University of South Florida

  CIS 4915:Supervised Research in Computer Sci

Research Report

|  |  |  |  |
| --- | --- | --- | --- |
| **Semester: Spring 2023** | | | |
| ***Writer:*** Geonhee Choi | | | |
| ***Class*** | *Section:* | 003 |
| *Professor* | Seungbae Kim |
|  | | | |
|  | | | |

# Abstract:

This research project aims to analyze public engagement with news channels on YouTube by focusing on three primary aspects: data cleaning of the YouTube dataset, topic modeling using titles and scripts of news videos, and constructing and analyzing the social network of news channels and their subscribers. By examining these factors, the study seeks to provide insights into audience preferences, patterns of interaction, and the relationships between news channels and their subscribers. The findings of this research can potentially guide content creators, news organizations, and policymakers in understanding the dynamics of public engagement with news content on YouTube and shaping their strategies accordingly.

# 1)Introduction:

* 1. **Background:**

In today's digital landscape, news channels have extended their reach to online platforms like YouTube, where millions access news content daily. As a crucial platform for news dissemination and public engagement, understanding the dynamics of public engagement on YouTube is vital for content creators, news organizations, and policymakers to develop effective strategies catering to audience preferences and fostering healthy interactions. This research project analyzes public engagement with YouTube news channels by focusing on three aspects: data cleaning of the YouTube dataset, topic modeling using news video titles and scripts, and constructing and analyzing the social network of news channels and subscribers. Through examining these factors, the study aims to provide insights into audience preferences, interaction patterns, and relationships between news channels and subscribers, potentially guiding stakeholders in understanding public engagement dynamics and shaping their strategies accordingly.

**1.2 Research Objectives:**

1)To clean and preprocess the YouTube dataset for effective analysis of public engagement with news channels on the platform.

2)To perform topic modeling on news video titles and scripts, identifying prevalent themes and trends in news content.

3)To construct and analyze the social network of news channels and their subscribers, revealing patterns of interaction and connectivity.

4)To gain insights into audience preferences, engagement patterns, and relationships between news channels and their subscribers.

# 2)Research Method:

**2.1 Data Collection:**

To begin the analysis of YouTube data related to the research topic, datasets were acquired from the following sources:

Channel-level dataset: https://bigdata-region.kr/#/dataset/276e2693-4aab-48d7-b419-68fc082a3e1a

Video-level dataset: https://bigdata-region.kr/#/dataset/d873e42d-9897-4626-b535-551ea6527f3a

Video-level comment data: https://bigdata-region.kr/#/dataset/750c5d47-5f44-4cae-8e86-1207f25e624a

Commenter-level dataset: <https://bigdata-region.kr/#/dataset/0c7e93e7-8271-4a54-8b62-1dd86824580a>

**2.2 LDA Topic Modeling:**

The data analysis stage of this research project involves processing and examining the cleaned dataset to identify patterns and insights into public engagement with news channels on YouTube. This section describes the methods and techniques applied during the data analysis process. I extracted each youtube title from dataset by python. This script is part of a larger data analysis process, focusing on extracting YouTube video titles from a dataset and cleaning the CSV files. The purpose of this script is to consolidate the data from multiple sources, remove any null bytes, and prepare the data for further analysis. After extracting of you tube title, I started to analyze the content of video titles and comments as Topic Modeling method. Especially, Latent Dirichlet Allocation (LDA), which classifies documents into a predefined number of topics based on the co-occurrence of words. By applying LDA, the most prominent topics from YouTube title and comments can be identified, shedding light on the audience's preferences and interests.

This section describes the methods and techniques applied during the data analysis process, with a focus on the topic modeling of video titles using Latent Dirichlet Allocation (LDA).

1. **Data Preparation**: The dataset containing video titles is loaded into a Pandas dataframe for further processing. The titles are cleaned by removing special characters and non-alphanumeric characters using regular expressions. This preprocessing step ensures that the text data is ready for tokenization and LDA modeling.
2. **Tokenization**: The cleaned text data is tokenized using the Konlpy library's Okt tokenizer. This step breaks down the text into individual words or tokens, which are then used as input for the LDA model. Tokenization is crucial for text analysis, as it allows the model to identify meaningful patterns and relationships between words.
3. **Dictionary Creation and Corpus Generation**: The tokenized text data is converted into a dictionary representation using the Gensim library's Dictionary class. This dictionary maps unique words in the dataset to integer IDs, which are used for efficient computation. Furthermore, the tokenized words are transformed into a bag-of-words representation, which serves as the input corpus for the LDA model.
4. **LDA Modeling**: The LDA model is trained using the Gensim library, with the prepared corpus and dictionary as input. The number of topics is set to 20, which determines the granularity of the topics generated by the model. The LDA model identifies the most dominant topics in the dataset based on the co-occurrence of words in the video titles, providing insights into the audience's preferences and interests.

LDA Topic modeling script Pseudo code:

1. Import necessary libraries (pandas, konlpy, gensim, matplotlib, and re)
2. Load the CSV file containing video titles into a Pandas dataframe
3. Reset the index of the 'title' column and store it in the variable 'text\_data'
4. Convert each document in 'text\_data' to a string format
5. Define a regex pattern to match non-alphanumeric characters
6. Remove special characters from each document in 'text\_data'
7. Initialize the Okt tokenizer from the Konlpy library
8. Tokenize each document in 'text\_data' using the Okt tokenizer
9. Create a dictionary of the words using Gensim's Dictionary class
10. Convert the tokenized word lists into a bag-of-words representation
11. Train the LDA model with the corpus, dictionary, and the specified number of topics (20)
12. Print the generated topics and their corresponding words

**2.3 Network Edge and Weight :**

This python script generates a video interaction network based on user comments on videos. It reads a CSV file containing user IDs (authorChannelId) and the video IDs (videoId) they commented on. The network is represented as a dictionary of dictionaries, with the video IDs as keys and their interactions (edges) with other video IDs as values. The edge weight represents the number of common users who commented on both videos.

This pseudocode provides a high-level overview of the steps involved in the given code. It focuses on the main tasks such as loading the data, preprocessing the text, tokenizing the documents, creating the dictionary and corpus, training the LDA model, and displaying the generated topics.

Network Edge and Weight pseudocode:

1. Import the required libraries (pandas and defaultdict from collections)
2. Set the input file path
3. Create a defaultdict 'network' to store the network information
4. Create a defaultdict 'users' to store the users and the videos they commented on
5. Read the input CSV file using pandas and store it in a DataFrame 'df'
6. Group the DataFrame 'df' by 'authorChannelId' and store the result in 'grouped'
7. For each group in 'grouped', do the following: a. Get the list of video IDs for each user and store it in 'video\_ids' b. Update the 'users' defaultdict with the user's video IDs c. Iterate through each video ID in 'video\_ids': i. Iterate through the remaining video IDs: 1. Increment the edge weight between the two video IDs in the 'network' defaultdict
8. Print the network information: a. For each video pair with edge weight greater than 900, print the video IDs and the edge weight

This code reads a CSV file containing information about users and the videos they commented on. It constructs a network where videos are nodes and the edge weights represent the number of users who commented on both videos. Finally, it prints the video pairs that have an edge weight greater than 900.

**2.4 Network Visualization:**

This script visualizes a weighted directed graph using the NetworkX library and Matplotlib for rendering. It takes a dictionary of edges as input data, where the keys are source nodes and the values are lists of tuples containing target nodes and edge weights. The script then creates a directed graph and adds the edges and their weights. Finally, it displays the graph with node labels, edge labels, and arrowheads to indicate direction.

Network Visualization pseudocode:

1. Import networkx and matplotlib.pyplot libraries
2. Define the edges of the graph in a dictionary format
3. Create a directed graph using the edges dictionary with networkx.DiGraph()
4. Iterate over the edges dictionary to add edges and weights to the graph using G.add\_edge(source, target, weight=weight)
5. Calculate the positions of the nodes using nx.spring\_layout()
6. Define the edge labels using a dictionary comprehension
7. Draw the graph using nx.draw()
8. Draw the edge labels using nx.draw\_networkx\_edge\_labels()
9. Turn off the axis using plt.axis('off')
10. Show the plot using plt.show()

# 3) Research Results:

**3.1 Final LDA output**: This LDA (Latent Dirichlet Allocation) model was trained on a dataset of video titles in Korean language, and it has generated 20 distinct topics. Each topic is represented by a list of words and their corresponding weights. The weights indicate the importance of each word within the topic. In this output, the topics and their top words are as follows:

테이블이(가) 표시된 사진

자동 생성된 설명

달력이(가) 표시된 사진

자동 생성된 설명

**3.2 Network Edge and Weight output:** The output above represents the relationships between different videos based on user comments. In this network, each video is represented as a node, and the relationships between the videos are represented as edges. The weight of each edge indicates the number of users who have commented on both videos, making the edge's weight a measure of the strength of the relationship between the two videos.In the output, each line shows the relationship between two videos, identified by their video IDs

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

3.3 Network Visualization output:

지도이(가) 표시된 사진

자동 생성된 설명

# 4)Conclusion:

This research project aimed to analyze public engagement with YouTube news channels by cleaning and preprocessing the YouTube dataset, performing topic modeling using news video titles and scripts, and constructing and analyzing the social network of news channels and subscribers. By examining these factors, the study provided insights into audience preferences, interaction patterns, and relationships between news channels and subscribers. The use of Latent Dirichlet Allocation (LDA) for topic modeling enabled the identification of the most prominent topics in the dataset, shedding light on the audience's preferences and interests. The network edge and weight script generated a video interaction network based on user comments on videos, while the network visualization script visualized the weighted directed graph using the NetworkX library and Matplotlib for rendering. Overall, the findings of this research project can guide stakeholders in understanding public engagement dynamics and shaping their strategies accordingly.