like joy, trust, anticipation and, in general, in a positive entity and also in emotions like sadness and negativity.
In our case, in a fairy tale, parents often could be either friends or enemies (assuming character isn't an orphan), ma it's curious this general double association.

3.3 Sentiment trend

During sentiment analysis, one of the most interesting operation to do is verify how emotions evolve in the story, using bing dictionary:

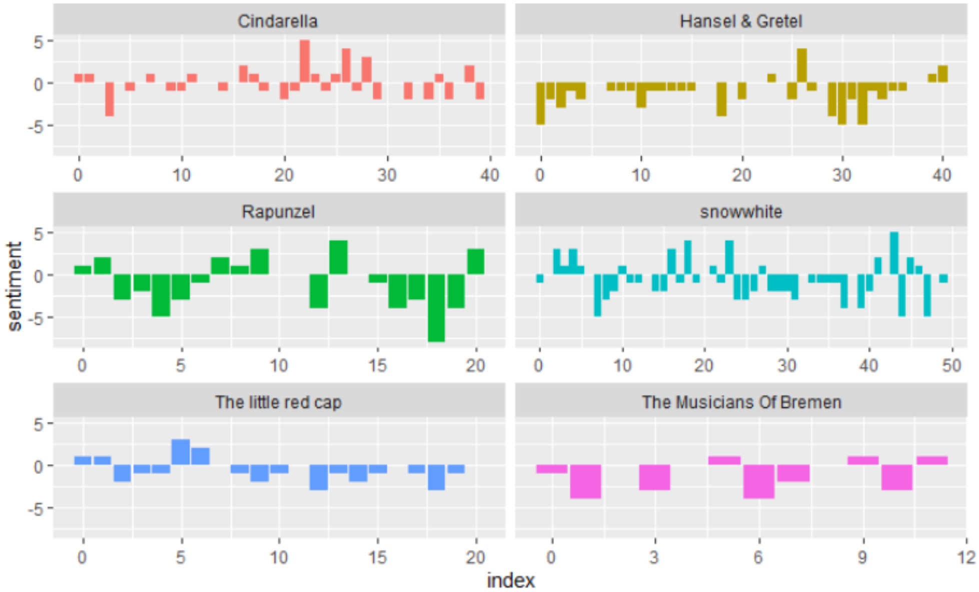

Figure 4: sentiment trend

From premises, I have expected the Grimm brothers' stories would be extremely negative. In fact, I've wanted to do this analyse to confirm this thesis. In the majority of the stories, positives bars focus on the ending and in one or few part in the middle.

"Rapunzel" is the fairy tale with the most long negative bars, so with the most concentration of negativity.
I didn't expected this result because when I was a child I've knew Rapunzel thank to the Barbie movie, with lots of magic, and thank to the Disney movie, full of adventures.

"Snow-white" is the story with the highest number of negative bars. This was a predictable result, even though I've expected "Hansel Gretel" would exceed it..

It's strange that happy endings isn't presence in all the fairy tales, but just in "Rapunzel", "Hansel Gretel" and "the musician of Breme"; this happens probably because Grimm brothers used mostly neutral words and this cause the lack of the bar.

4 Zipf law

Zipf law supports the frequency of any word is inversely proportional to its rank in the frequency table.
Thus the most frequent word will occur approximately twice as often as the

second most frequent word, three times as often as the third most frequent word, etc.
This texts confirm this theory:

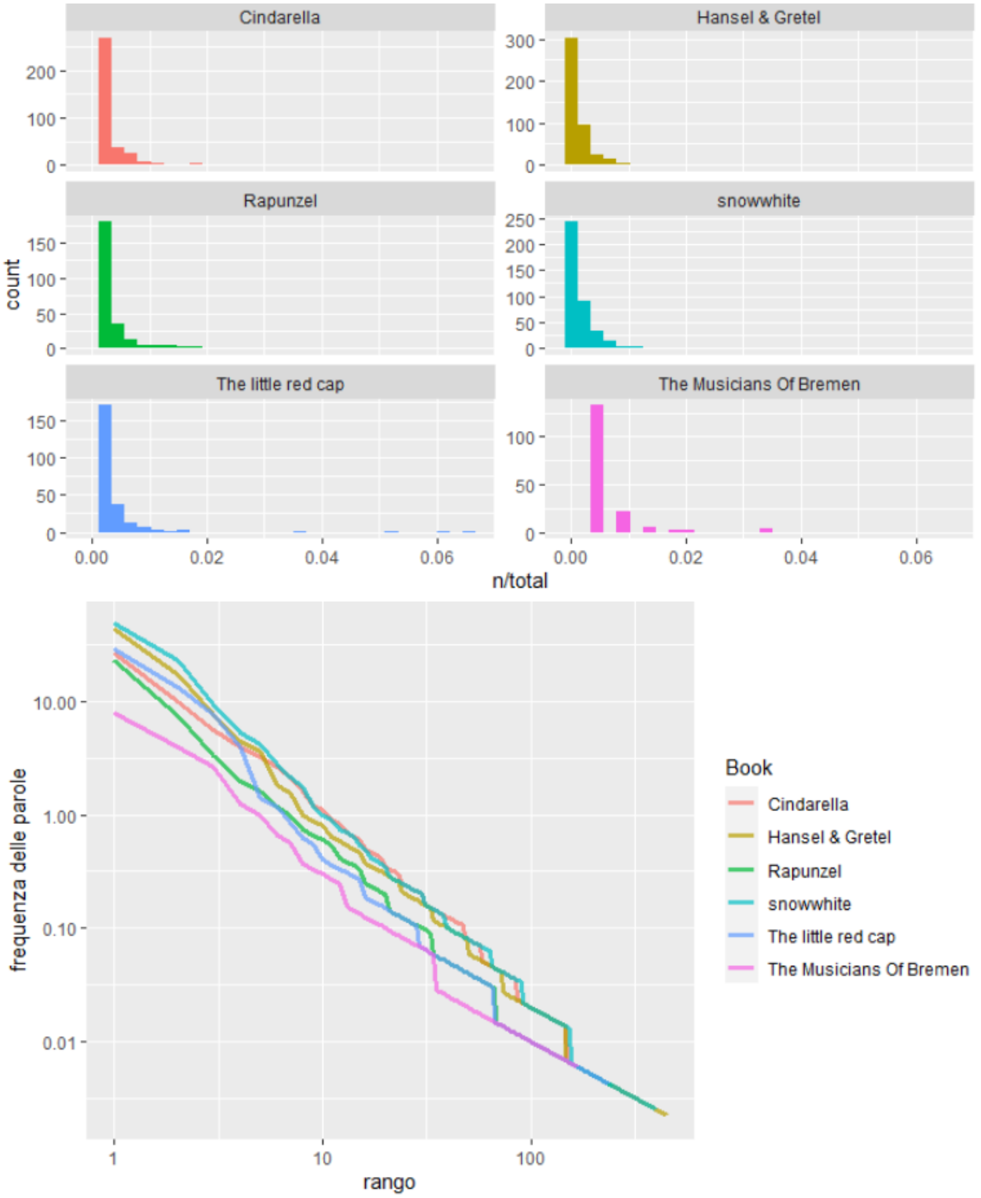

Figure 5: Zipf law

5 TF-IDF

An important operation in text mining is to associate a grade of importance to words, even for understanding the main topic of the single text.

La term frequency is the frequency of the word in a single document.
Il document frequency is the frequency of the word in all documents.
A word is very important in a specific document when it has an high term frequency and a low document frequency, so the ratio between this two integer number indicates the importance of the word itself.
The importance of a word in a document o in a collection of them is measured by tf-idf:

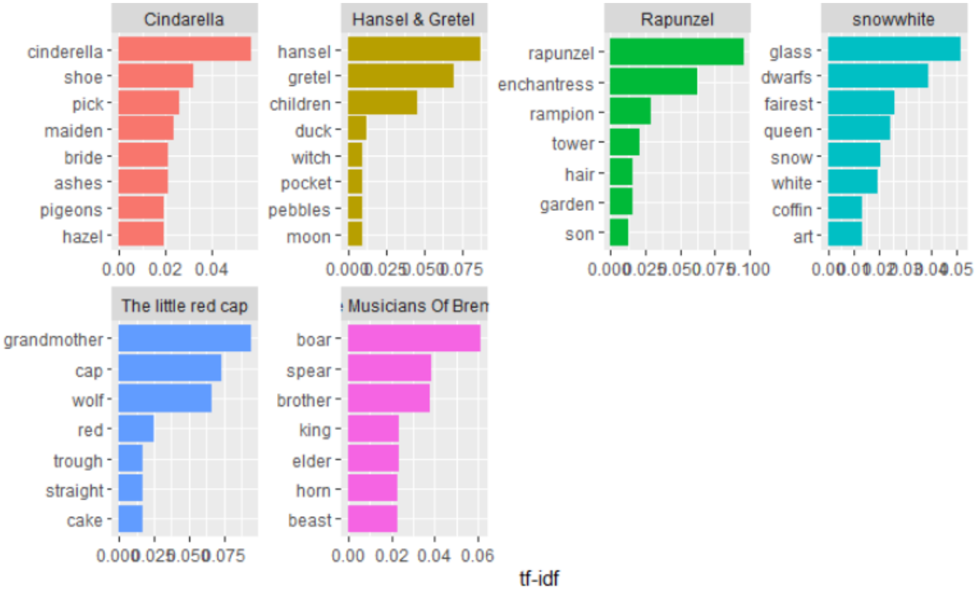

Figure 6: TF-IDF

From this analyse, we can notice that the first characterizing words of a story are just the names, objects and other character of the fairy tales, like:

- shoe, ashes or bride for ”Cinderella”;
- children or witch ”Hansel Gretel”;
- enchantress, tower or hair for ”Rapunzel”;
- dwarf, queen or coffin for ”Snow-white”;
- grand-mother, wolf and the red cap for ”The little red cap”;
- king or beast for ”the musician of Breme”.

6 Topics division

As one the last operation for the fairy tales data-set, I’ve used a machine learning technique of unsupervised learning to divide fairy tales into 3 topics for analyze the associations that an objective machine could do:

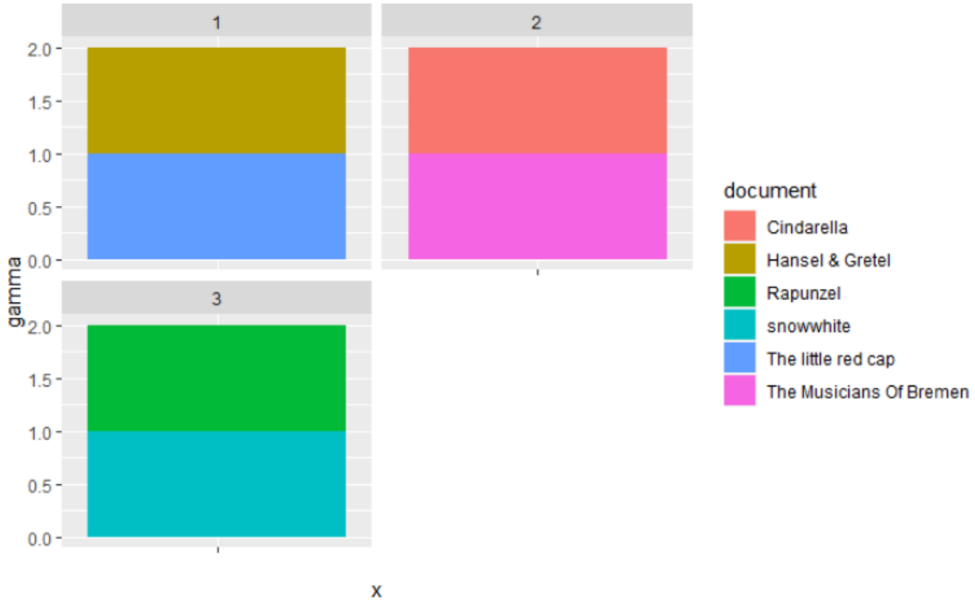

Figure 7: topics division

Since this is an unsupervised learning technique, there is no a precise classification for the subdivision of fairy tales into topics, they could be multiple, also because it is not a mathematical formula.

A possible interpretation of this classification could be:

- ”Hansel Gretel” and ”The little red cap” are united in the same topic probably because both the stories are connected by the sentiment of fear and both the main character are far from their family.
- ”Cinderella” is connected with ”the musician of Breme” probably because all the main character are sad, mistreated by their own family and all of them are pursuit happiness and a better condition of life.
- ”Rapunzel” and ”Snow-white” are connected by the fact that both the princess are isolated from their family in a tower or in a house in the wood and they need a prince, who can save them.

In my opinion, this classification is making sense.

7 So, the original fairy tales by the Grimm brothers' are more negative than the version of the stories, distributed to children?

I've analyzed this comparison in two part, similar the the previous ones, but taking into consideration both the two version of each stories.
The results are shown below:

7.1 Emotions trend

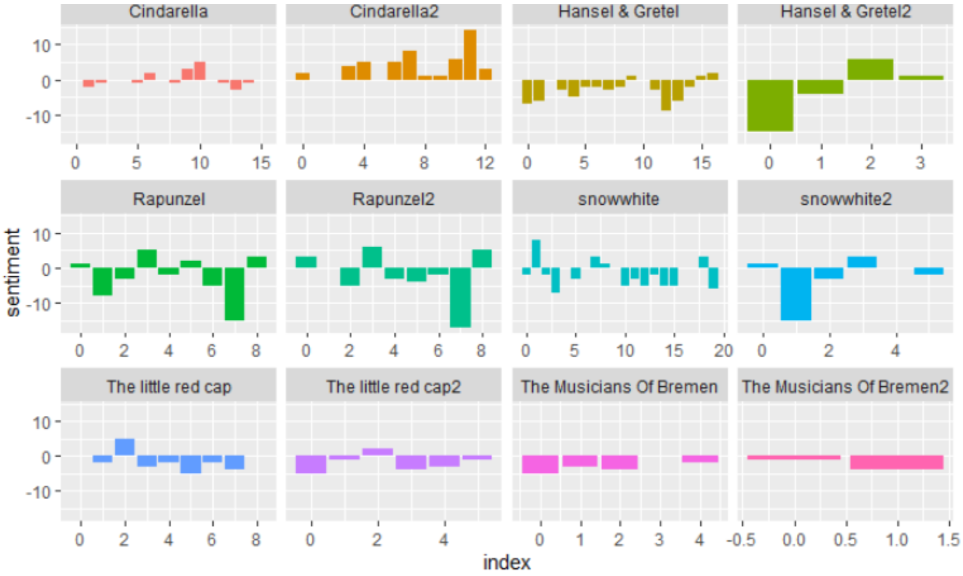

Figure 8: Emotion trend

This outcome confirm my theory that the version of fairy tales distributed to children are typically more positive than the original ones, especially for "Cinderella", but I have to admit I have expected to a worse result with more differences.
The "Rapunzel" result continue to astonish me.

7.2 Topics division

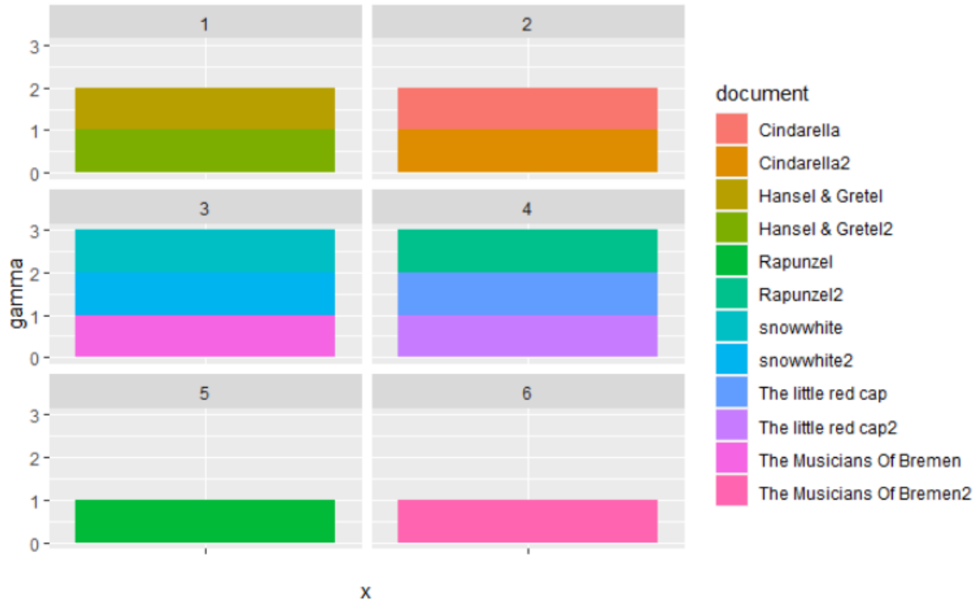

Figure 9: Topics division

With this last operation I've wanted to verify if the two version of the same story would be united in the same topic.
This happens for "Cinderella", "Hansel e Gretel", "Snow-white" and "The little red cap", while "the musician of Breme" and "Rapunzel" are divided in two different topics. the first story is united with "Snow-white" (like in the previous topic division) and the second story is united with "the little red cap", probably because both the fairy tales have animals like main character.
This differences could be interpreted like a possible change of plot between the two version, which obstacle the correct association, even though the sentiment trend is pretty the same.

8 Conclusion

This analysis shows the two version of the fairy tales are typically more positive than the original ones by the Grimm brothers', but in a minor way of the expected. This is a good result, due to the mean age of the audience.

Anyway, a dose of negativity in a fairy tale is normal, actually it is essential for the creation of an interesting story and of unbeatable heroes.