

TV SERIES RECOMMENDATION SYSTEM WITH MACHINE LEARNING

Dr. Paul Raj D^{#1}, Arshad Ahad B^{*2}, Alfred Shadrach Samuel A^{*3}, Deepak Kumar L^{*4}

[#]Professor, Head of Department, ^{*}UG Scholar Department of Computer Science
R.M.K College of Engineering and Technology, India

Abstract— The growth of series has largely emerged in recent times over many years. People have started to allocate separate time for watching series. As the popularity of series have been expanded, there are numerous amounts of series with different genres and plots, nearly every week a new series is getting released. But this throw's a confusion to the person to decide which series he should start watching. In addition, the offer is nowadays so big that is difficult for some shows to get some exposure due to "blockbusters" or very hype programs getting most of it (Long Tail Phenomenon). Though there are many top-rated series, not everyone has the same taste. That is why some shows are not considered by many people even if they might be a good and enjoyable match for them. So, the idea is to recommend a person according to the wish of his/her genre by segregating a series' genre through the subtitles. The main idea behind this technique is to use Latent Dirichlet Allocation method in Machine Learning.

Keywords — Series, Genre

I. INTRODUCTION

Our aim is to create a website that recommends series to users according to their wish of genre by predicting the genre of a series which is done by running the subtitles and determining the percentage of genre by each word.

1.1 Data Acquisition

Subtitles are acquired from tvsubtitles.net because there are no limited number of downloading in that website. The data from tvsubtitles.net does not follow the required hierarchy we have to make a script to format it in any desired way. There are duplicate subtitles in some TV Shows, they are erased using python scripting.

1.2 Formatting the Data

The downloaded file will be of the format *.srt. Subtitle synchronization, subtitle effect tag, special characters must be removed from the downloaded file. The format must also be changed from *.srt to *.txt. We keep important tokens such as nouns, verbs, adjectives, etc. and removing others such as pronouns, determinant, etc. It allows us to remove information less words for segregating the genre. Now we our words, they are counted and written in a file next to their number of occurrences, this will later be used for determining the genre.

1.3 Analysing the Data

In further process, the refined text data will be executed by Latent Dirichlet Allocation which will return the type of genre as output.

1.4 User Interface



Creating a HTML, Java Script embedded website where the user can access these data according to his preferences.

II. LITERATURE SURVEY

2.1 Content-Boosted Collaborative Filtering

It gives an approach to combine content and collaboration to enhance existing user data and give better performance than a pure content based predictor.

2.2 FAB Technique

An adaptive recommendation service for collection and selection of web pages. It makes the system more personalized and combines the benefits of content analysis with shared user interests.

2.3 Bayesian Hierarchical Model (BHM)

Proposes a faster technique to gather a huge number of individual user profiles even if feedbacks available are less. It uses various parameters of BHM for optimization of joint data likelihood.

III. PROPOSED SYSTEM

1. Latent Dirichlet allocation

The method used in this system is Latent Dirichlet Allocation. It is a generative probabilistic model of a corpus. The idea is to represent document with random mixtures over latent topics, where each topic is characterized by a distribution over words.

Latent Dirichlet allocation (LDA) assumes the following generative each document w in a corpus D :

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n :
 - a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
 - b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n .

The Dimension of k is assumed known and fixed in the Dirichlet Distribution. The word probabilities are represented as $k \times V$ matrix β where $\beta_{ij} = p(w^j=1|z^i=1)$. N is Independent from other variables like θ, z . thus eliminating randomness in the subsequent development.

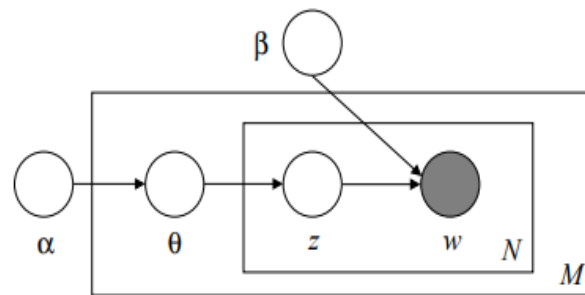
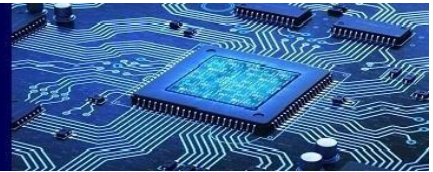
A k -dimensional Dirichlet variable θ can take any value in $(k-1)$ simplex, and has the following probability density on simplex.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$$

parameter α is k -vector where as $\Gamma(x)$ is the Gamma function. With the parameters α and β the joint distribution with topic θ , set of Z topics and set of W words:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^N p(z_n|\theta) p(w_n|z_n, \beta),$$

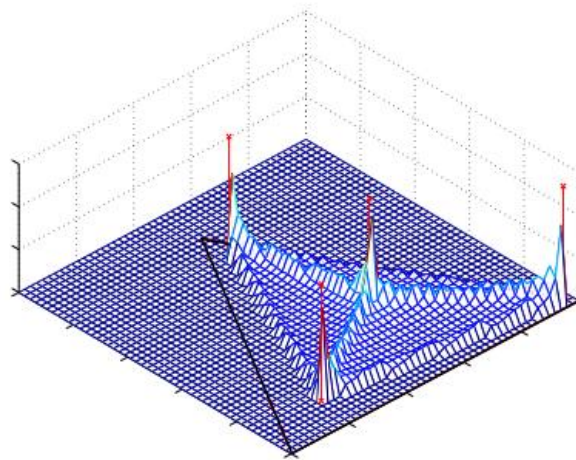
In the graphical representation of LDA . the plates are represented as boxes, the outer box represent the document, whereas the inner box represents the words and topics in the Document.



2. Continuous mixture of unigrams

The LDA model in the above figure is more elaborate than two-level models. However by marginalizing over hidden topics z , LDA can be two-level model as:

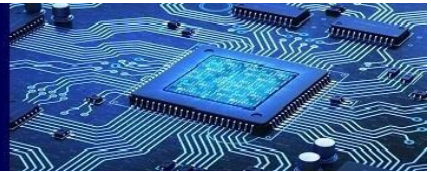
$$p(w|\theta, \beta) = \sum_z p(w|z, \beta)p(z|\theta).$$



A model thickness on unigram appropriations ($w|\theta, \beta$) under LDA for three words and four subjects. The triangle installed in the x - y plane is the 2-D simplex addressing all conceivable multinomial conveyances more than three words. Every one of the vertices of the triangle relates to a deterministic appropriation that allocates likelihood coordinated of the words; the midpoint of an edge gives likelihood 0.5 to two of the words; and the centroid of the triangle is the uniform dispersion over each of the three words. The four focuses checked with a x are the areas of the multinomial appropriations $p(w|z)$ for every one of the four points, and the surface appeared on top of the simplex is an illustration of a thickness over the $(V - 1)$ -simplex (multinomial appropriations of words) given by LDA.

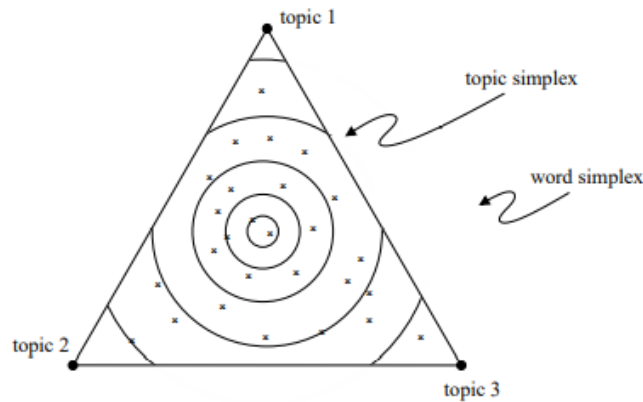
1. Geometric interpretation.

A decent method of representing the contrasts among LDA and the other inert subject models is by



thinking about the math of the inactive space, and perceiving how a report is addressed in that math under each model.

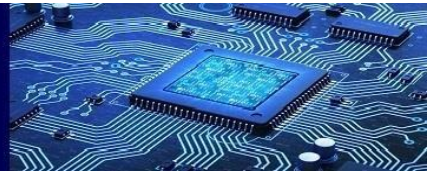
The point simplex for three subjects installed in the word for three words. The corners of the word simplex compare to the three dispersions where each word (respectively) has likelihood one. The three places of the theme simplex compare to three various conveyances over words. The combination of unigrams puts each archive at one of the sides of the theme simplex. The pLSI model instigates an experimental dissemination on the point simplex meant by x . LDA places a smooth circulation on the theme simplex indicated by the form lines.



2. Application and Results

The observational assessment of LDA in a few issue Domains - Document demonstrating, report characterization, and shared separating. Taking all things together of the combination models, the normal complete log probability of the information has nearby maximize the focuses where all or a portion of the blend segments are equivalent to one another. To maintain a strategic distance from these nearby maxima, it is imperative to introduce the EM calculation suitably. In our investigations, we instate EM by cultivating each contingent multinomial appropriation with five records, reducing their compelling complete length to two words, and smoothing across the entire jargon. This is basically an estimate to the plan depicted in Heckerman and Meila (2001).

Let us consider an example paragraph from AP corpus. Each colour represents a different factor from the words generated putatively. The text will look like as figure given below:



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

From this stage the percentage of each factor is calculated from the given formula

$$\frac{\text{total number of words in a factor}}{\text{total number of words in the file}} \times 100 = \% \text{ of single factor}$$

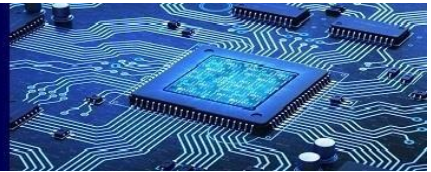
and is shown as the output for the user to decide whether it is recommended or not according to his preferences.

IV. CONCLUSION

We can conclude that our Series Recommendation System has a unique feature that no other proposed system has which is segregating the genre of a series with the help of subtitles to determine the percentage of genre present in a series. This system helps the user to find the genre of series he/she wishes to seek. This refined result shows accurate genre which is totally obtained from the subtitle of a series not the description or storyline, in which some cases the description won't match with the genre. Our system removes this defect. So the desired outcome of our system would be percentage of action, comedy, drama, mystery, romance, etc.

V. RESULT

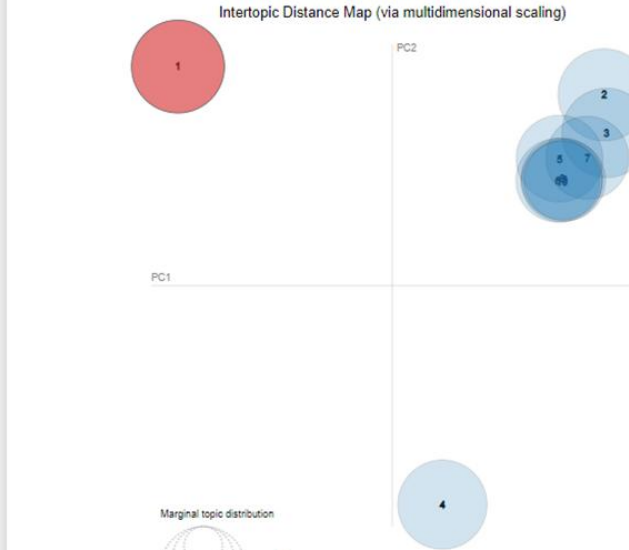
We provide a recommendation not according to our wish or the machine's choice, it is completely up to the user because of the user enters the amount of percentage of genre he/she likes to watch which will produce multiple results as a list of TV Series' that satisfies the user's criteria. For our implementation, we have used Latent Dirichlet Allocation method which will segregate the words present in the subtitle and allocate the resultant into specific genre which will finally display the percentage of genre present in a TV Series.



jupyter first attempt on sub Last Checkpoint: 03/10/2021 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3


Intertopic Distance Map (via multidimensional scaling)



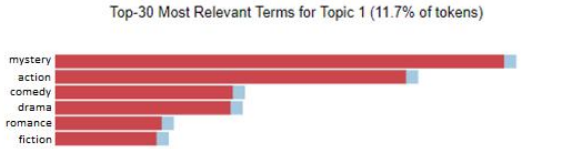
PC2

PC1

Marginal topic distribution




Top-30 Most Relevant Terms for Topic 1 (11.7% of tokens)



Overall term frequency (blue bar)
Estimated term frequency within the selected topic (red bar)

Series Recommendation Website x +

localhost/series/index.php

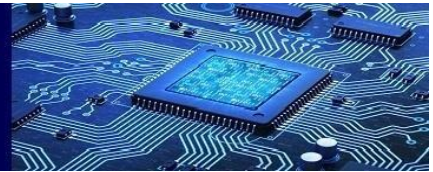


Series Recommended For You are :

Name	Action	Comedy	Romance	Mystery	Fiction	Drama
Breaking Bad	30	12	9	28	1	20
Money Heist	36	12	21	28	0	3
Lucifer	17	30	24	21	8	0
Sherlock	18	12	4	48	0	18
Friends	1	82	16	0	0	1
How I Met Your Mother	5	34	30	1	2	22

Type here to search

07:12 26-03-2021



REFERENCES

1. Prem Melville, Raymond Mooney, Ramadaas Nagarajan [21]
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.
3. P. Diaconis. Recent progress on de Finetti's notions of exchangeability. In *Bayesian statistics, 3* (Valencia, 1987), pages 111–125. Oxford Univ. Press, New York, 1988.
4. G. Salton and M. McGill, editors. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
5. M. Leisink and H. Kappen. General lower bounds based on computer generated higher order expansions. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference, 2002*.
6. C. Morris. Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–65, 1983. With discussion.
7. J. Dickey. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78:628–637, 1983.
8. M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.
9. C. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. pages 159–168, 1998.
10. G. Ronning. Maximum likelihood estimation of Dirichlet distributions. *Journal of Statistical Computation and Simulation*, 34(4):215–221, 1989.
11. D. Heckerman and M. Meila. An experimental comparison of several clustering and initialization methods. *Machine Learning*, 42:9–29, 2001.
12. R. Kass and D. Steffey. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84 (407):717–726, 1989.