Japanese Text Analysis
with NHK (Japan Broadcasting Corporation) Show Information

By: Miho Hunter

ABSTRACT:

The goal of this project is to analyze NHK's show information using Natural Language Processing (NLP) and topic modeling techniques including NMF, LSA, and LDA. K-Means was used for clustering the shows into different genres. Finally, I repeated the classification using BERT.

DESIGN:

A model that understands the Japanese Texts and helps extract insights from the data can potentially be used for many applications such as churn rate prediction, customer sentiment analysis, customer segmentation, etc. To come up with such a model, I have used the show information from NHK. The general pipeline of the process is: data collection, text pre-processing, topic modeling, and clustering.

DATA:

The dataset was obtained using NHK's program list API which allows us to get show time, title, subtitle, and content for TV, radio and internet radio scheduled for the next 7 days. I have collected information about 9,695 unique shows that aired (or to be aired) between 6/27/22 and 7/17/22.

ALGORITHMS:

Text Pre-processing:

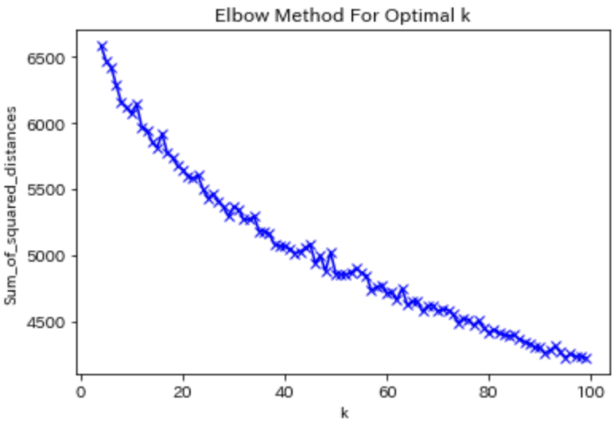
I used neologdn (Japanese text normalizer) to normalize text, replaced numbers with 0, removed alphabets, and identified stop words. Using MeCab, I tokenized and extracted nouns. Vectorization was done using Tfidfvectorizer.

Topic Modeling:

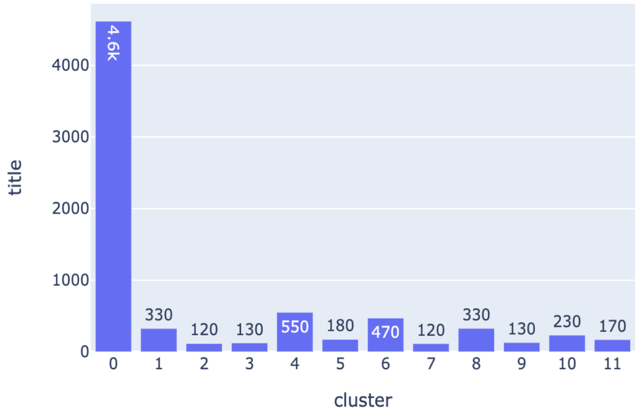
NMF, LSA, and LDA models were created and compared. NMF and LSA showed similar results. LDA did not identify some of the topics as well as the other two models such as "politics" and "kids".

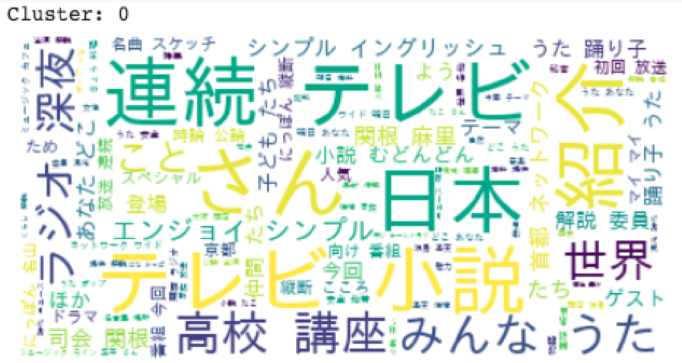
K-Means Clustering:

Elbor method was used to identify optimal k.



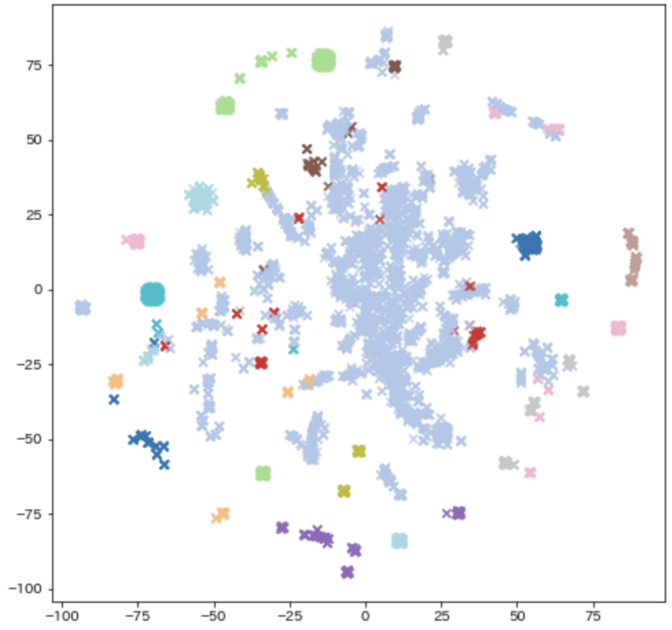
For K=12, most shows were classified as one category.





The model successfully identified genres like “education/kids” and “politics”.

Finally, I used t-sne to visualize the clusters. We can see some clusters are mostly on their own, while some are intermingled with other clusters.



BERT:

Lastly I tried a BERT model to see how much better it can do with the classification of the shows. I divided the data into training (80%) and test (20%). I used one of BERT's pre-trained models, then it was fine-tuned with the BertForSequenceClassification model. *bert-base-japanese-v2* (trained using 30M Japanese sentences from Wikipedia by researchers at Tohoku University) was fetched. The model was trained across 5 epochs. The highest genre probability was compared with the true value (the show data had genre information) to calculate accuracy. Genre prediction accuracy was 92% for the final model. The training was limited to 5 epochs due to time constraints, but the accuracy would go up if I increase the number of epochs.

Note: The code I used was taken from the article "Multi Class Text Classification With Deep Learning Using BERT" by Susan Li at https://towardsdatascience.com/multi-class-text-classification-with-deep-learning-using-bert-b59ca2f5c613

Genre	Accurately Predicted	In Test Dataset	Accuracy
0. Documentary	175	202	**87%**
1. Welfare	0	12	**0%**
2. Music	131	141	**93%**
3. News	446	447	**100%**
4. Drama	45	49	**92%**
5. Information	85	88	**97%**
6. Comedy	29	45	**64%**
7. Sports	47	59	**80%**
8. Hobby/Education	375	386	**97%**
9. Theater	0	5	**0%**
10. Anime	25	31	**81%**
11. Other	23	23	**100%**
12. Movie	0	3	**0%**

TOOLS:

NHK's program list API for data collection, Mecab for Japanese text segmentation, Neologdn for Japanese text normalization, sklearn for decomposition and topic modeling, Matplotlib, Plotly, and Word Cloud for visualization.

CONCLUSIONS:

As my first project working with NLP techniques, I was happy with being able to get some reasonable topics and genres from my dataset, such as politics, music, travel, and kids using TF-IDF and topic modeling techniques although the majority were classified as one “general” category. It was interesting to see some clusters are mostly on their own, while some are intermingled with other clusters from t-SNE visualization. In comparison, BERT is very powerful, but it is a very heavy model. There is a tradeoff between accuracy and speed.