Movie Reviews: A First Vectorized Representation

W. Andy Holst

School of Professional Studies, Northwestern University

MSDS 453: Natural Language Processing

Dr. Syamala Srinivasan

July 11, 2021

Movie Reviews: A First Vectorized Representation

With the advent of computers in the mid twentieth century, the idea of using these machines to "perform useful tasks involving human language" was simultaneously conceived (Jurafsky & Martin, 2014 p. 1). Two almost uniquely human traits are our ability to communicate information and knowledge to our fellow humans (in written and spoken form), as well as to use tools and technology to enhance our own brains beyond the singular. It is through all of the different human languages that exist in today's world that our communication occurs. So, it naturally follows that teaching computers how to communicate and perform useful tasks as a tool is a desirable outcome. That, in a nutshell, is what Natural Language Processing is all about.

This first paper in Natural Language Processing (NLP) begins with the very basic idea of identifying common terms, themes and ideas within a "corpus", or collection of documents. It is important that there is a common corpus vocabulary so that information and knowledge are effectively communicated. With this goal in mind, we embark on analyzing a corpus (200 documents) of movie reviews that have been collected. Our end goal and final application of analyzing this corpus, while not yet entirely clear, reveals some of the challenges and tools that result. With the work described in this paper, we have taken the first step towards better NLP by identifying terms which are good candidates for a useful corpus vocabulary.

Data

The data utilized in this research consist of 200 movie reviews that have been collected and compiled by students in this NLP course. Each movie review (the "text") is at least 500 words long and includes the following metadata associated with each text:

- **Document ID** (numbering 1 to 200)
- **Unique Document Name**

- **Student Name** (20 students total, each student contributed 10 movie reviews)
- **Movie Title Name** (20 different movies included, based upon the above distribution)
- **Genre of Movie** (4 genres were selected, and movies are equally distributed across each genre: Action, Comedy, Horror, and Sci-Fi)
- **Review Type** (positive or negative, each student contributed 5 of each, so there are equal numbers of positive reviews and negative reviews)

The data preparation, exploratory data analysis and visualization steps undertaken to prepare the dataset for subsequent research, analysis, and modeling were as follows:

- A Google Colaboratory Notebook (Python 3.7.10) was created.
- Python Packages imported as necessary for pre-processing and modeling, including: Numpy (1.19.5), Pandas (1.1.5), Sci-kit Learn (0.22.2), Matplolib (3.2.2), NLTK (3.2.5), Gensim (3.6.0), and Re (2.2.1).
- Much of the exploratory data analysis involves printing strings of characters from each text. These strings are compiled into lists and dictionaries for review and analysis.

As a very first aspect of my analysis, I noticed that some documents had included strings and words which had experienced some type of concatenation during import. Here is an example of the concatenations observed, a listing of the longest words found in Text #75:

```
['whygravityfound',
 'astrophysicists',
 'withinterstellar',
 'noncondescending',
 'ofinterstellaris',
 'ininterstellaris',
 'callinterstellara',
 'ininterstellarbegin',
 'boyhoodandmoonlight']
```

When I investigated further, I realized that multiple encodings were present in the original texts. These different encodings included Western European (Mac) and Western European (Windows), to Unicode (UTF-8). So, as a pre-processing step, I manually converted each text from the corpus into a Unicode (UTF-8) format. After correcting for this encoding issue, a corrected listing of the longest words found in Text #75 is now as follows, much improved:

```
['interstellar',
 'performances',
 'overpopulated',
 'gravitational',
 'heartbreaking',
 'heartbreaking',
 'introspective',
 'sentimentality',
 'Astrophysicists']
```

Research Design and Modeling Method(s)

In Step 1, the goal is to look for terms that would be prevalent across documents. I chose to first implement a series of very simple data cleaning steps applied to each text in the corpus, which are generally described below:

1. Separate each document into individual tokens, using built-in `.split()` function.
2. (*Added later*) Apply lemmatization to tokens using `WordNetLemmatizer()`
3. Convert tokens into lower case using built-in `.split()` function.
4. Remove punctuation using regex `re` function `.escape(string.punctuation)`.
5. Remove non-alphabetic tokens using built-in `.isalpha()` function.
6. Remove stop words using NLTK `stopwords.words('english')`.
7. Remove single letter words.
8. (*Added later*) Remove custom stop words based on corpus

Following the text cleaning described above, I then created a Term-Document Matrix with Frequencies. This initial TDM Frequency Matrix produced 17,125 words in my vocabulary, with the twenty most common words (note this was prior to adding lemmatization):

```
[('film',616),('one',542),('movie',477),('like',428),('time',376),
('even',255),('also',242),('much',233),('story',227),('us',227)]]
```

After removing several more custom “stop words” that were custom for this particular vocabulary, the updated vocabulary now consists of 14,597 words. I nominate my very first list of prevalent terms that are present in this movie review corpus:

```
[('film',766),('movie',671),('time',399),('make',257),
('character',256),('even',253),('story',249),('way',246),
('bond',246),('scene',183),('action', 174)]
```

So the themes that emerge are both inherently obvious, i.e. these are all documents that are about films and movies, and they involve characters and stories. So of course these are prevalent terms in the corpus. But, there are other concepts starting to emerge, such as “make”, “time” and “even” and “way”. “Even” is a peculiar, unexpected word but maybe it will one day make more sense than it does right now. The first genre association has emerged, with the word “action” showing up as a prevalent term. And one final conclusion that I might draw from this list is that “Bond” movie reviewers LOVE to use the word “Bond” in their reviews!

With the TDM Frequency Matrix providing a small glimpse into what the corpus themes are, I turn next to the Term-Document Matrix using TF-IDF as a methodology. This is more

challenging to explore, since there are 200 different TF-IDF lists generated, one for each document. So comparing a TF-IDF list seemed to me to make the most sense as follows:

1. Prepare TF-IDF listings for all 200 documents
2. Take the fifty (50) top TF-IDF candidates from each document
3. Compile these 200*50 = 10,000 top words into a Bag-of-Words
4. Look at the Top Ten of the Bag-of-Words List.

Using this methodology, I nominate my second list of prevalent terms that are present in this movie review corpus:

```
[('film',57),('movie',49),('horror',23),('family',22),('action',22),
('alien',19),('character',19),('cruise',18),('planet',17),('time',17)]
```

So, “film”, “movie” and “character” are still very prevalent terms. From the TF-IDF perspective, the themes of the movies are now starting to emerge, with “horror” and “action” being the most direct terms to their genre. One could posit that “family” is common to comedy movies (probably less so in horror!). And sci-fi films likely center around “aliens”, “planets” and “time.” And one final conclusion that I might draw from this list is that “Cruise” movie reviewers LOVE to use the word “Cruise” in their reviews! It may be worth exploring this more, but the TDM TF-IDF approach does seem to better elicit the themes across the genres better than TDM Frequency approach, likely because the “tamping” down of very prevalent terms across all documents that just drown out the true themes of the reviews. I would suggest exploring this further in the next few assignments.

Turning to the Word2Vec and Doc2Vec methods, I found these methods to be less intuitive and more difficult to elicit corpus-wide insight into prevalent terms. The fact that

Word2Vec fundamentally functions best upon ***huge corpa*** on the order of billions and billions of words may explain this challenge with our corpus (<100,000 words). But instead what I found to be really interesting and helpful was the ability of Word2Vec to identify similar terms. Here are a few of the interesting relationships that Word2Vec was able to determine from the movie review corpus:

[Input]	`w1 = 'wick'` `model_w2v.wv.most_similar(positive=w1)`

[Output]	`[('john', 0.9999902844429016),`
	` ('reef', 0.9999902248382568),`
	` ('new', 0.9999886155128479),`
	` ('action', 0.9999885559082031),`
	` ('film', 0.9999884366989136),`
	` ('character', 0.9999884366989136),`
	` ('he', 0.9999883770942688),`
	` ('man', 0.999988317489624),`
	` ('star', 0.9999881982803345),`
	` ('even', 0.9999881386756897)]`

[Input]	`w1 = 'emma'` `model_w2v.wv.most_similar(positive=w1)`

	`[('cruella', 0.9999895095825195),`
	` ('film', 0.9999880194664001),`
	` ('even', 0.9999877214431763),`
	` ('new', 0.9999876022338867),`
	` ('character', 0.9999874234199524),`
	` ('make', 0.9999873638153076),`
	` ('time', 0.9999872446060181),`
	` ('bond', 0.9999871850013733),`
	` ('world', 0.9999871850013733),`
	` ('story', 0.9999870657920837)]`

Results

As described above, I have nominating my official list of the most prevalent terms present in this movie review corpus, choosing the TDM TF-IDF Methodology:

```
[('film',57),('movie',49),('horror',23),('family',22),('action',22),
('alien',19),('character',19),('cruise',18),('planet',17),('time',17)]
```

Analysis and Interpretation

From the TF-IDF perspective, the themes of the movies are now starting to emerge, with “horror” and “action” being the most direct terms to their genre. One could posit that “family” is common to comedy movies (probably less so in horror!). And sci-fi films likely center around “aliens”, “planets” and “time.” It is worthy of exploring more the how and why the TDM TF-IDF approach does seem to elicit the themes across the genres better than TDM Frequency approach. Perhaps something related to the “tamping” down of very prevalent terms across all documents which drown out the true themes of the reviews. I would suggest exploring this further in the next few assignments.

From a Word2Vec perspective, these methods appear to be less intuitive and more difficult to elicit corpus-wide prevalent terms than the TDM TF-IDF methodology, as specifically applies to our corpus. The fact that Word2Vec fundamentally functions best upon ***huge corpa*** on the order of billions and billions of words may explain this challenge with our corpus (<100,000 words). There are certainly many facets to explore in future research papers to understand Word2Vec better, and I look forward to presenting results at a later date.

Conclusions

We began this assessment with the very basic idea of identifying common terms, themes and ideas within a “corpus”, or collection of documents. It is important that there is a common corpus vocabulary so that information and knowledge are effectively communicated. With this goal in mind, we analyzed a corpus (200 documents) of movie reviews that have been collected. Our end goal and final application of analyzing this corpus, while not yet entirely clear, reveals some of the challenges and tools that result. With the work described in this paper, we have taken the first step towards better NLP by identifying terms which are good candidates for a useful corpus vocabulary.

References

Jurafsky, D., & Martin, J. H. (2014). Chapter 1: Introduction. In *Speech and language processing* (Second Edition, pp. 1–16). essay, Pearson Prentice Hall.

Lane, H., Hapke, H. M., & Howard, C. (2019). *Natural language processing in action: understanding, analyzing, and generating text with Python*. Manning Publications Co.