Report

The problem was to match and extract movie titles from the user survey responses with given contents (titles).

As the problem definition is unique, there is no particular method available that we can use right way, it isn’t a typical keyword extraction or named entity recognition problem.

My approach was to solve it using simple word matching algorithm and still achieve the best results, steps as follows,

1. First the user response is subjected to text cleaning process using NLTK and regex libraries which basically involves emoji removals, special character removals and finally stop words removal. Note, some of the stop words are to be excluded from the NLTK’s standard stop word list before removing stop words. To be particular in the sample content “Ma” was one the title which is removed when applied with standard stop list. So, I excluded them manually before processing the text.
2. Different combinations of words in sequence are created and then brute-forcedly matched with all the content titles using “levenshtein**”** distance measure from python library called Levenshtein. The Levenshtein distance is a measure that tells us how different two strings are. The higher the number, the more different the two strings are. The main reason behind choosing this algorithm was its similarity score consistency for small to lengthy sequence of text. It performed very well even when the length of the strings were short.
3. Thresholds are set in matching with appropriate title which dynamically adjusts as per the length of the sequence. If there isn’t any absolute matches with the title, the near most appropriate or approximately matched title is returned with again threshold set which is a constant one.

This novel approach has passed all the test cases i.e., was able to match and retrieve all the titles from user responses. Please find the output table in the page 2.

The code, data and outputs are uploaded to the git repository.

Link to the repository - https://github.com/kriskrishnaa/title_extractor.git

Table 1: Output

Response	Content_names
Fear the walking dead,Supernatural (huge fan and sad it has finished),The Gentlemen, Outlander	{'Fear the Walking Dead': 100.0, 'Supernatural': 100.0, 'The Gentlemen': 100.0, 'Outlander': 100.0}
A lot! -good doctor -gangs of London - the gentleman -ma -spies in disguise	{'The Gentlemen': 88.88888888888889, 'The Good Doctor': 100.0, 'Gangs of London': 100.0, 'Ma': 100.0, 'Spies in Disguise': 100.0}
Miss scarlet and the duke,knifes out,Dublin murders	{'Miss Scarlet and the Duke': 100.0, 'Dublin Murders': 100.0, 'Knives Out': 83.33333333333333}
History drama-Vikings,Kid friendly-Casper,Sometimes the conversations while watching Neon can get serious but we all end up having fun together, :)	{'Vikings': 100.0, 'Casper': 100.0}
The Undoing,Game of thrones,Outlander, Vikings,CB Strike (and most all British dramas) Westworld	{'Outlander': 100.0, 'Vikings': 100.0, 'The Undoing': 100.0, 'Game of Thrones': 100.0, 'C.B. Strike': 90.0, 'Westworld': 100.0}