Genre Discovery: Using Naive Bayes Modeling to Predict Genre

Modelling language is a large field with many applications to understanding not only how people are writing, but what they are writing about. It can be used to analyze the culture, mythos or ideas that are common in a group of people. Modelling language can also be used to analyze how we write, with the labeling of text presenting a major problem to companies today. While this analysis is a fun application of natural language processing, it represents some of the larger challenges common in labeling data. As more and more text data is produced in places like social media, and in journalism, the need for these types of analyses are more important.

Genre has been a way of sorting various forms of art since Ancient Greece. It can describe form (such as poetry, or biography), audience (for example young adult or children's literature), or most abstractly, a set of cultural expectations of a body of works. For example, science fiction is a genre well known for focusing on potential what ifs of science. If could be the power of science in a police state, such as in George Orwell's 1984. It could be, and more commonly is, space travel, in works like Star Trek. These cultural norms can also be part of the form of the genre, with genres like mystery that are based on a common plot point (in this case solving a mystery).

Thus, genre presents a unique classification problem. Not only is genre often poorly defined, it is often mixed together within a work in order to create a new sub genre. For instance, the Harry Potter series is fantasy, for it's magical elements, but it is also children's literature, as the main characters are all children for a majority of the books. This amalgamation of genre creates interesting challenges to machine learning, as a single set of features can represent many different labels.

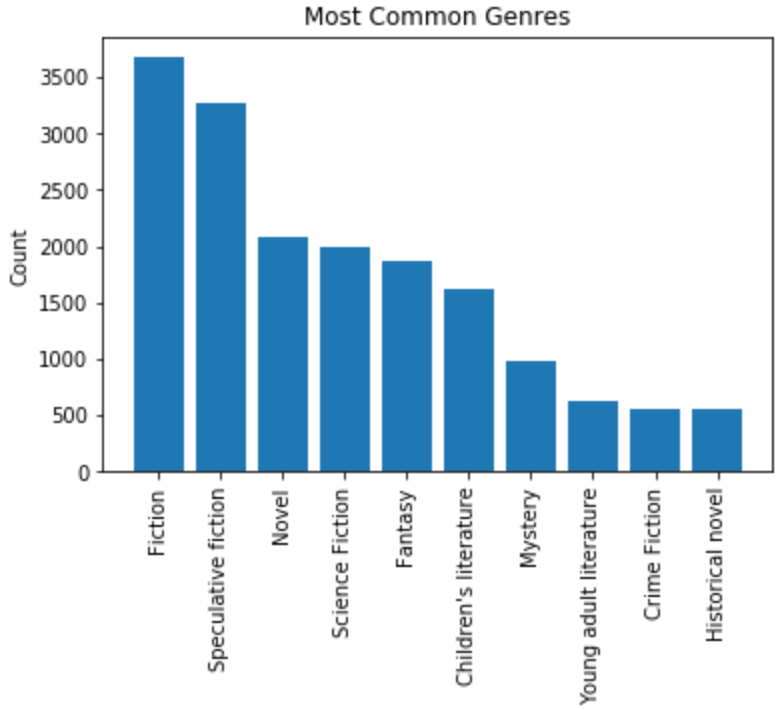
Initial Data Processing and Analysis

The dataset, originally from the work of David Bamman and Noah Smith at Carnegie Melon University, pulled 16,559 summaries of books from Wikipedia, along with the author, title and Wikipedia summary. It also includes a list of genres, pulled from Freebase, for each book.

Initially, these genres were cleaned to be readable, and then investigated in two categories. The first are genres of form, such as memoirs, biographies and novels. These are forms that are more about how the work is written instead of the content. These genres can be used to sort a subset of the data into fiction and non-fiction based on the forms. The second are genres of content, such as science fiction, fantasy, or horror. These genres were broadened to include works that fell into sub categories in order to more broadly categorize the works. For example,

sword and planet is a subgenre of science fiction, so works labelled as sword and planet were tagged into science fiction in order to provide larger genre categories.

Initially, all of the genre labels were considered, regardless of genre type.



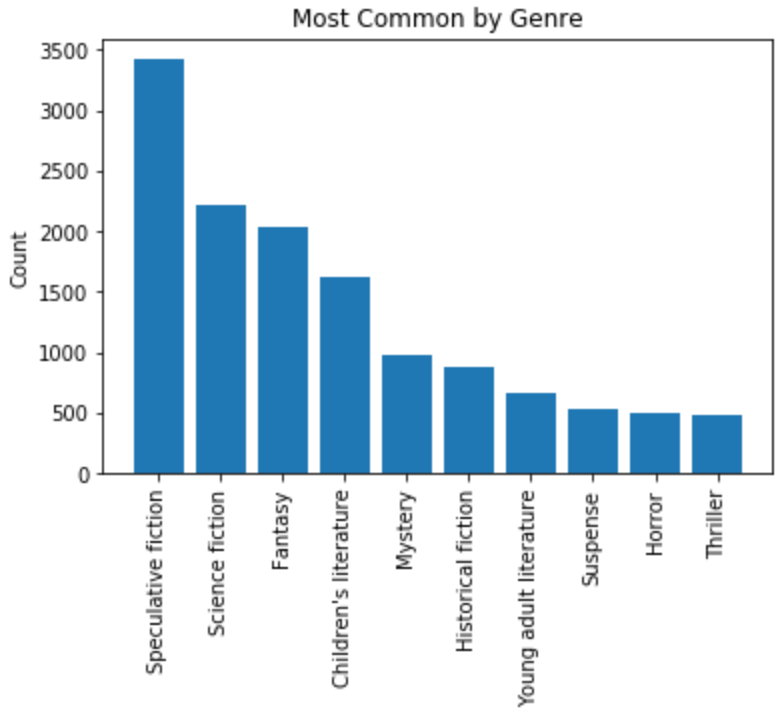
Above is the breakdown of the most common genres at the macro level across the entire cleaned dataset. Both fiction and novel were in the top 10, which are synonymous, and are genres of form. The rest are genres of content, with common genres falling out in an unsurprising order. Speculative fiction, one of the hardest genres to easily define, sits at second largest. As this genre represents any novel that has differences from the real world, it overlaps with many of the other genres, and thus is very common in the labelling. Science fiction and fantasy are also high, as some of the larger, more well known content based genres. The rest of the top ten rounds out with more reality based genres, like mystery, and historical fiction.

The top two genres have over 3000 tags a piece, and represent more works than the 6 in the bottom of the top ten combined. The drop off continues into other genres, which may get as low as one tag in the entire dataset.

Genres of Form:

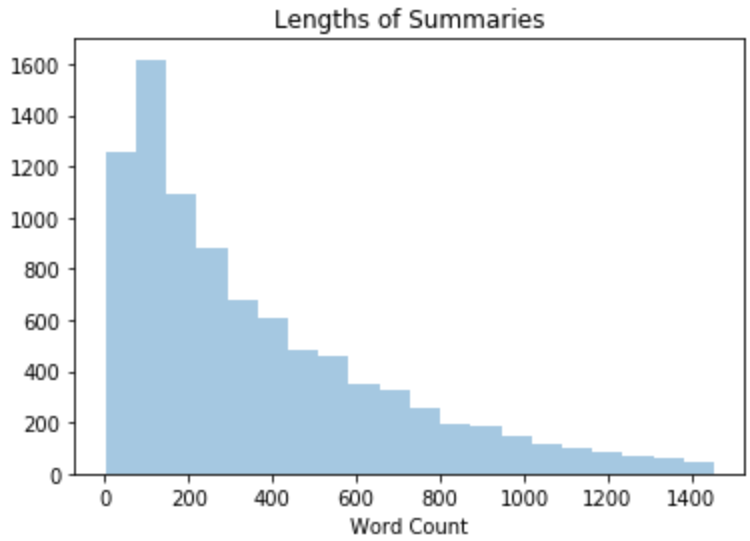
Once the fiction and non fiction labels of form are combined, they occasionally cause problems as a single summary may be tagged by forms that fall into the two different categories. After accounting for these rare cases, the summaries were labeled, and it was found that there are far more fiction summaries than nonfiction. These discrepancies is size of groups are accounted for later in the analysis.

Genres of Content:



After the initial plot, subgenres were combined with their parent genres, and the genres of form were removed, producing the plot above. Notice that 3 new genres have joined the plot: suspense, horror and thriller. These 10 genres were the final genres chosen to be modelled on, as they represented the majority of the dataset, and all had large enough sets to be able to model on.

An Aside on Word Count:



9000 of the 16,559 summaries were plotted by word count in order to get an idea of the average length of summaries on Wikipedia. While there is a major spike between 100 and 200 words, the average value is 419 words, as the tail of the data is significant. (ADD INFO HERE ABOUT THE TWO EXTREMAS CASES: LOW AND HIGH)

Models and Predictive Words:

The Models:

In order to appropriately model the data for each genre, it was cleaned in a series of steps. After correctly interpreting the genres and separating them out from each other (as done for the initial analysis), the summaries were cleaned to remove proper nouns and punctuation. Proper nouns were removed in order to prevent common character names, especially those from series, or common names in general, from affecting the model. These values were then vectorized using both a count vectorizer and Tfidf, in order to assess which would be better for model performance. In all but two of the cases, the count vectorizer was either equivalent or better.

Before creating the vectorizer and model, the sets were reduced to create parity between the labels for each of the categories. So, for example, in a smaller set like mystery, which only had around 500 cases, the non label was randomly chosen to be of the same size, so that the much larger non-mystery set would not rule the model. This was applied to all of the genre models, with the exception of speculative fiction, which was almost identical in size to it's non-labelled counterpart.

A final model, accessible online at sgunners.pythonanywhere.com, all of the models into one hierarchical model. This model first cleans new text, and then puts it through the fiction vs. nonfiction model. If the models predicted the text as fictional, if it then fed to the second tier of

genre to predict which subgenre is the more likely. The words that indicated the most likely genre are placed in a word cloud, with their size representing their relative impact on the genre choice.

Predictive Words:

Predicting Fiction vs. Nonfiction:

With a dataset labelled based on the forms of fiction versus non-fiction, a Naive Bayes model was applied to two different vectorizers to try and predict on text, as well as find the most predictive words in the data sets. The final result has an accuracy on the training set of 80.7% and 78.8% on the testing set.

```
Non Fiction Words
              argues 0.93
              theory 0.91
             chapter 0.88
             example 0.86
            describes 0.82
              author 0.81
             subject 0.80
               issue 0.80
          individual 0.79
               spent 0.79

Fiction Words
                ship 0.10
             reveals 0.19
           discovers 0.20
          mysterious 0.21
              refuse 0.22
              battle 0.22
                kill 0.22
          apparently 0.23
              agrees 0.23
              rescue 0.23
```

The above are the top 10 non-fiction words and top 10 fiction words, with the probability of the words being associated with the non-fiction tag. The fiction words all trend towards language about adventures and journeys, with things like rescue, reveals, and mysterious in the indicative words. On the other hand, non-fiction summaries seem to contain words that might be seen in literature that is expository in nature, like theory, example, or subject.

Predicting Fiction Genres:

Speculative fiction:

Speculative fiction is a genre that is a bit hard to define, thus making this classification the classification with the lowest accuracy values with only a 71.9% accuracy in training and testing sets. Despite the low accuracy, it's predictive words make sense when observed.

```
Speculative fiction
              specie 0.78
         civilization 0.77
               orbit 0.76
               demon 0.76
             magical 0.76
              colony 0.76
              galaxy 0.75
               space 0.74
             humanity 0.74
              planet 0.73

Non Spec
              french 0.20
                bond 0.21
             russian 0.22
           detective 0.24
            murderer 0.24
              letter 0.24
             british 0.25
               hotel 0.25
                bank 0.26
                lady 0.26
```

The most predictive words are on the left, while the right is a word cloud of the predictive words, with more inclusions. Based on the most predictive words (and least predictive), it can be seen that speculative fiction contains more words that relate to groups of people, like civilization, humanity, and colony. These are all common tropes in speculative writing. In opposition are words about types of people, like french, russian and british. These will come up in a later word cloud as well. Also notice that while lemmatization was applied, it is not perfect, as specie is not a word, but is still included.

Science fiction:

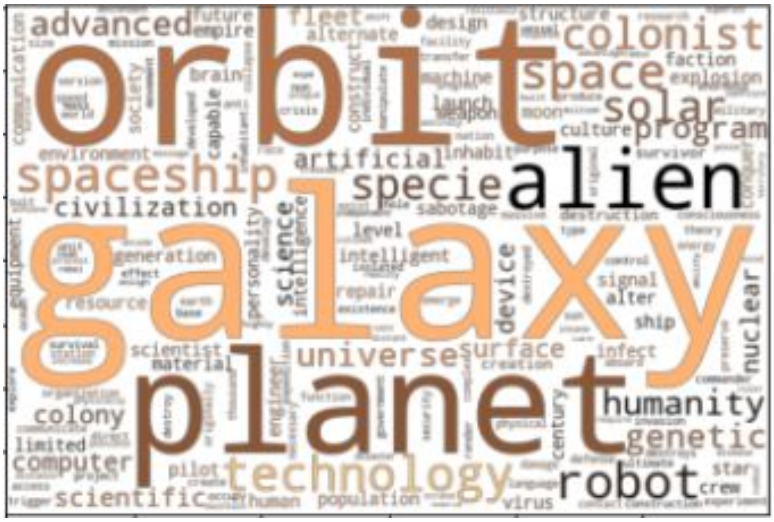
Science fiction is a more concrete genre, though the data set also includes some interesting subgenres, like steampunk. The Naive Bayes model does a bit better, with an accuracy of 79.4% on training, and 78.5% on the testing set.

Sci Fi

galaxy	0.99
orbit	0.97
planet	0.97
alien	0.95
spaceship	0.95
space	0.94
robot	0.94
technology	0.93
colonist	0.93
solar	0.93

Non Sci Fi

aunt	0.08
lawyer	0.16
detective	0.18
french	0.18
grandmother	0.18
letter	0.19
money	0.20
marry	0.20
wedding	0.21
birthday	0.22



Science fiction is often a genre about battles and epics, which can be seen in the predictive words, which deal with galaxis, and space, and aliens. The speculative words deal more with relationships and events, with words like aunt and grandmother as low indicators, as well as events like wedding and birthday.

Fantasy:

The fantasy model has an accuracy of 77.4% on the training and 74.5% testing set, similar to the accuracy of the speculative fiction model.

Fantasy

magical	0.97
magic	0.96
dragon	0.94
spell	0.94
sword	0.87
kingdom	0.84
god	0.84
warrior	0.83
quest	0.81
king	0.80

Non Fantasy

american	0.18
money	0.18
government	0.22
officer	0.23
agent	0.23
police	0.23
car	0.24
case	0.26
military	0.27
suicide	0.28

Unsurprisingly, magic is the most common word in the fantasy genre, with a 96% probability of being in a fantasy summary. The other common words also fit common themes of fantasy, with words like dragon and spell. For a genre that is known for its epic stories, it is unsurprising that words that would be associated with it fit in to those stories. The non-fantasy words seem to point towards a couple of different potential themes, with a more police procedural sort of feel, like officer, agent and case. Notice that both magic and magical are included, as an error in the lemmatizer.

Children's Literature:

Children's literature is an interesting case, where it is a common secondary genre, with books like Harry Potter that could fit into both children's literature and fantasy. This model hovered around an accuracy of 71.2% in training and 68.4% in testing sets.

Children's Literature

school	0.75
boy	0.70
parent	0.68
play	0.68
girl	0.65
mother	0.65
little	0.64
house	0.63
adventure	0.63
home	0.63

Non Children's

planet	0.15
contact	0.25
information	0.29
mission	0.33
future	0.33
murder	0.34
ship	0.34
case	0.34
men	0.35
various	0.35

The common words here are also not a surprise, with school, boy, and parent. It shows that a lot of children's literature tends to focus in on the events of a child's life. The other side also makes sense, as children's literature doesn't often combine with the stories that talk about murder, or missions (like mystery, or thriller).

Mystery:

Mystery included both classic mysteries, like Agatha Christie's work, as well as spy fiction and crime fiction, which include mysteries in their plot. The Naive Bayes model of the genre was 83.8% accurate on the training set and 82.0% on the testing set.

Mystery

murderer	0.92
detective	0.89
solve	0.87
lawyer	0.87
killer	0.87
murder	0.85
investigation	0.85
case	0.84
crime	0.83
mystery	0.82

Non Mystery

planet	0.06
space	0.10
destruction	0.15
alien	0.17
slave	0.18
king	0.18
earth	0.19
soldier	0.20
quest	0.20
journey	0.20

All of the indicative words in mystery fit the common tropes of the genre, with the common people of murderer, detective and lawyer as well as the common ideas like investigation, case and crime. From the words that are not indicative, it can be posited that there is not a lot of crossover between mystery and science fiction, as many of the science fiction words are not indicative of mystery.

Historical fiction:

Historical fiction has the most obvious definition of the genres in the data set, with historical pieces being set in the past. The model also showed one of the largest differences between the training and testing set, with an accuracy of 81.5% in training, and 78.9% in testing after tuning the model.

```
Historical Fiction
               french 0.89
           historical 0.82
              english 0.82
            political 0.80
              british 0.77
               priest 0.76
                slave 0.76
             marriage 0.76
              soldier 0.76
              wealthy 0.76

Non Hist Fic
               planet 0.01
                human 0.09
             creature 0.12
                  fly 0.18
               police 0.18
              ability 0.19
              problem 0.20
                world 0.20
                 game 0.25
              destroy 0.25
```

Historical fiction often deals in the relationship of governments, so french, english, political and british all fit that trope. The indicative words also include slave, which may indicate the era that many of the summaries in the data set were placed in, as well as soldier, which may indicate a common setting as well.

Young adult literature:

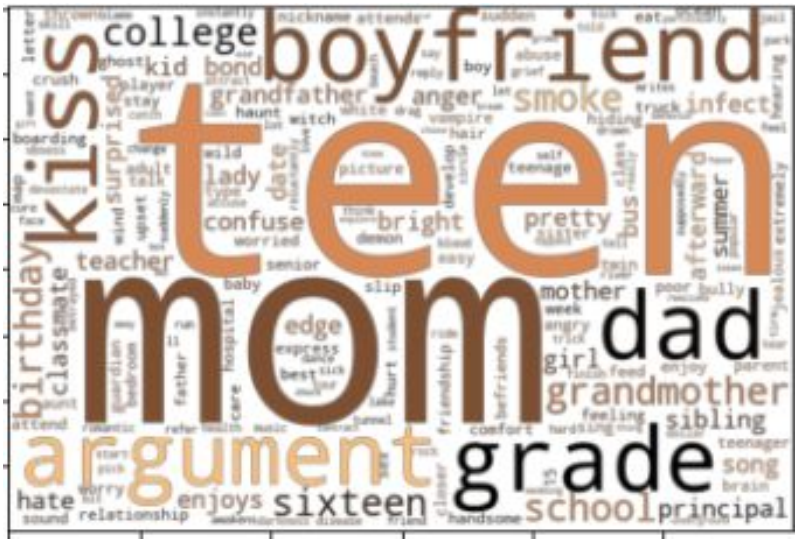
While similar to children's literature, young adult literature (or YA), tends towards more adult themes, and is geared towards teenage audiences. The model had an accuracy of 79.9% on the training set and 77.3% on the testing set.

Young Adult

teen	0.90
mom	0.88
boyfriend	0.87
kiss	0.87
dad	0.85
grade	0.82
argument	0.81
grandmother	0.81
college	0.80
sixteen	0.79

Non Young Adult

humanity	0.06
alien	0.13
murderer	0.14
invasion	0.14
nuclear	0.15
troop	0.16
nation	0.16
intelligence	0.17
ally	0.17
assistance	0.17



Just like children's literature, YA has predictive words about family, like parents. It also has relationship words that are more adult themed, like boyfriend, and kiss. These words may be a result of the coming of age stories that are common in YA, where the protagonist learns more about what it means to be an adult. The non-indicative words pull heavily from science fiction (with humanity and alien at the top) as well as mystery, with murderer.

Suspense:

Suspense has many common themes with mystery (and later, thriller), though suspense often deals with a mounting sense of dread or urgency as the protagonist digs further into the main mystery or problem of the novel. The model performed at 80.8% accuracy on the training set and 77.9% on the testing set in this case.

Suspense

lawyer	0.90
phone	0.86
investigation	0.85
murderer	0.84
blackmail	0.84
murder	0.84
police	0.82
detective	0.82
case	0.81
estate	0.80

Non Suspense

planet	0.06
human	0.18
space	0.20
defeat	0.21
century	0.23
journey	0.23
army	0.23
ship	0.23
city	0.24
animal	0.24

Suspense has very similar indicative words to mystery, though it includes detective, blackmail, phone and estate in its top 10. It also places lawyer significantly higher, suggesting many of the summaries had lawyers as main or important characters in their plot. It seems to separate itself from war-like terminology, with words like defeat, and army in the non-indicative words. It also seems to avoid science fiction, with planet and space in the opposing words.

Horror:

Horror is a well known genre to most, for its very specific presentation in visual media. In this case, the model struggled a bit with over fitting again, with a training accuracy of 76.0% and a testing accuracy of only 75.4%.

Horror

vampire	0.91
blood	0.80
room	0.65
trap	0.64
night	0.62
reveals	0.62
fear	0.61
kill	0.61
strange	0.61
figure	0.61

Non Horror

war	0.26
plot	0.29
party	0.32
land	0.32
police	0.34
ship	0.34
journey	0.35
work	0.35
marry	0.35
leader	0.36

Horror deals with the moment of death more than some of the other genres, with words like blood and kill in its top ten. It also builds tension, with words like trap, reveals, fear and strange. Finally, a commonly used character in the summaries of the set was the vampire, making it very indicative of this genre. Horror doesn't tend to include police, or war, unlike some of the other genres before it.

Thriller:

To finish out the chunk of related genres, thriller falls into a similar vein as suspense and mystery. Thrillers are often differentiated as containing an immediate threat from the onset to the main character. The model has 75.2% accuracy on training and 72.3% accuracy on testing.

```
Thriller
              agent 0.82
               team 0.82
      investigation 0.77
               shot 0.75
        information 0.72
            mission 0.72
               plot 0.72
              shoot 0.71
                car 0.71
         government 0.71

Non Thriller
            journey 0.25
              human 0.28
               city 0.30
          encounter 0.33
             travel 0.33
              great 0.34
             little 0.35
             living 0.35
               away 0.35
             mother 0.36
```

Thriller has very similar words to mystery and suspense before it, with things like investigation. But, the words seem to lean more towards government level characters, with words like agent, plot, and government. Also note that in the word cloud, few words are truly bigger, showing that there are a large number of words that are used in some but not all thrillers, like american, or hospital.

Application to the Web and Twitter

After the initial analysis of the models, they were combined to create a web app that allows the user to interact with the models in two different ways. The first takes Twitter usernames and looks at their most recent 100 tweets in order to predict the most likely fiction genre that the writer is using. The second takes any text and checks it first to see if it is fiction or nonfiction, and then if it is fiction, it checks for the likelihood of various fiction genres.

Twitter Test:

Once a username is submitted, the website pulls the most recent 100 tweets by the person. The system then removes retweets to keep to the writing done by the actual user. Then, the tweets go through a set of cleaning steps that lemmatize and vectorize all of the tweets together, as well as separately. Looking at all of the text together, the system predicts on all of the genres,

and produces a likelihood for each, of which the top 5 are graphed. The top genre is chosen, and predicted across the single tweets, in order to find the tweet that indicates the top genre best.

The top tweet, top 5 tweets, and a word cloud associated with the top genre are displayed for the user. There are also 3 options of presets, for the accounts of former president Barack Obama as well as two fiction-based accounts, Good Captain and Mayor Emanuel.

Summary Test:

The summary test produces similar results to the Twitter test, with the addition of the fiction vs. nonfiction split. If the work is fiction, the system will provide similar results to those of the Twitter test, with both a graph of the top 5 genres as well as a word cloud associated with the top genre shown. It also applies the same cleaning steps as the Twitter test.

Conclusion:

Genre is consistently hard to define, as are its common tropes. But, as this analysis shows, the language used is much easier to see. The word clouds above show this the most starkly, as nearly all words included fit the basic ideas associated with the genre they are representing. This language is common amongst thousands of summaries, despite being written by different authors about different works. The consistency of these results suggests that Western culture has a common language in how it represents genres that have become the norm. It also shows how consistent that common language is, since it is so obviously seen in the results.

From this analysis, there are a few potential routes of further analysis. First would be to get actual text from literature and use it to predict genre. There may be more variability from author to author, as the conventions and idiosyncrasies of the writer affect the word choice. It may also be found that the common language found in this analysis would be further cemented by the addition of longer text examples. Another possible route of analysis could be to apply this same framework to summaries in other languages. It may be found that the concepts in those languages follow a similar pattern of common sense answers. It may also be found that they show similarities and differences to English, and may also have their own large genres that are distinctive from those genres common in English based media.

To explore this analysis yourself, use the applied model on both summaries as well as Twitter usernames at sgunners.pythonanywhere.com.