Japanese Text Analysis

Classifying NHK(Japan Broadcasting Corporation) shows

Miho Hunter
7/13/22

TOPIC OVERVIEW

Wouldn't it be useful to have a model that would extract insights from data in foreign language?

Let's build one for 日本語 (Japanese)!

TABLE OF CONTENTS

1. Data
2. Process and Tools
3. Model
4. Results

DATA

NHK's show information

NHK: Japan Broadcasting Company

9695 unique shows

aired between 6/27/22 and 7/17/22

9859 tokenized words

Process and Tools

Gather show info

-NHK's show info API

Topic modeling Clustering

NMF, LSA, LDA
K-Means Clustering

Text Pre-processing

Normalize text (Neologdn),
Tokenize and extract nouns with MeCab, vectorize with Tfidfvectorizer

BERT

Hugging Face library,
bert-base-japanese-v2
(pre-trained model),
BertForSequenceClassification
model for fine tuning

Top 100 Words

Topic Modeling 1: NMF

Topic 0	ニュース, 日本, きょう, 関東甲信越, 中国, スペイン	**News**
Topic 1	天気, 予報, 交通, 情報, 全国, 海上	**National News**
Topic 2	気象, 情報, 通報, 交通, 関東甲信越, カラフル	**Local News**
Topic 3	英語, 基礎, レベル, 中学生, 中高生, 前年度	**English**
Topic 4	選挙, 選出, 議員, 参議院, 政見, 一部	**Politics**
Topic 5	国内外, 内容, 変更, スポーツ, 中継, 時間	**Sport News**
Topic 6	教授, 英会話, 講師, 大学, ジェニー, スキッドモア	**English Conversation**
Topic 7	うた, 踊り子, 地帯, 安全, しな, ポップ	**Songs, Music**
Topic 8	音楽, カフェ, ミュージック, 話題, 世界, リクエスト	**Music**
Topic 9	講師, 非常勤, 中国語, ハングル, 講座, キム	**Asian Languages**
Topic 10	様々, 届け, 時間, 情報, 全国, 解説	**Information**
Topic 11	イングリッシュ, エンジョイ, 関根, 麻里, シンプル, 司会	**English Lesson**
Topic 12	体操, 多胡, 歳児, 対象, エンターテインメント, 教育	**Kids**

Topic Modeling 2: LSA

Topic 0	ニュース, 情報, 気象, 予報, 天気, 交通	**News**
Topic 1	天気, 予報, 情報, 交通, 気象, 全国	**National News**
Topic 2	気象, 情報, 時間, 国内外, スポーツ, 内容	**Sports News**
Topic 3	英語, 基礎, レベル, 中学生, 中高生, 講師	**English**
Topic 4	選挙, 選出, 議員, 参議院, 政見, 一部	**Politics**
Topic 5	気象, 情報, 関東甲信越, 通報, 講師, 関東	**Local News**
Topic 6	講師, 教授, 英会話, 大学, ジェニー, ソレイシィ	**English Conversation**
Topic 7	うた, 踊り子, 地帯, 安全, しな, ポップ	**Songs, Music**
Topic 8	音楽, 届け, 様々, カフェ, 時間, ミュージック	**Music**
Topic 9	講師, 非常勤, ハングル, 中国語, キム, ウナ	**Asian Languages**
Topic 10	様々, 届け, 時間, 情報, 全国, 首都	**Information**
Topic 11	イングリッシュ, エンジョイ, 関根, 麻里, シンプル, 司会	**English Lesson**
Topic 12	体操, 多胡, 歳児, 対象, エンターテインメント, 教育	**Kids**

K-Means Clustering

60% in this “GENERAL” category

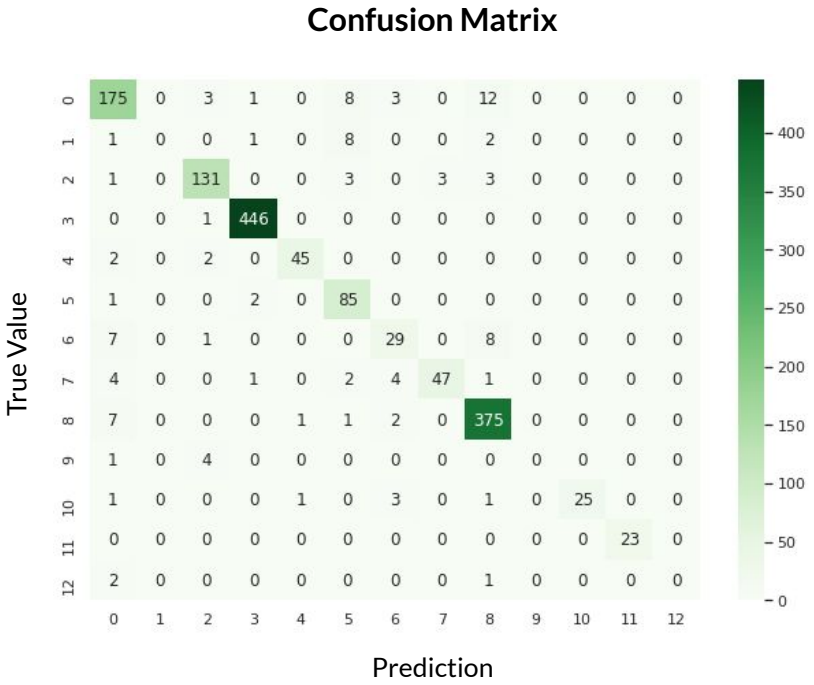
Now let's predict show's genre using

BERT State of the art language model for NLP!

- Divide data into training (80%) and test (20%)
- Train the model across 5 epochs and use one with the highest accuracy
- Use one of BERT's pre-trained models, then fine tune with *BertForSequenceClassification* model
- Fetch *bert-base-japanese-v2* (trained using 30M Japanese sentences from Wikipedia by researchers at Tohoku University)
- Compare highest probability genre with the true value

Results Genre prediction with 92% accuracy!

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%**



CONCLUSION

TF-IDF, topic modeling techniques can be applied to Japanese texts.

Using K-Means, I was able to cluster TV shows into different genres.

BERT is very powerful! But it's a very heavy model.

Speed - Accuracy tradeoff