Nick Videtti

nvidetti@syr.edu

Assignment 7

Naive Bayes and Support Vector Machines on Corpora

SciKit Learn’s Document Vectorizers, Naive Bayes, Support Vector Machines, and Confusion Matrices

6/1/2023

| Which Streaming Service to Choose? | |
| --- | --- |
| **Introduction** | Not too long ago, live television was a very popular phenomenon and was what many considered the best way to watch television. Whether it be an awards show, a game show, American Idol, Fear Factor, or any other major program that made up live television in the 2000’s, there was something about the fact that it was being watched in real-time that made it seem a little more special than the usual reruns of Seinfeld and The Simpsons. Families would rush through their to-do lists to make sure they didn’t “miss their show” and often tensions would arise if they did miss it. While sports fans still go through this phenomenon to this day, the general appeal of live television has severely worn off over the years to the average person. It is actually arguable that live television has become an undesirable option to most people.    How did it get to this point? The simple answer is that people realized that streaming movies and shows is a much more convenient option. Anything that is new is likely to be a better version of what was available in the past, which is exactly what happened when streaming became an option instead of live television. While live television is still as prominent as ever, most live shows are also available to stream just a few hours later, if not immediately. Many people will either opt for the latter, or will even hold off on watching their shows until all episodes in a season are available to “binge” during just a couple sittings, similar to the hungry uncle at Thanksgiving dinner that can’t stop talking about how he skipped breakfast and lunch in preparation.    Streaming services have now become the main way to watch television for the majority of people, which has resulted in an overwhelming overflow of content. Streaming services do not need to follow a schedule the way that a channel on TV does, which means they are not held back to time blocks, they can give many more options over the course of a day than TV, they can offer all options throughout the day, and they are not pressured to only release one new episode of a series per week. Even while only considering one of the roughly half-dozen major streaming services, the amount of content is breathtaking.    There are a few options for the modern-day television watcher, but the two extremes will be considered. One extreme is to subscribe to all major streaming services to eliminate the chance of missing out on any content, but that ends up costing almost as much as a cable bill that gives hundreds of channels (which is appealing despite the aforementioned limitations with television channels). The other extreme is to simply pick a single streaming service to subscribe to, which saves on costs and will still provide enough content to watch continuously for the rest of the time that the service is in existence. The only question for this option is which streaming service is the one to pick?    Considering the single-service method, the question of which to pick will be investigated. In this study, movie reviews from an arbitrary streaming service will be inspected. These reviews are made up of both positive and negative reviews, and will be used to identify some keywords that will assist in searching for the best and/or the worst streaming service. While this will not identify a single service to pick, it will give some good information on finding which streaming service is likely to have the best content and which has the worst content. |
| **Analysis** | The reviews that will be used come in the form of two corpus files of documents, one for positive reviews and one for negative reviews. Each file has roughly 12,000 documents, but only the first 100 will be used in these analyses due to high latency while performing these analyses with a larger number of documents.  The documents will first need to be read into a Python session. The user is prompted to select the file for the negative corpus and then for the positive corpus. For the corpora files, each document in the corpus is looped through and read into a list. Each element in the list is a list itself, containing the document and “+”/”-” for positive/negative reviews. These lists will be used as the data source for a pandas DataFrame containing the columns “SENTIMENT” and “CORPUS”, where “SENTIMENT” is either “+” or “-” and “CORPUS” is the document/movie review.  Next, data cleaning will be needed. This will involve three steps, each of which will result in another column being added to the dataframe. First, the “CORPUS” column will be split into a list of its words using NLTK’s word tokenizer and the result will be stored in a new column called “WORDS”. The following step will take the “WORDS” column and clean/filter them by converting to lowercase, removing line break characters, removing NLTK’s English stop words, removing phrases that contain numbers, and removing phrases that do not contain letters. This result will be stored in another new column called “CLEANED\_WORDS”. The final step will be to combine the lists in “CLEANED\_WORDS” into a single string and to store that result in a new column called “CLEANED\_CORPUS”.  Once the data are prepped for the analyses, a function will be created in order to reduce code redundancy. This function will be applied twice, once for unigrams and once for bigrams. This function will create four different vectorizers and perform a cross validation on models using all of the applicable vectorizers on 5 different algorithms. The four vectorizers are count, count binary, TFIDF, and TFIDF binary. The five algorithms will be Multinomial Naive Bayes, Bernoulli Naive Bayes, Support Vector Machine with a polynomial kernel, Support Vector Machine with a RBF kernel, and Support Vector Machine with a linear kernel. Confusion matrices and three other metrics (minimum accuracy, maximum accuracy, and average accuracy) will be generated for each of these vectorizer/algorithm combinations, resulting in 18 confusion matrices and 54 metrics. All of this will be printed to the user, as well as the most accurate model(s), based on average accuracy.  After this function is created and applied to the data for both unigrams and bigrams, the top 10 most positive unigrams/bigrams and top 10 most negative unigrams/bigrams will be found. These will be found using a count vectorizer that will find which unigrams/bigrams occur most often in each sentiment classification. There is likely a much better method where the most accurate model can be applied, but this was a good first step. |
| **results** | The results of the analyses were as follows.  **Cleaned DataFrame - First Few Rows**      **Unigrams - Multinomial Naive Bayes**    **Unigrams - Bernoulli Naive Bayes**    **Unigrams - Support Vector Machine (Polynomial Kernel)**    **Unigrams - Support Vector Machine (RBF Kernel)**    **Unigrams - Support Vector Machine (Linear Kernel)**    **Unigrams - Accuracy DataFrame**      **Unigrams - Best Model(s)**    **Bigrams - Multinomial Naive Bayes**    **Bigrams - Bernoulli Naive Bayes**    **Bigrams - Support Vector Machine (Polynomial Kernel)**    **Bigrams - Support Vector Machine (RBF Kernel)**    **Bigrams - Support Vector Machine (Linear Kernel)**    **Bigrams - Accuracy DataFrame**      **Bigrams - Best Model**    **Unigrams - Top 10 Most Positive and Top 10 Most Negative**    **Bigrams - Top 10 Most Positive and Top 10 Most Negative**    There are obviously some significant caveats to the results of these analyses, none more relevant than the fact that only about 1% of the available data are used in this study (200 of roughly 24,000 reviews). Another issue is the method of how the top 10 positive and negative features are found. It is likely a better idea to somehow incorporate either the best model or determine the best vectorizer to use, and it is seen that raw count vectorizer was not the best vectorizer for unigrams or bigrams. In a future study, both of these issues would look to be addressed to find some more meaningful and accurate results. |
| **conclusions** | There were 200 reviews inspected, 100 positive and 100 negative. Based on those, the following was determined.   |  | Best Streaming Service | Worst Streaming Service | | --- | --- | --- | | What to Search For | movie  film  one  story  titanic  time  like  good  see  well  james cameron  robin williams  love story  king masks  night listener  toni collette  leonardo dicaprio  special effects  street performer  good movie | movie  film  one  like  good  even  would  really  bad  story  akshay kumar  anil kapoor  kareena kapoor  action scenes  krishna acharya  ali khan  vijay krishna  saif ali  justin timberlake  yash raj |   It seems that searching the actors and directors of positively/negatively reviewed movies is a good suggestion based on the small amount of reviews that were inspected. This is actually pretty useful given the actors of films on streaming services is something that a prospective customer can easily find. A future study would likely include a much larger number of reviews and give even better suggestions than these, but until then, a good option is to use these searches to determine the best streaming service. |