R. Reussner, A. Koziolek, R. Heinrich (Hrsg.): INFORMATIK 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021 1271

SubRosa: Determining Movie Similarities based on Subtitles

Jan Luhmann[1] Manuel Burghardt[2] Jochen Tiepmar[3]

Abstract: For streaming websites, media shopping platforms and movie databases, movie recommendation systems have become an important technology, where mostly hybrid methods of collaborative and content-based filtering on the basis of user ratings and user-generated content have proven to be effective. However, these methods can lead to popularity-biased results that show an underrepresentation of those movies for which only little user-generated data exists. In this paper we will discuss the possibility of generating movie recommendations that are not based on user-generated data or metadata, but solely on the content of the movies themselves, confining ourselves to movie dialog. We extract low-level features from movie subtitles by using methods from Information Retrieval, Natural Language Processing and Stylometry, and examine a possible correlation of these features' similarity with the overall movie similarity. In addition we present a novel web application called *SubRosa* (`http://ch01.informatik.uni-leipzig.de:5001/`), which can be used to interactively compare the results of different feature combinations.

Keywords: Movie Similarity; Subtitles Processing; Information Retrieval; Stylometry; Natural Language Processing

1 Introduction

With a rapidly increasing number of movies produced each year, a growing number of film industries worldwide[4] and new possibilities of distribution via streaming websites, such as *Netflix* and *Amazon Prime*, or on-demand services like *Vimeo* and *Youtube Movies*, recommendation systems for movies have become an essential tool to enhance the user experience: Users who are eager to discover and watch movies unknown to them would be completely lost at the attempt to single-handedly pick the ones they are interested in from the mass of released movies. Currently used movie recommendation systems are largely based on collaborative filtering [BL07; SK09]. A major challenge for collaborative filtering recommendation systems is the so-called cold start problem. A cold start, i.e. the case that a movie does not yet have enough user ratings or that a user did not yet rate enough movies to calculate any recommendations using collaborative filtering, is mostly handled using content-based approaches:
A basis of movie similarities can be determined for initial recommendations using metadata

[1] Leipzig University, Computational Humanities Group, Leipzig, Germany, jan.luhmann@gmx.net
[2] Leipzig University, Computational Humanities Group, Leipzig, Germany, burghardt@informatik.uni-leipzig.de
[3] Leipzig University, Computational Humanities Group, Leipzig, Germany, jtiepmar@informatik.uni-leipzig.de
[4] UNESCO Institute of Statistics. (2016) Record number of films produced. `http://uis.unesco.org/en/news/record-number-films-produced` (date accessed: 2019-08-29)

doi:10.18420/inf2020_119

provided by the movie distributor such as genre tags, list of cast and plot synopsis, as well as tags regarding plot, style and mood of a movie [Sa01].
However, recommendation systems using this approach still suffer from a popularity bias: Any movie that has not been sufficiently tagged by users or whose metadata is only fragmentary is a lot less likely to survive a cold start. To increase diversity and novelty in movie recommendations, it would be greatly beneficial to be able to estimate movie similarities independently of any human-supplied attributional data but based on the content of movies themselves.
In addition of the actual video and audio data of a movie, a third resource which can represent one aspect of a movie, i.e. its dialog, is its subtitles. Of course, subtitles can only contain a fraction of the information which a movie's dialog provides, completely missing information about speakers, intonation, facial expressions etc. But since it still may contain information about dialog topics and manner of speaking, and since English subtitles – even for little known films – are widely available today through online platforms such as *OpenSubtitles*[5] and can be processed inexpensively and efficiently in comparison to video or audio data, we will here discuss and explore the possibility of detecting movie similarities using feature extraction from subtitles. The applied methods of feature extraction are related to Natural Language Processing (NLP), Information Retrieval (IR) and Stylometry.

1.1 Related Work

Before we present the experimental setup, we discuss a number of works that use subtitles and movie scripts as a basis to calculate similarities between movies.
Blackstock; Spitz [BS08] propose a method for classifying 399 movies by NLP-related features extracted from movie scripts. Their method is partly stylometric, examining the ratios and distributions of occurrences of grammatical word forms using Part-of-Speech tagging (POS), partly statistical using features derived from speaker annotations which are present in movie scripts, and partly based on Named Entity Recognition (NER), for the analysis of identical named entities. Movies are classified by genre using Maximum Entropy Markov Models and Naive Bayes. While stylometric features achieve the best results, the overall accuracy is relatively low. The authors conclude that a larger and more diverse dataset would have improved their results. However, freely and digitally available movie scripts are much harder to obtain than subtitle files, and they are also more difficult to process.
Nessel; Cimpa [NC11] propose a movie recommendation engine which uses an Inductive Inference-based method of calculating similarities among subtitle texts of 290 pre-selected movies. The results of the evaluation experiments look promising, although it is difficult to say how their approach would perform on a more diverse dataset.
Bougiatiotis; Giannakopoulos [BG17] examined the correlation of movie similarities and features extracted from subtitles and audio, using a dataset of subtitle files of 160 movies.

[5] http://www.opensubtitles.org

Bag-of-Words (BOW) representations of subtitle text were used for calculating topic models. Segments of audio data were classified by event (music, speech, noises etc.) and in case of music classified by music genre. In evaluations, extracted audio features only yield very low accuracy scores. The two most accurate results are generated by topic modeling using Latent Dirichlet Allocation, and by simple tf-idf weighting of BOW representations.

2 Experimental Setup

2.1 Dataset

Our dataset consists of English subtitles for 5,914 movies. These movies are all among the 10,000 most rated movies on *IMDb*[6]. Despite our motivation to tackle a popularity bias, we chose rather well-known movies to be able to better assess the results of our experiments. We decided to use such a large and diverse dataset because it may improve the quality of some models and more accurately represent a later real-world application.
Subtitles were kindly made available by *OpenSubtitles*. They claim that their database only offers files that can be freely and legally distributed. For each movie they provide us with several versions of subtitles. Frequently, subtitles contained OCR errors (optical character recognition), encoding errors, and for our purposes unwanted data like speaker annotations, music lyrics, authorship tags by the subtitle creator and HTML markup. To minimize the occurrence of such data, subtitles for all movies undergo a heuristic cleaning, validation and selection process, which leaves us with one selected subtitle file for each of 5914 movies, formatted as a SubRip file (*.srt). Additionally, metadata (*IMDb*-ID, title, release year, genres, runtime, number of ratings, rating) for these movies are obtained from *IMDb*.

2.1.1 Preprocessing

For some of our feature extraction methods which are described in Sect. 2.2 it is necessary to convert a movie's subtitles into a continuous text. This is done by simply joining all dialog lines to a single string, separated by whitespace. The text is then further processed using *spaCy*'s language processing pipeline[7]. In our case, this pipeline consists of the following stages: tokenization, POS tagging, dependency parsing, NER, lemmatization. For each movie, we store its sequence of tokens, lowercased and without punctuation, additionally in lemmatized form, and their corresponding POS tags. We filter out interjections, and named entities which refer to persons, organizations or places because these would heavily distort the results of the Bag-of-Words Model discussed in Sect. 2.2.1.

[6] http://www.imdb.com
[7] http://www.spacy.io/

2.2 Methods for Feature Extraction

In this section the applied methods for feature extraction are presented. Their configurations and the adjustments of their parameters were determined experimentally, i.e. by testing them on randomly selected samples of our dataset and examining the results.
There is one method for topical analysis of documents, in this case with the motivation to address the question: *What are the characters of a movie talking about?* (see Sect. 2.2.1). Four methods are aimed at stylistic analysis, in an attempt to address the question: *How are the characters of a movie talking?*, considering aspects of lexicality, syntax and speech rhythm (see Sect. 2.2.2 to 2.2.4 and 2.2.6). A sixth method is aimed at an analysis of emotions: *Which emotions are the characters expressing in their speech?* (see Sect. 2.2.5)

2.2.1 Bag-of-Words Model (BOW)

The Bag-of-Words model is a simple approach for representing text documents in Information Retrieval and Natural Language Processing. In our model, the subtitle text for a movie (a document) is represented by its set of unigrams of lemmatized tokens, weighted by sublinear-tf-idf scaling [MRS08] which is a logarithmic variation of tf-idf scaling. By logarithmizing the term frequency factor, the actual number of occurrences of a term in the document has a less drastic effect on the weighting. To filter out stop words and to reduce dimensions of our document representation, all terms which occur in more than 95% of all documents are ignored. This limit is adjusted to the occurrence of "traditional" stop words. Likewise, all terms which occur in less than 2.5% of all documents are ignored. This limit is adjusted to the occurrence of terms containing spelling mistakes. Finally remaining are 4,952 terms, so that documents are represented by this model as vectors of dimension 4,952.

2.2.2 Distribution of Stop Words (SWD)

One method that has long been successfully used in stylometry, particularly in studies of author attribution, is to measure the rates of occurrences of stop words in a given document and to understand the distribution of these as a fingerprint of the author's writing style. This was mainly developed by John Burrows during the late 1980s [Bu87]. This method is applied on our dataset of unlemmatized tokens. Based on the NLTK stop word list and the most common terms in our corpus, we manually generate a set of 87 stop words. Document representations are then computed in the same way as our BOW-Model but considering only the terms of the stop word set, weighted simply by their term frequency.

2.2.3 POS Tag Trigrams (PTT)

The method of extracting features related to syntactic structures is commonly used for genre classification of text documents. Genres in this case are usually fiction, academic text, news text, conversation, etc., where the syntactic structure is a discriminating factor between genres, regardless of the topic of texts. Santini [Sa04] proposes the use of trigrams of POS tags which "are large enough to encode useful syntactic information, and small enough to be computationally manageable ", so we choose to do the same in this work. The POS tags emitted by *spaCy* are from the "OntoNotes 5 version of the Penn Treebank tag set." [8] Only trigrams within sentence boundaries are considered. Similarly to the method presented in Sect. 2.2.2, we are interested in the distribution of the frequencies of globally frequent, common features among our documents. For this reason we ignore all trigrams that occur in less than 90% of our documents, resulting in 417 trigrams, weighted by their frequency.

2.2.4 Stylometric & Statistical Measurements (SSM)

Next we calculate a model of document representations based on various ratios and measurements which are popular in stylometric studies and which are once again often successfully used for genre classification and authorship attribution. These include Shannon Entropy [Sh48] of term probabilities, standardized type-token ratio [Jo44; TC13], average length and ratios of lengths of sentences and words.

2.2.5 Sentiment Analysis (SEA)

Using sentiment analysis of a document of our dataset we may estimate the emotional arc of the respective movie, assuming that this arc is in some way reflected in the dialog. We use the well evaluated open-source tool *VADER Sentiment*[9] [HG14] for calculating a *compound sentiment score* for a given text, which reflects positive or negative sentiment on a scale from 1 to -1. Since emotions usually evolve very intensely over the course of a movie, it is not useful to calculate the sentiment of an entire document of movie dialog at once. Instead, for each second of a movie we calculate the sentiment of the dialog spoken, resulting in a sentiment curve (Fig. 1). Smoothing by simple moving average (window size 10) is applied. As features of a curve, we calculate its mean value, its first and third quartile, its rate of zero crossings and the rates of zero crossings of the first and second derivative. The latter metrics are derived from signal processing [Ch88] and may be replaced in future work by more robust metrics to measure entropy, volatility and oscillation of time series.
Fig. 1 shows the highly different sentiment curves for the horror movie "Evil Dead" (2013) and the romantic musical "La La Land" (2016).

[8] `https://spacy.io/api/annotation#pos-tagging` (date accessed: 2019-08-29)
[9] `https://github.com/cjhutto/vaderSentiment` (date accessed: 2019-08-29)

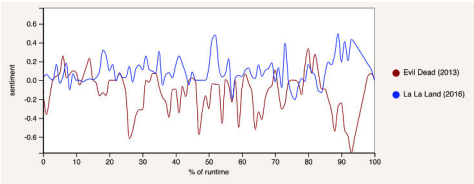

Fig. 1: Detail view of Sentiment Analysis curves in *Sub Rosa*.

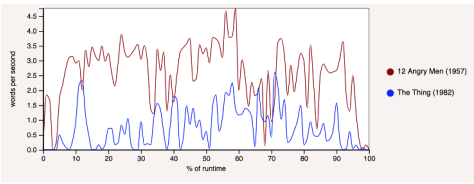

Fig. 2: Detail view of Speech Tempo Analysis curves in *Sub Rosa*.

2.2.6 Speech Tempo Analysis (STA)

It is intuitively understandable that distinctive features of a movie may be the tempos at which characters speak as well as the frequency and duration of speech pauses. These may correlate with the rhythm of a movie's editing which is crucial to its style and atmosphere [Va85]. For each second of a movie, we calculate the speech tempo by approximating how many words are spoken during the second, resulting in a speech tempo curve (Fig. 2). Smoothing by simple moving average (window size 10) is applied to the curve. We extract the same features as from the SEA curves.
Fig. 2 shows the highly different speech tempo curves for the slow-paced horror movie "The Thing" (1982) and the dialog-driven courtroom drama "12 Angry Men" (1957).

2.3 Similarity Measurement

For each of the mentioned models we have generated a unique feature vector. To determine the similarity between two movies' subtitles, we calculate the distances between the six respective vectors (BOW, SWD, PTT, SSM, SEA, STA). For the BOW model, cosine distance is used. For all other models, cosine delta is used as a distance metric. Compound distance scores are the weighted averages of all models' distance values. The weights can be set manually in the web application that is described in the next section.

2.4 Web Application for Interactive Evaluation

So far, we have been using various parameters to determine the similarity of movies based on their subtitles. Our primary research goal in this project and future work is to learn more about the expressiveness and validity of the different features, i.e. which feature or combination of features (and different weights of the features) provides the best similarity measure for movies? And accordingly: Are there differences for movies from different directors, genres or dates? To investigate these questions in an exploratory way, we designed a web application called *Sub Rosa*, which is available via `http://github.com/bbrause/subrosa`. We also provide a live demo of the application that can be found

at http://ch01.informatik.uni-leipzig.de:5001/. *SubRosa* allows users to adjust the weighting of feature models based on which compound distance scores are calculated. Users can request the nearest neighbors of any movie which are visualized in a graph in which each node represents a movie and the length of the edge between each two movies is proportional to the square of the compound distance score calculated between them. In addition, specific movies can be compared in more detail by providing information about the most frequent 200 tokens in the BOW model, POS and stop word distributions, stylometric features, sentiment and speech tempo scores.

3 Results

Beyond the exploratory testing of the *SubRosa* web interface, we can estimate whether our models succeed to determine similarity between movies by analyzing two-dimensional projections of all vectors of the models. These were made by first reducing to 50 dimensions using Truncated SVD[10] and then further reducing to two dimensions using t-SNE for visualization [MH08]. I The color of each data point matches the genre of the movie it represents. Interactive plots for each single model as well as for some weighted combinations of models are available via https://chart-studio.plot.ly/~bbrause/#/

Fig. 3 shows a plot of all six unweighted features. Although some genres, e.g. Comedies and Horror movies, seem to form (partial) clusters, most of the other genres are highly scattered in the feature space. In other words: movie similarity based on dialogs does not manifest itself clearly in the existing genre definitions, i.e. a Comedy with a war theme (e.g. "The Men Who Stare at Goats", 2009) is rather rated similar to other War movies than to other comedies.

Fig. 4 provides an overview of plots that were created for each singular model. It shows that each models produces rather different similarity patterns. In a way the BOW model stands out, as it produces at least some visible clusters of movies. Taking a closer look at those clusters, Fig. 5 underlines the previous observation: although some genres tend to form clusters (e.g. Western), most clusters are formed by either a common cultural theme (e.g. Indian movies) or some other kind of theme (e.g. Religion or War).

An obvious next step would be a more detailed evaluation of the quality of our models, which will be possible if a ground-truth dataset of human-estimated similarities of movies can be found.

4 Conclusions

We presented an experimental setup to determine the similarity of movies based on different features that can be extracted from the subtitles. This is not only important to generate objective recommendations for movie consumers, but can also help to aid computer-based

[10] https://scikit-learn.org/stable/modules/decomposition.html (date accessed: 2019-08-29)

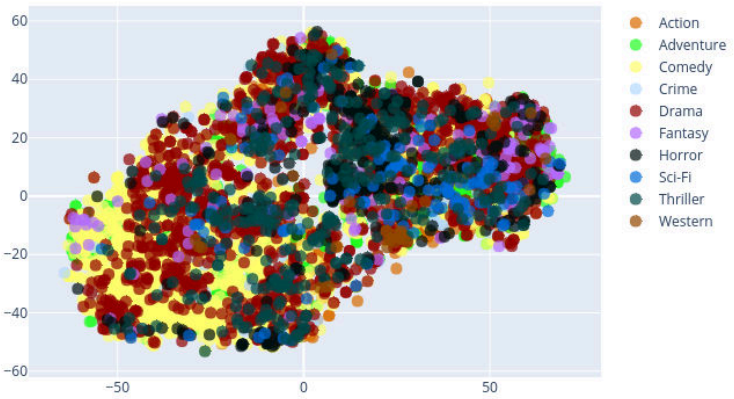

Fig. 3: All feature models concatenated (unweighted), 2D projection using Truncated SVD and t-SNE.

film studies, which are currently a trend in the Digital Humanities[11]. *SubRosa* is intended as an interactive tool that allows users to experiment with with different features and weights, to see how different parameter settings have an effect on the results. In addition, we created some very basic plots that show that obvious similarities (beyond genre classifications) can be identified with our different models. However, the question remains: which features or combination of features yields the best results for the detection of similar movies?
As was indicated in the previous section, we are planning to do a systematic evaluation with all of the parameter configurations as an obvious next step, as we would like to know what is the best setting to detect movie similarity on the dialog level. A possibility for proper evaluation of our experiment may be provided by the "MovieLens" dataset by the research lab *GroupLens* (University of Minnesota, USA)[12]. It offers user-generated tags, mostly related to style, mood, plot or setting, for 58,000 movies. Similarities of movies regarding their tags may be compared with their similarities regarding our models.

5 Acknowledgements

We would like to thank the *OpenSubtitles* team for providing part of their subtitle database, which made this work possible in the first place.

[11] See the SIG AudioVisual material in Digital Humanities, `https://avindhsig.wordpress.com/`.
[12] `http://grouplens.org/datasets/movielens/`
(date accessed: 2019-08-29)

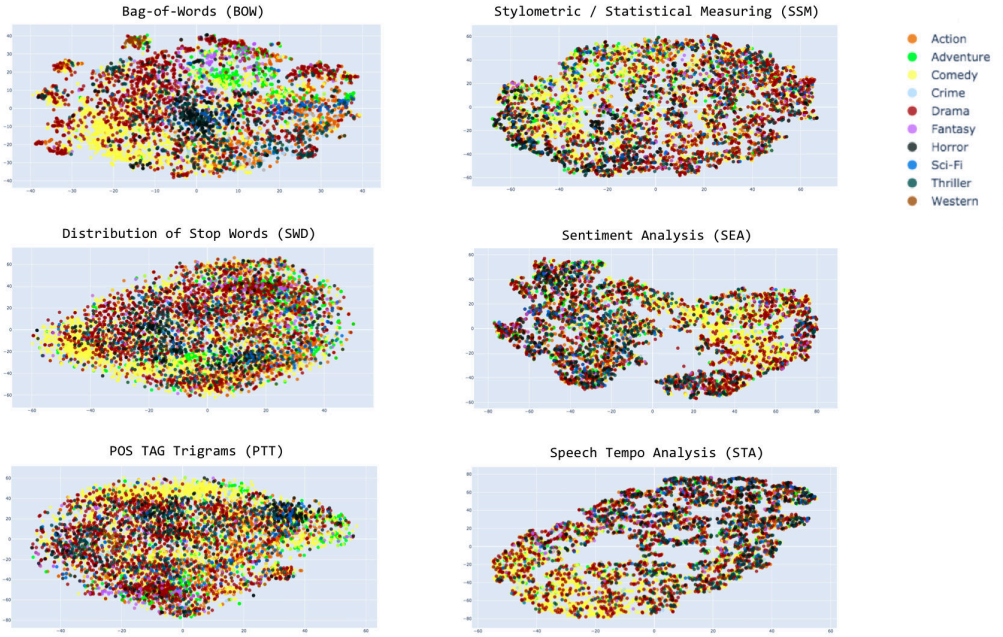

Fig. 4: Plots for the six separate feature models.

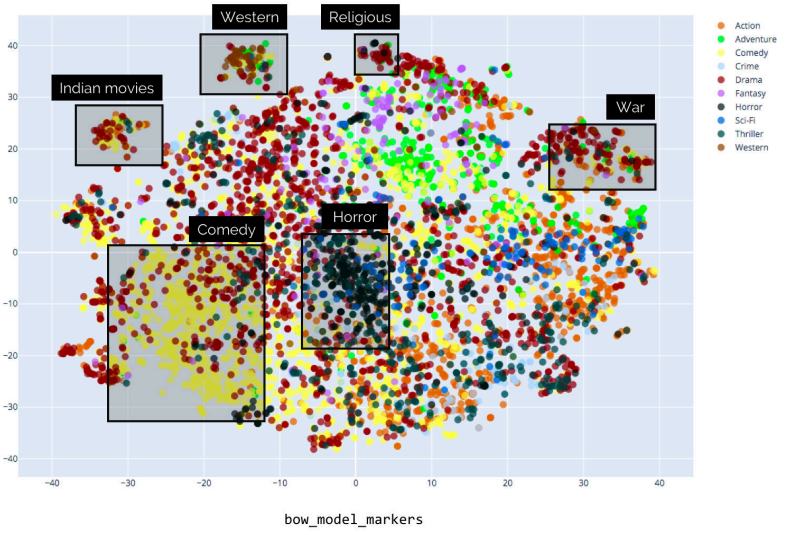

Fig. 5: Annotated BOW plot.

References

[BG17] Bougiatiotis, K.; Giannakopoulos, T.: Multimodal Content Representation and Similarity Ranking of Movies, arXiv preprint arXiv: 1702.04815, 2017.

[BL07] Bennett, J.; Lanning, S.: The Netflix Prize, 2007.

[BS08] Blackstock, A.; Spitz, M.: Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features, 2008.

[Bu87] Burrows, J.: Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Clarendon Press and Oxford University Press, 1987.

[Ch88] Chen, C.-H.: Signal processing handbook. CRC Press, 1988.

[HG14] Hutto, C. J.; Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media. 2014.

[Jo44] Johnson, W.: Studies in language behavior: A program of research. Psychological Monographs 56/2, pp. 1–15, 1944.

[MH08] van der Maaten, L.; Hinton, G.: Visualizing Data using t-SNE, 2008.

[MRS08] Manning, C.; Raghavan, P.; Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, 2008.

[NC11] Nessel, J.; Cimpa, B.: The MovieOracle - Content Based Movie Recommendations, 2011.

[Sa01] Sarwar, B.; Karypis, G.; Konstan, J.; Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the tenth international conference on World Wide Web - WWW '01. the tenth international conference. ACM Press, Hong Kong, Hong Kong, pp. 285–295, 2001.

[Sa04] Santini, M.: A Shallow Approach To Syntactic Feature Extraction For Genre Classification. In: Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK 2004). 2004.

[Sh48] Shannon, C. E.: A mathematical theory of communication. Bell system technical journal 27/3, pp. 379–423, 1948.

[SK09] Su, X.; Khoshgoftaar, T. M.: A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence 2009/1, Jan. 2009.

[TC13] Torruella, J.; Capsada, R.: Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. Procedia - Social and Behavioral Sciences 95/, pp. 447–454, 2013.

[Va85] Van Leeuwen, T.: Rhythmic Structure of the Film Text. In (Dijk, T. v., ed.): Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication. Walter de Gruyter, pp. 216–232, 1985.