Bhuvana Ramaswamy

Data Analysis II
Prof. Tharsen
Winter 2023

A Strag's Guide to Adaptations

Mar 10, 2023

Project Overview

This project seeks to explore what features predict the film adaptability of literary works through top-level and more granular approaches. For the top-level approach, an ANN was built to predict movie ratings based on a dataset that combined film/literary work reception information available Goodreads and IMDB. The complementary subproject focused on adaptations of the Douglas Adams's *Hitchhiker's Guide to the Galaxy*. Through this analysis, we may also attempt to identify the factors that could predict the adaptability of a literary work to emerging forms of art such as virtual reality environments and interactive films.

Data Sources

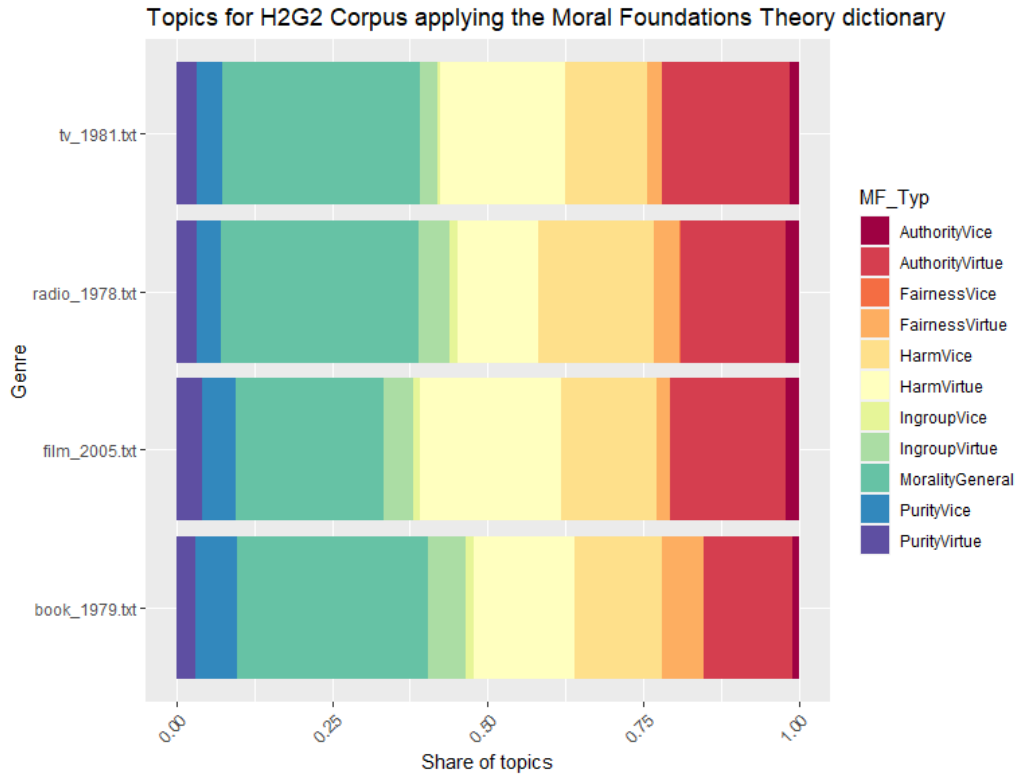
The data sources for the ANN project consisted of a dataframe created by stringjoining Goodreads and IMBD data from Kaggle. For the Hitchhiker's Guide to the Galaxy, the radio script, television screenplay, book, and film screenplay were used.

Methods

- Artificial Neural Network
- Topic-specific dictionaries
    - Moral Foundations Theory
    - Newsmap Dictionary
- Topic Modelling
    - Latent Dirichlet Allocation (LDA)
    - Structural Topic Model (STM)
- Textual Statistics
    - Zipf's Law
    - Parts-of-Speech comparatives
    - Flesch-Kincaid
    - Type-Token Ratio
    - Hapax Richness
    - Tf-idf
    - Euclidean Distance

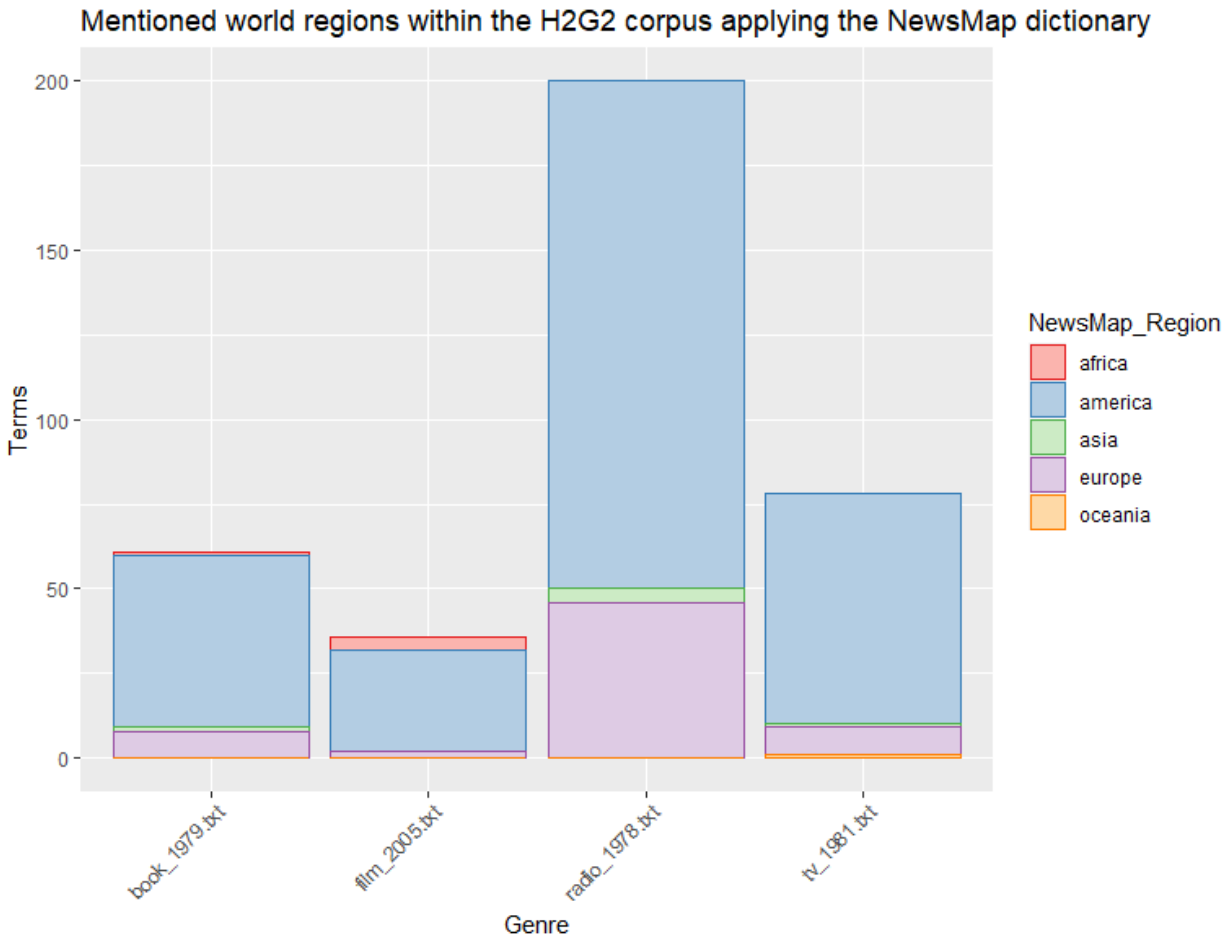
Topic Dictionaries - Moral Foundations Theory

Moral Foundations Theory was postulated to interrogate reasons for cross-cultural variations in the manifestation of human moral reasoning, as well as explain recurrent moral themes that cut across cultures. Although I'm not inclined to comment on the original applications of the theory, I think it's entirely possible that similar "cross-medium" manifestations of cultural values occur in the course of adaptations, especially of literary work. The creation of literary work is almost always a solitary exercise, as opposed to adaptations into other genres such as film, radio and television, all of which involve either teamwork with other co-creators or significant participation from stakeholders external to the primary creator. In a way, the MFT pits "indivisualistic" societies against "collectivistic" ones, and the purpose of this enquiry was to verify if there were parallels to this dichotomy in an individualistic work environment and one that requires teamwork and cooperation, and whether this could be observed in the creative outputs produced under these work-culture, rather than sociocultural constraints.



In the case of the Hitchhiker's Guide to the Galaxy (hereafter referred to as H2G2), it's interesting to note that representations of ingroup virtue and ingroup vice seem lowest in the genres that require the most teamwork, film and television. While a radio play would normally fall into the same category, we know that this particular show, was written mostly without other contributors and so falls into the same category as the book. \

Topic Dictionaries - Newsmap



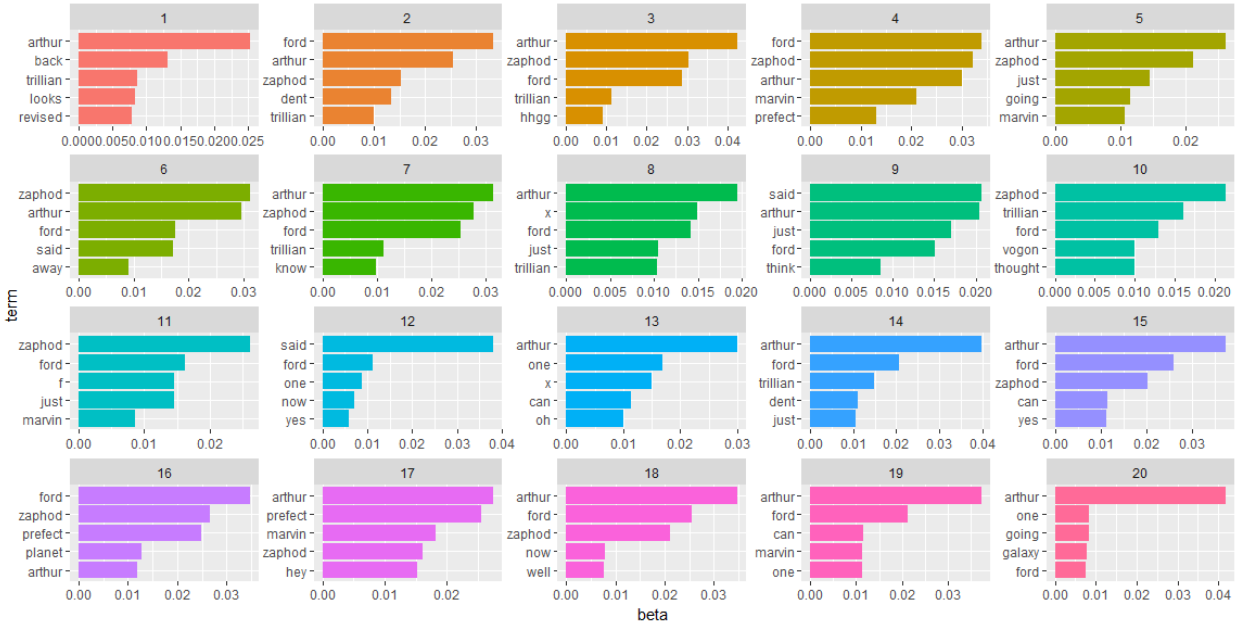
The NewsMap dictionary uses a semi-supervised model to detect and classify geographical locations in texts. It seemed pertinent to amps these locations because films, and to a lesser extent television shows often target a more universal audience than books or radio shows, which are invariably optimized to a target audience rooted in the creator's own economic region or location. In this case, the H2G2 adaptations conform to the rule *and* disprove it, with the radio

show having more mentions of the Americas but it also contains far more references to Europe as expected. It's also interesting to note that the film and television show contains the least number of references to the work's place of origin, Europe. This is to be expected, given that television and film especially cater heavily to the North American copyrights region. There are obvious drawbacks to using such a classification in the context of a sophisticated work of speculative fiction, where the speculative elements mirror Earth cultures in ways too complicated to be captured computationally at this point in time.

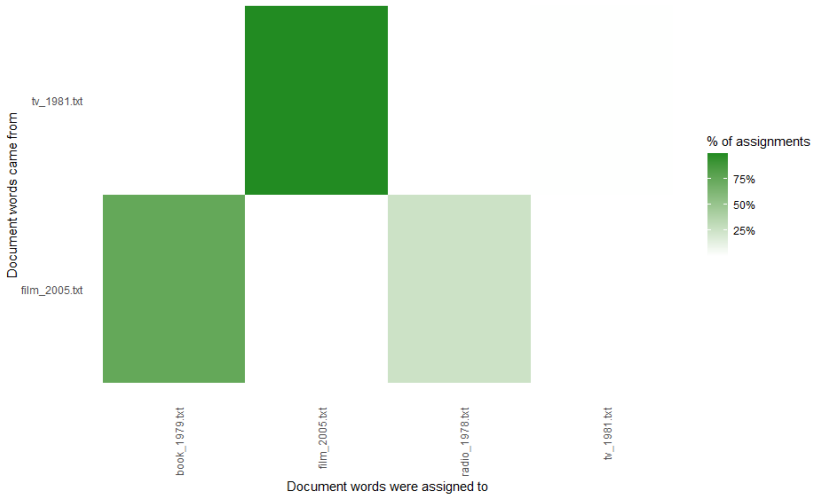
Topic Modelling - LDA

LDA is a topic modelling technique that treats each document as a mixture of topics and each topic as a mixture of words. Since the story of H2G2 remains largely constant across its various adaptations, I was curious to see how closely the various topics from the adaptations would resemble each other. From the results I obtained, my instincts prove right. There isn't a great deal of topical distance(?) between the various genres.

```
> get_terms(LDA_fit_20, 5)
     Topic 1    Topic 2    Topic 3    Topic 4   Topic 5  Topic 6  Topic 7
[1,] "arthur"   "ford"     "arthur"   "ford"    "arthur" "zaphod" "arthur"
[2,] "back"     "arthur"   "zaphod"   "zaphod"  "zaphod" "arthur" "zaphod"
[3,] "trillian" "zaphod"   "ford"     "arthur"  "just"   "ford"   "ford"
[4,] "looks"    "dent"     "trillian" "marvin"  "going"  "said"   "trillian"
[5,] "revised"  "trillian" "hhgg"     "prefect" "marvin" "away"   "know"
     Topic 8    Topic 9  Topic 10   Topic 11 Topic 12 Topic 13 Topic 14   Topic 15
[1,] "arthur"   "said"   "zaphod"   "zaphod" "said"   "arthur" "arthur"   "arthur"
[2,] "x"        "arthur" "trillian" "ford"   "ford"   "one"    "ford"     "ford"
[3,] "ford"     "just"   "ford"     "f"      "one"    "x"      "trillian" "zaphod"
[4,] "just"     "ford"   "vogon"    "just"   "now"    "can"    "dent"     "can"
[5,] "trillian" "think"  "thought"  "marvin" "yes"    "oh"     "just"     "yes"
     Topic 16  Topic 17  Topic 18 Topic 19 Topic 20
[1,] "ford"    "arthur"  "arthur" "arthur" "arthur"
[2,] "zaphod"  "prefect" "ford"   "ford"   "one"
[3,] "prefect" "marvin"  "zaphod" "can"    "going"
[4,] "planet"  "zaphod"  "now"    "marvin" "galaxy"
[5,] "arthur"  "hey"     "well"   "one"    "ford"
```



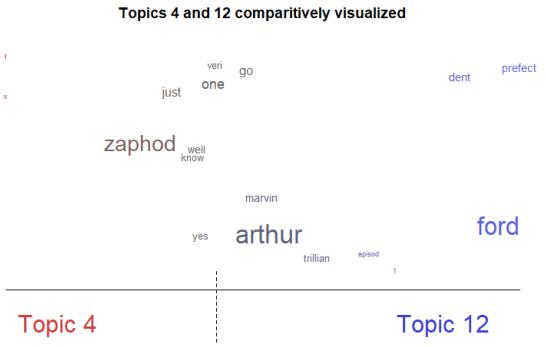
However, interestingly enough, upon producing a confusion matrix which visualizes how often topics from one genre were assigned to another, I realized that the film and television had far more influence in the assignment process than the book or the radio show. This may indicate that those topics that seem pertinent to the story, ie, the characters, appear with more significance in the screen adaptations than the others. I made the choice to retain dialogue tags in all the texts simply because removing them would also remove the whole of characterization

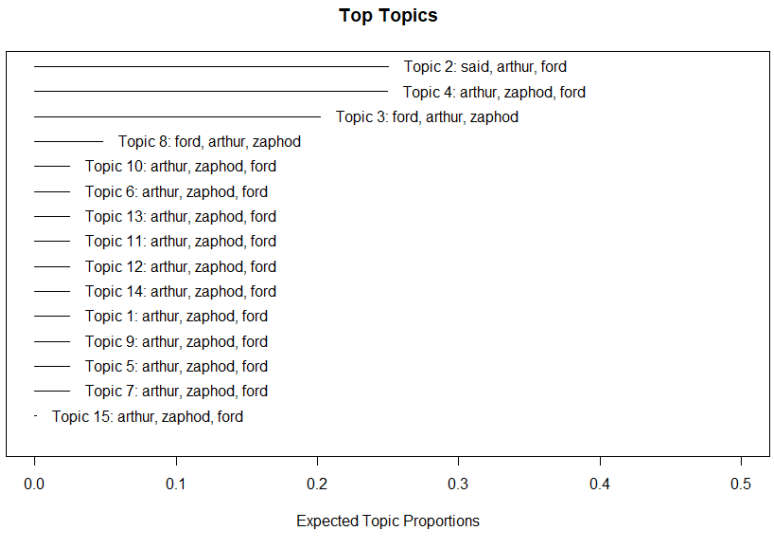


from the document.

Topic Modelling - STM

STM is another popular topic modelling approach that yielded useful results in this analysis. The main advantage of STM over LDA is that by allowing us to incorporate metadata into our model, it eliminates the reliance on hyperparameters that LDA requires.



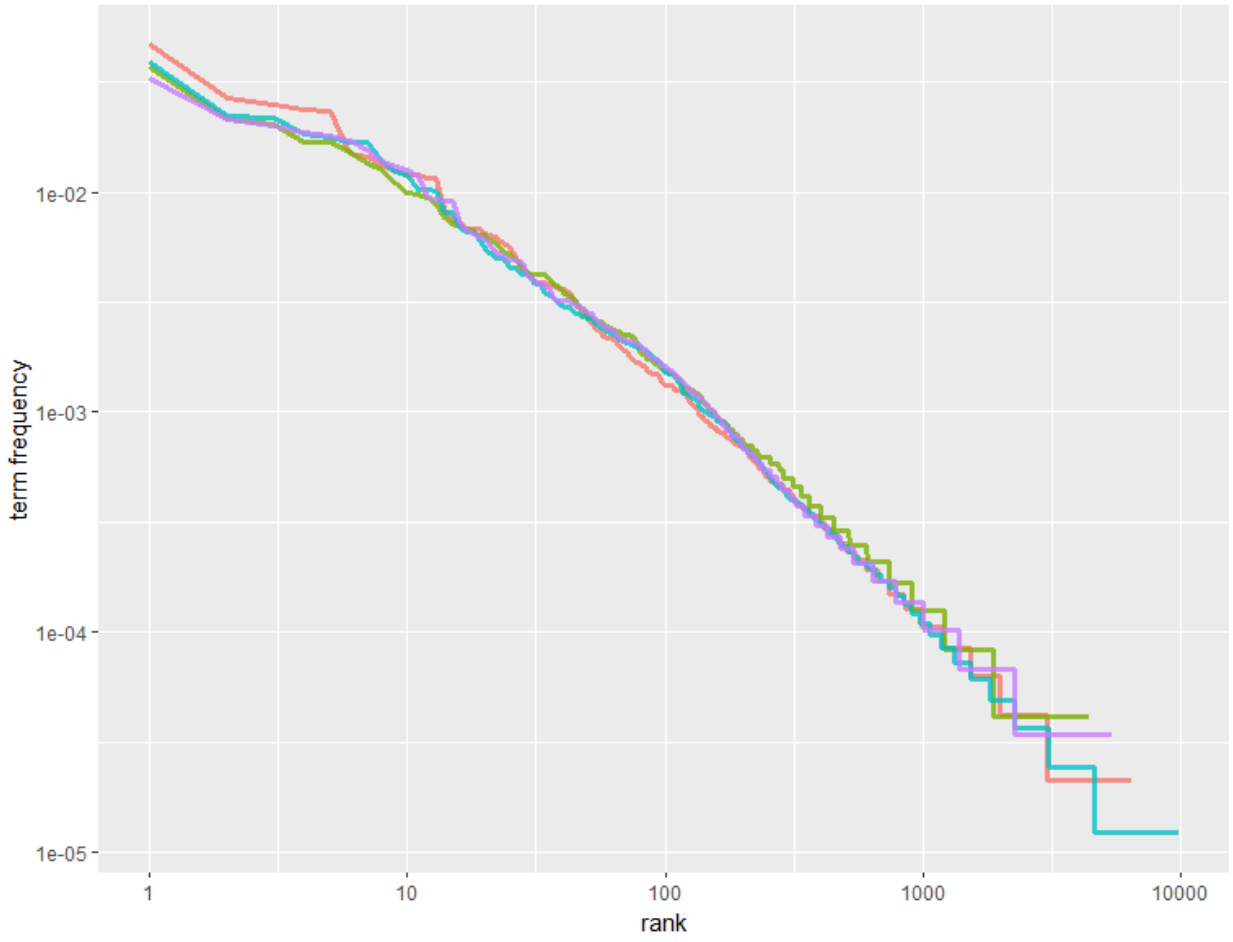


The topic modeling again reveals almost no variation, even less than the LDA. This makes sense given that the character names were retained in the texts, but the lack of variation in the representation of these characters across the topics also points to a certain constancy in character representations across adaptations in H2G2.

Textual Statistics

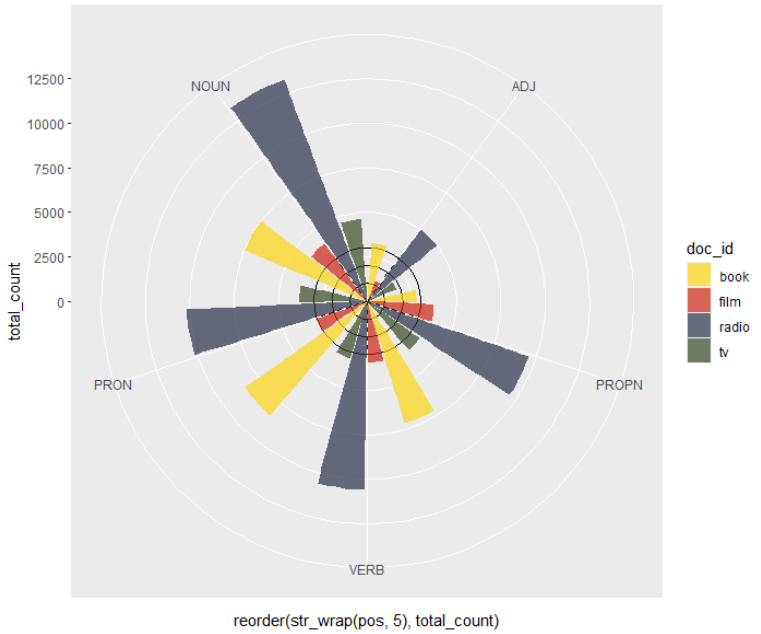
Zipf's Law

Zipf's Law states that for a large enough piece of text, the frequency of use n of any word decreases with its rareness r in the text in an approximately hyperbolic way. Comparing the adaptations against each other and against normal English usage, allows us infer lexical variation and the degree of aberrations in the adaptations. It was expected that the film and television adaptations will have smoother curves than the book or the radio show. However, there seems to be more or less similar degrees of variation in the tail ends of all the genres.



Parts of Speech Comparatives

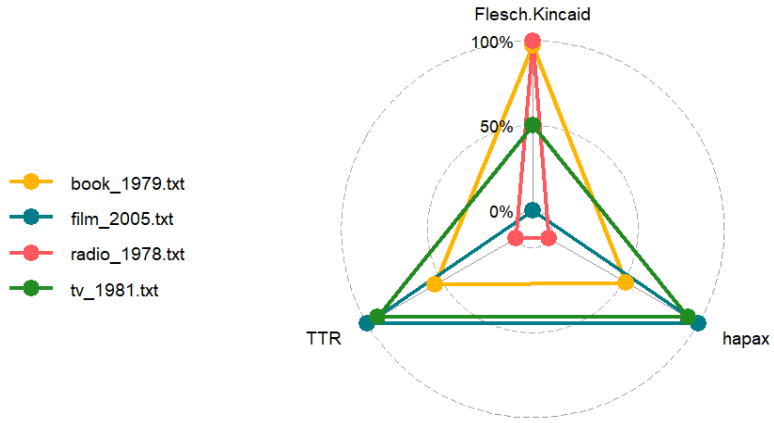
```
> posdf4plot
   doc_id   pos total_count
1    book   ADJ        3302
2    book  NOUN        7450
3    book  PRON        6149
4    book PROPN        2782
5    book  VERB        7179
6    film   ADJ        1225
7    film  NOUN        4049
8    film  PRON        3060
9    film PROPN        3764
10   film  VERB        3447
11  radio   ADJ        5025
12  radio  NOUN       13312
13  radio  PRON       10173
14  radio PROPN        9573
15  radio  VERB       10575
16     tv   ADJ        1769
17     tv  NOUN        4635
18     tv  PRON        3863
19     tv PROPN        3600
20     tv  VERB        3345
>
```



Unsurprisingly, verbs appear far less significant to the TV and film adaptations, where action is depicted visually rather than through character dialogues. Since adaptations to audiovisual mediums require significant action, it seems reasonable to look for literary works with a relatively high noun-verb ratio. However, an attempt to distinguish between thinking and doing verbs may be necessary to optimize results.

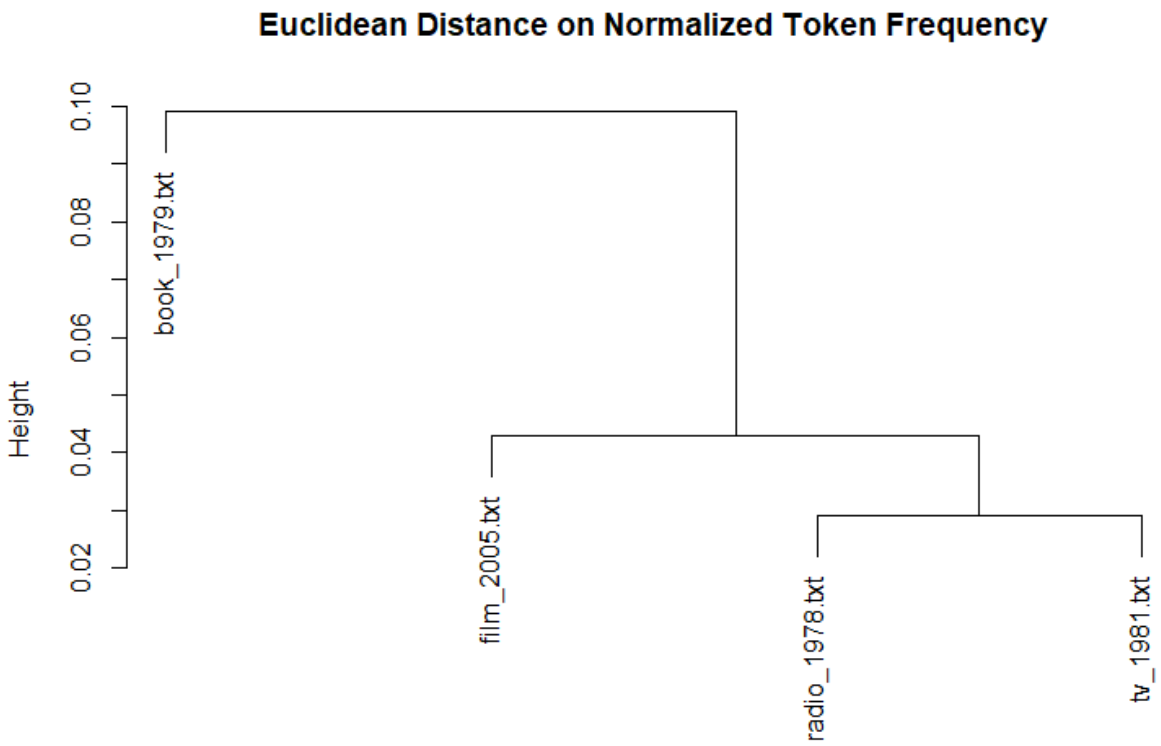
Textual Statistics - Flesch-Kincaid, TTR, and Hapax Richness

```
        document Flesch.Kincaid      hapax        TTR
1  book_1979.txt       5.913495 0.06628125 0.13117704
2  film_2005.txt       4.122678 0.08313549 0.15777608
3 radio_1978.txt       5.963394 0.04822247 0.09973406
4    tv_1981.txt       5.042302 0.08063957 0.15361561
```



Flesch-Kincaid is a common readability score assigned to indicate how difficult a text is to read and is used widely in publishing to assess grade levels for children's and young adult fiction. TTR is the ratio obtained by dividing the types (the total number of different words) occurring in a text or utterance by its tokens (the total number of words) and is an indicator of lexical variation. Hapax richness is another measure of lexical variation and is calculated by Hapax dividing the number of words that occur only once by the number of total words. The plot above is a normalized representation of all three measures. Logically, the book should have the biggest triangle, ie, the highest lexical variation (represented by TTR and hapax richness) and the highest Flesch-Kincaid score. However, this does not appear to be the case. The biggest triangle surprisingly belongs to the TV show. This could also be due to the emphasis on conveying world-building, a major pull due to humor, in the adaptations.

Euclidean Distance

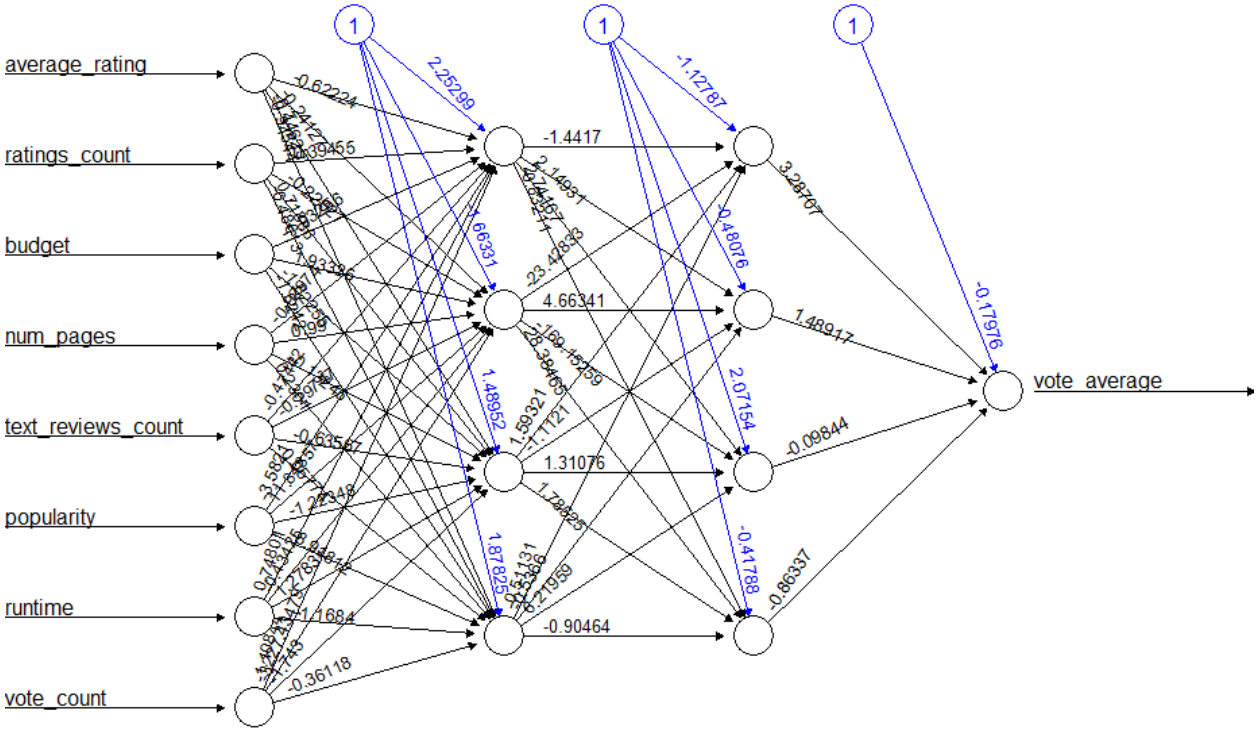


It's not surprising that the radio show and the television are more closely related than the film and the book is more removed from the others. The radio and the television show cover a similar period of narrative time, in addition to representing similar kinds of narrative distance with the audience, as opposed to the film or the book.

ANN

The objective of the ANN was to predict movie ratings based on reader reception of a literary work and the popularity metrics of the film itself. The two datasets were merged using stringjoining methods and revisions were made to the default stepmax and threshold parameters to facilitate convergence. The ANN achieved a surprising 78% accuracy rate,

probably due to the addition of the popularity variables from the IMDB datasets, which though not strongly correlated with ratings independently, appear to bolster predictiveness otherwise.



Scope for Further Research

While it may never be possible to identify metrics that predict adaptability or adaptation success, such work can provide inroads into research about emerging genres such as virtual reality spaces and interactive filmmaking etc, especially in helping existing creators transfer their skills effectively and informedly. These new media will likely funciton in opposition to traditional adaptations, precisely because they have arisen to fulfil needs not met by television or film. Crucial to that need is immersion, represented by sophisticated world-building and the minimization of narrative distance, as opposed to action or characterization, which have been of totalizing significance to film/TV adaptations.