Overview of PicTropes, a film trope dataset. Preliminary technical report

Rubén H. García-Ortega

23rd June 2018

Abstract

The following report provides a descriptive analysis on the dataset PicTropes, that links 5,925 films with 18,270 tropes from the database *DBTropes.org*, which in turn collects the information from the wiki *TVTropes.org*. This preliminary study also includes and discusses graphical data distribution and top lists of both film-by-trope and trope-by-film relations, including further questions and proposing extensions of the dataset in order to explain the results.

1 Introduction

A trope can be defined as a recurring narrative device [1], whether it is technique, a motif, an archetype or a *cliché*, used by the authors to achieve specific effects that might vary from increasing the interest, surprising, recall familiarity, entertaining, etc, in their creative works, like books, films, comics or videogames. Some tropes are broadly adopted, academically studied and promoted, as the *Three-act Structure* formulated by Syd Field [2], the *Hero's Journey* studied by Vogler [3], the *McGuffin* popularized by Hitchcock [4] and the *Chekhov's Gun* developed by the russian writer with the eponymous name [5], but there are thousand of not-so-widely used tropes as well, discovered and catalogued everyday by professionals and enthusiastic of the storytelling; their study is organic, dynamic and extensive, for this reason our reference is a live wiki called *TVTropes.org* [6], that is being collecting thousand of descriptions and examples of tropes from 2014 until now. The semantic network of knowledge behind *TVTropes.org* is huge and complex; it massively links hierarchies of tropes to their usage in creations for digital entertainment. However, the data is only available through its web interface, that's why, in order to make it usable by the scientific community, Kiesel [7] extracted all their data to a database so-called *DBTropes.org*.

DBTropes.org is released as a NTriples formatted RDF file that can be downloaded directly from their official site [7]. The last release available was built in July 2016 and contains 2,1057,602 RDF statements, a large amount of data that makes it hard to load in a RDF visualization tool. For this reason, we extracted part of its information to a new dataset in JSON format, readable by most of the programming languages in a friendly manner, that is called PicTropes and contains just the films and the name of the tropes they use. This dataset can be used to build a recommendation system.

The goal os this report is to extract valuable statistical data from the dataset PicTropes that shall be used in further researches and experiments related to machine learning and narrative generation. In particular, the results will be directly applicable to different researches in the context of the PhD *Bio-inspired techniques for procedural generation of backstories in literature and open world videogames*.

Following the principles of the Open Science, the dataset and the source files of the report are released under the *Attribution-ShareAlike 3.0 Unported* License (CC BY-SA 3.0) in a public repository [8]. The values and graphs included in the current report are dynamically calculated using pweave, scipy, numpy and matplotlib in order to allow the reproducibility.

2 Pre-processing the data

In order to ease the data analysis we will use two data structures: a dictionary of films where the values are lists of the tropes they use and the reversed dictionary, where the keys are the tropes and the values are lists of films that they are used in. The class that handles the python code can be found in the repository [8].

3 Descriptive analysis of the number of tropes per film

The dataset contains 5,925 films with the tropes that have been found in each of them. The number of tropes in a film goes up to 515 (*GuardiansOfTheGalaxy*); however, the great majority of films have just a few dozens of tropes (43.434 on average). As shown in Figure 1 the data fits a log-logistic distribution (*location*=1.945, *shape*=0.054, *scale*=29.292). The features of the descriptive analysis can be found in Table 1 whereas the top-25 ranking of films by the number of tropes is provided in Table 2 for further analysis.

Number of films	5,925
	Tropes per film
Minimum	1
Maximum	515
Mean	43.434
Median	29.0
Q1	16.0
Q2	29.0
Q3	52.0
Variance	2,133.35
Skewness	3.332
Kurtosis	17.373

Table 1: Descriptive analysis of the number of tropes per film

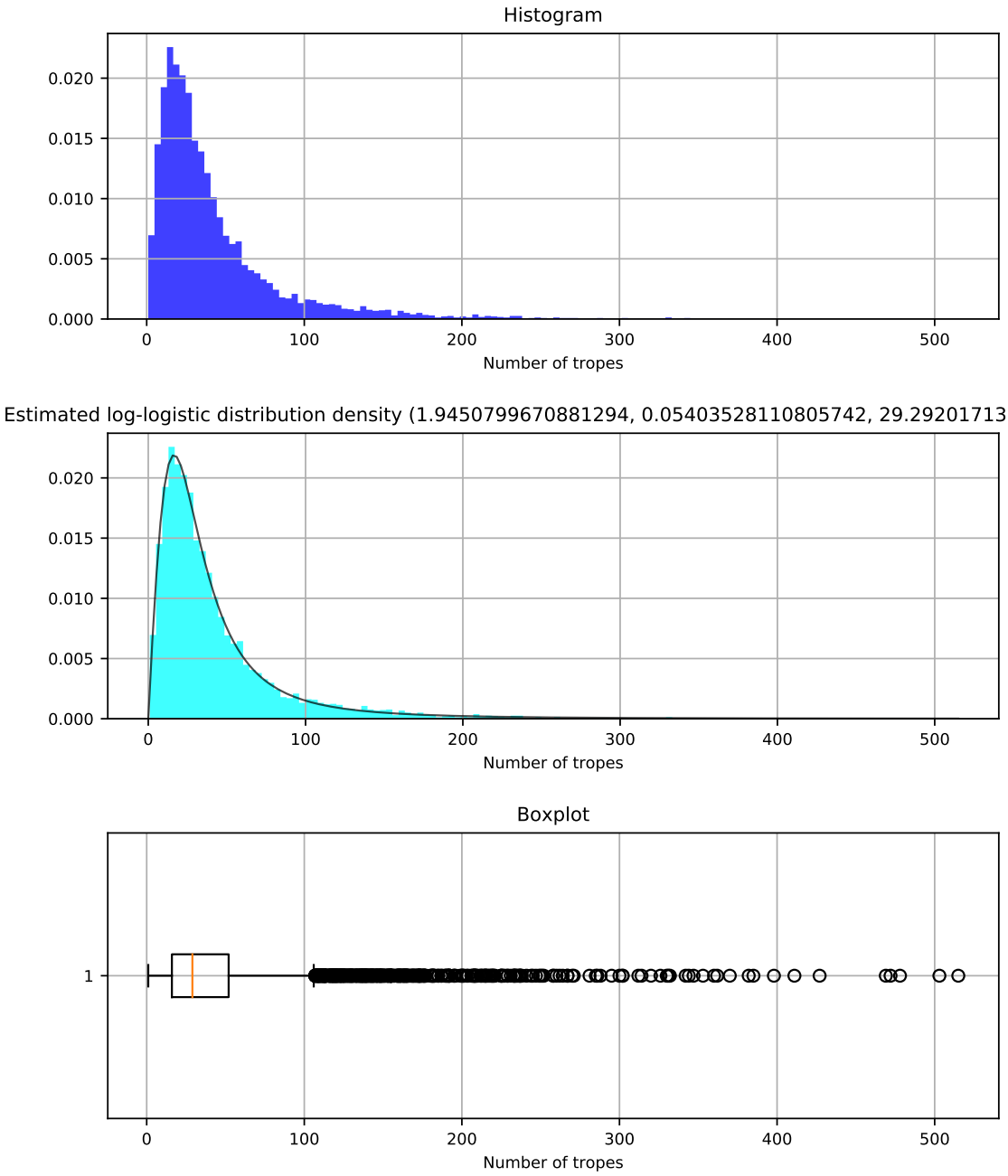

Figure 1: Distribution of the number of tropes per film

Position	Film (short name)	N. tropes
1	GuardiansOfTheGalaxy	515
2	TheDarkKnightRises	503
3	XMenDaysOfFuturePast	478
4	CaptainAmericaTheFirstAvenger	472
5	XMenFirstClass	469
6	Thor	427
7	SherlockHolmes	411
8	TheLordOfTheRings	398
9	PacificRim	385
10	CaptainAmericaTheWinterSoldier	382
11	WhoFramedRogerRabbit	370
12	TheDarkKnight	362
13	TronLegacy	360
14	StarTrek	353
15	StarTrekIntoDarkness	347
16	Skyfall	344
17	TheGodfather	342
18	JurassicWorld	332
19	Serenity	331
20	BackToTheFuture	330
21	Inception	326
22	IronMan3	320
23	AustinPowers	314
24	GalaxyQuest	312
25	ManOfSteel	302

Table 2: Top-25 ranking of films by the number of tropes

4 Descriptive analysis of the number of films per trope

The dataset contains 18,270 tropes with the films where they have been found. The number of films where a specific trope appears goes up to 1502 (*ShoutOut*); however, the great majority of tropes are only found in a few films (14.086 on average). As shown in Figure 2 the data fits a folded Cauchy distribution (*location*=0.13, *shape*=1.0, *scale*=3.735). The features of the descriptive analysis can be found in Table 3 whereas the top-25 ranking of tropes by the number of film they appear in is provided in Table 4 for further analysis.

Number of tropes	18,270
	Films per trope
Minimum	1
Maximum	1,502
Mean	14.086
Median	5.0
Q1	2.0
Q2	5.0
Q3	12.0
Variance	1,464.794
Skewness	11.758
Kurtosis	245.951

Table 3: Descriptive analysis of the number of films per trope

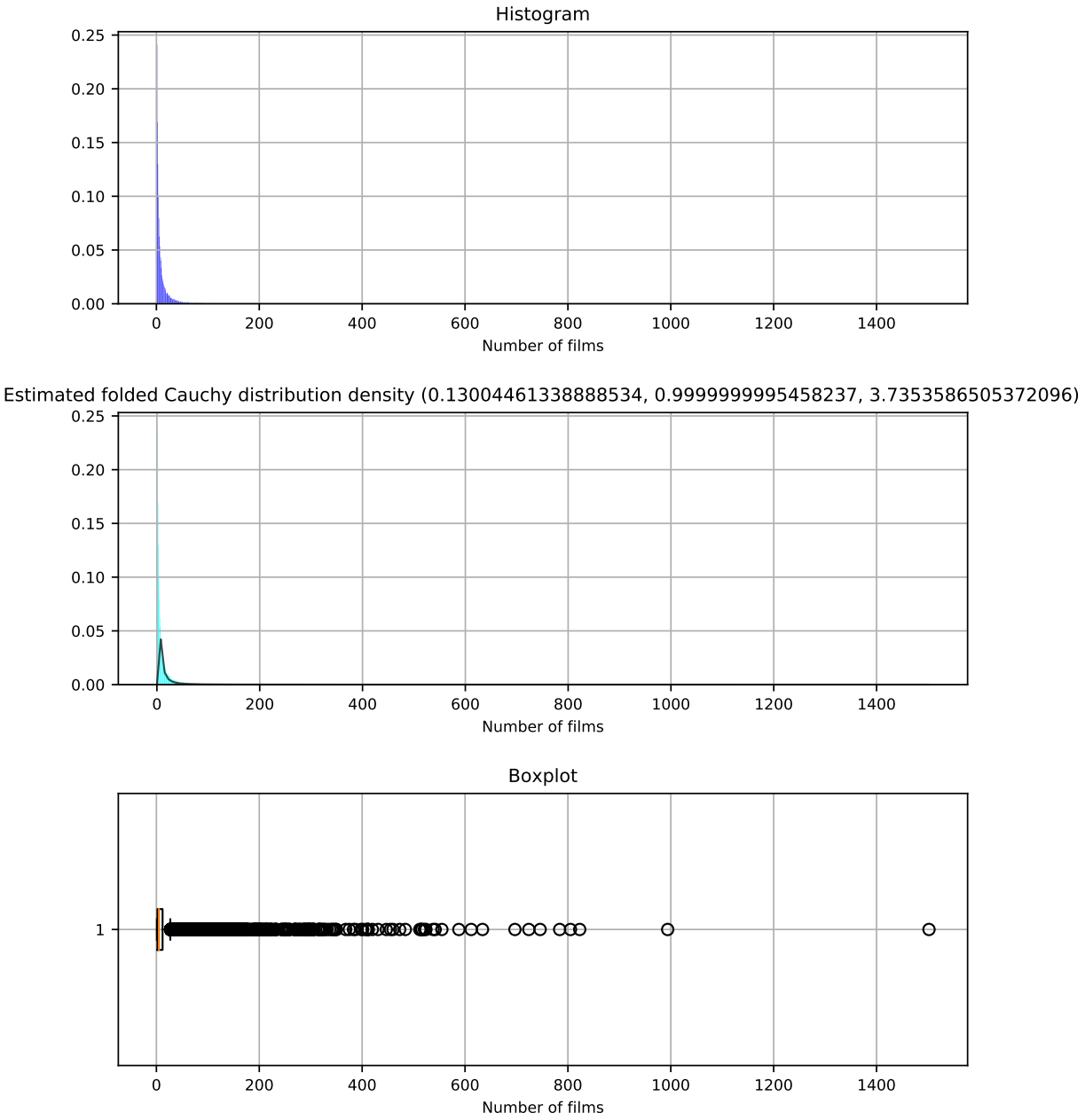

Figure 2: Distribution of the number of films per trope

Position	Trope (short name)	N. films
1	ShoutOut	1502
2	ChekhovsGun	994
3	OhCrap	823
4	DeadpanSnarker	805
5	Jerkass	784
6	Foreshadowing	746
7	LargeHam	724
8	BittersweetEnding	697
9	TitleDrop	634
10	BigBad	612
11	MeaningfulName	588
12	BerserkButton	555
13	TheCameo	542
14	WhatHappenedToTheMouse	538
15	RunningGag	524
16	TooDumbToLive	521
17	DownerEnding	516
18	FanService	516
19	KarmaHoudini	514
20	GroinAttack	512
21	BrickJoke	484
22	BookEnds	473
23	MoodWhiplash	460
24	KickTheDog	455
25	PrecisionFStrike	447

Table 4: Top-25 ranking of tropes by the number of films

5 Discussion

The descriptive analysis of the data shows that, on average, a film contains 43.434 tropes and that a trope is contained in 14.086 films. However, the distribution in both cases, is long tailed:

In the case of the films, some of them include hundred of tropes, but the majority include just some dozens. That's clearly visible in the top-25 ranking, where the difference between the top (*GuardiansOfTheGalaxy*) and the bottom (*ManOfSteel*) is 41.359 %. The reason why this difference is so big is not clear: It could be because they actually use more tropes or because the bigger the fandom is the more the analysis is performed and the more the tropes are found. It is specially noticeable that 22 out of the 25 belong to the adventure genre, that 10 out of the 25 are about super-heroes of the comics, that most of the films belong to this millennium and are super-productions. In order to fully analyse the results, we would need extra meta information from external sources, for example, the average votes by the community, the release date, the genres or the popularity.

When we analyze the tropes, the effect of the long tail is even more noticeable and the difference between the top (*ShoutOut*) and the bottom (*PrecisionFStrike*) in the top-25 ranking is 70.24 %. The trope *Shout out*, deliberate allusions or references to other sources of inspiration, is by far the most used and it appears in the 25.35 % of the analyzed films. Again, the explanation of the distribution graph would require extra information, and, perhaps, a good ontology of tropes by different features.

Finally, it is important to remark that the last dump of *DBTropes.org* was extracted in 2016 and that's the reason why we cannot find newer films. The information in *TVTropes.org* could've been modified in many ways since then so we could get more accurate information if a new version of *DBTropes.org* were available or if we build a way to extract the information.

6 Conclusions

The dataset PicTropes allowed shows...

References

[1] C. Baldick, *The Oxford dictionary of literary terms*. OUP Oxford, 2015.

[2] S. Field, *Screenplay*. Delacorte New York, 1982.

[3] C. Vogler, *The Writer's journey*. Michael Wiese Productions Studio City, CA, 2007.

[4] F. Truffaut, A. Hitchcock, and H. G. Scott, *Hitchcock*. Simon and Schuster, 1985.

[5] P. M. bitsilli, *Chekhov's art, a stylistic analysis*. Ardis, 1983.

[6] "Tv tropes." (visited on 2018-07-11).

[7] M. Kiesel, "Dbtropes – skipforward." (visited on 2018-07-11).

[8] R. H. Garcia Ortega, "tropes open data." (visited on 2018-07-11).