Wilson Housen

Prof. Ide

May 22nd, 2018

Computational Linguistics

An Exploration of James Joyce's Work

One of the first resources made available to us, or that we were made aware of, was Project Gutenberg, which offers free ebooks in a number of different file formats for download. Seeing that these files could be used as corpora for programs such as NLTK piqued my curiosity. As a student, but also as an english major, I have been writing essays analyzing fiction for most of my life. I thought, what could I find and show from a literature perspective using natural language processing? Do what degree could these tools be used to understand things about literary texts?

I started out by considering which author or authors would be the best subjects for this kind of analysis. I was restricted by convenience, as it might be hard to find utf-8 text documents of authors that do not appear in Project Gutenberg, so I mainly looked for authors in Project Gutenberg. The first author I thought of, and found, was James Joyce. This was helpful to me personally, as I am familiar with two of Joyce's works, and I have at least attempted to read Ulysses. Plus, Joyce's dense, wordy style coupled with an evolution over the course of his works make him a prime candidate for this kind of statistical approach. While for this project I focus on James Joyce, I also found other authors whose works I threw into the ring. The second I went for was Virginia Woolf. While I am less familiar with her works, she serves as a good point of reference compared to Joyce, as she also writes within the modernist movement, and produced

her work in the early twentieth century. For a third source, I went a little further away from modernism. I wanted to stick with that era of literature, and fiction, so I chose some of the Holmes stories that Arthur Conan Doyle wrote, which were a natural fit as I have read many of them. I also liked this choice as these detective short stories were all stylistically similar, and that sort of consistency would lend itself to comparison. Multiple authors were used specifically as predictive points of reference for the classifier used to determine authorship. Both authors were from the same general time period, and at least one was also of the modernist movement, so their writing would form a good basis for the determination of authorship relative to Joyce and Joyce's work.

For existing work, I focussed on examples of natural language processing that were employed to analyze something about Joyce's work. The first work cited does exactly that. Joyce's novels and short stories have a reputation for being dense and hard to read, so the author of this study, Mario Borja, decided to hold that idea up to the scrutiny of natural language processing. He looked at whether textual difficulty was quantifiable, and if so how difficult were Joyce's texts. Starting with Dubliners, Joyce's popular series of short stories, he calculates their readability with the existing Flesch's Reading Ease Score, an algorithm that was thought up to determine the readability of the text. Given total words, total sentences, and total syllables, the formula is, roughly: 206 - 1.015(words / sentences) - 84.6(syllables / words), where the higher the score is, the easier a text is to read. In addition to Dubliners, Ulysses and Finnegan's Wake, Joyce's more experimental and complex pieces, are analyzed. Surprisingly, no individual short story or novel rank under a sixty on the algorithm, meaning that they all range from not very hard to easy to read, which is the conclusion his forced to make. While Mario expected this, I did

not, and wonder how adept this algorithm is at finding this kind of difficulty. The algorithm was designed for the readability of non-fiction after all; been produced to analyze technical manuals and expanded to insurance forms and other papers. It could be argued that this goes back to the algorithm itself and its focus on words and syllables. It takes no account for sentence structure, which along with diction and figurative speech is where much of the complexity of literature comes from. This is not to say that this is an invalid solution; the algorithm does measure a kind of complexity of text, and that is relevant by itself. I would say, however, that a better algorithm could be found to describe literary complexity rather than this form of non-fiction complexity.

The second piece I examined was Chris Beausang's work on the narrators of Ulysses. I found this work notable for two reasons: comparison of the narrators was something I had thought to look into, and that Beausang was using this statistical analysis to create a refutation of an interpretation of Ulysses. This piece is the closest to what I set out to do: use natural language processing to facilitate a literary analysis of the text. Specifically, he used what he described as the 6 narrators of Ulysses to refute the idea that Ulysses was wisdom literature: that Ulysses is a piece offering a reflection of what modern man should be. He argues that this wisdom literature narrative arises from the fusion of two narrators: Stephen and Bloom, who somehow become Blephen and Stoom. So, to analyze how close the narrators really become when they become Blephen and Stoom, he uses word frequencies and components to form a correlation matrix, then uses cluster analysis to visually display how close the narrators are to one another. If the wisdom narrative were to hold true, Blephen and Stoom narrations would be somewhere in the middle of the Stephen and Bloom ones. If not, then that theory can be argued to be less tenable. The correlation matrix, once formed, has both a significant spread and a gendered partition, which is

interesting, but also supports Beausang's hypothesis that the narrators do not actually come together as the narrators are so widely spread as to be unable to be described as the middle of one another. This statistical platform allows him to conclude that they are not the compromise or the summation of the two, and proceeds to describe what kind of a literary text Ulysses really is, given this fact: it is a book about how we do not live rather than how we do live. While the conclusion of this piece is not as directly linked to the hypothesis as in the Borja piece, I appreciated how linked the literary analysis was to the statistical analysis. It is hard to refute literary theory on text alone, and bringing in something more concrete as a mathematical, natural language approach bolsters a refutation that was necessary for the presentation of Ulysses as the opposite of this wisdom literature.

As corpora, I chose nine texts, three from each author. From James Joyce, I took Ulysses, A Portrait of the Artist as a Young Man, and Dubliners, two novels and a collection of short stories respectively. From Arthur Conan Doyle, I took the Bruce-Partington Plans, a Study in Scarlet, and The Hounds of Baskervilles, a short story and two novels respectively. From Virginia Woolf, I took Jacob's Room, Night and Day, and The Voyage Out, all novels. Each text was published between eighteen eighty eight and nineteen twenty two, and therefore represent a slice of the literature of the turn of the century. The choice of contemporaneous texts was on purpose so as to provide appropriate comparisons when attempting to classify authorship. It would make sense to provide a wider corpora that also included examples from the same time period and genre.

For tools, I used two tools. The first I used was WEKA, mainly for the classification and ease of use. It provided me with a built in Naive Bayes classification model. The second tool I

used was an arff-writer designed by the github user Waltaskew specifically to read in Project Gutenberg files and generate features for classification. This was also a convenience tool, as it was an easy way to go through the data and create WEKA readable files.

One of my hypotheses about James Joyce was on the similarity of his texts, and the predictability of authorship. The timeline of James Joyce's working on his titles creates an interesting question: of the three, Dubliners is published first in nineteen fourteen, then A Portrait of the Artist as a Young Man in nineteen sixteen, then Ulysses in nineteen twenty two. Given Portrait is closer to Ulysses chronologically, it can be said that it must be closer stylistically; especially given Joyce's reputation for an experimental style that only grew more complex as he aged (Finnegan's Wake, his most complex and labyrinthian text, was his last, in nineteen thirty-nine.) But, Ulysses was originally planned to be a short story in Dubliners, ending the collection instead of Joyce's famous short story The Dead. Deciding against it, he shelved the idea but eventually returned to it, developing it into a full length novel rather than a short story. Based on this, it can also be said that Dubliners had similar stylistic and creative impetus to Ulysses. So, which text ends up being more similar to Ulysses? To answer this question, I decided on a parameter: to use both as training texts, along with the other turn of the century pieces, and see which one can predict the author of Ulysses. The idea being, that if one text is determinative of authorship, then the training and test texts must have some sort of stylistic similarity, given the right features. As quote-unquote control tests, I also tested whether Doyle's works were predictive of Doyle's other works, and if Woolf's works were predictive of Woolf's other works, and if both Dubliners and Portait in tandem were predictive of Ulysses. I ran the classifier twice, once with four features and once with seven features. The initial four features

were word count, lexical diversity, number of unique words, and authorship. Too those four I added in the second round of testing average sentence length, number of object and subject pronouns, and average word length. The process was just running naive bayes classifier on the training corpora, then seeing if it predicted the author of the test text correctly or not.

For the first round, with four features, neither Portrait, Dubliners, or both were able to correctly predict that Ulysses was authored by Joyce. This outcome I would attribute to Ulysses itself, which is a notoriously complex book, and it would be unsurprising to me that it is stylistically unrecognizable compared to Joyce's earlier texts. The controls all came out positive: Woolf's work successfully predicted all of Woolf's other work, and Doyle's work also successfully predicted all of Doyle's other work. With Doyle this makes sense: all of his texts are detective serials with no grand evolution of style or content. I more or less expected this outcome, and it shows me that my process might be functioning somewhat as hoped. I did not really have an expectation for Woolf's work; modernism is an experimental and varied genre, but maybe Woolf is just more stylistically contained than Joyce, who writes almost with a mundane, natural air in Dubliners but explodes in his experiments of Ulysses. I took all of these results with a grain of salt, however, due to the limiting number of features.

The second round of testing were more curious. To the original four, average sentence length, average word length, and number of subject and object pronouns were added. Dubliners once again failed to predict the Joyce-ness of Ulysses, as does Portrait, and both of them combined. What is curious is the other controls. Woolf's texts remained predictive, but the classifier now fails to predict both Bruce Partington Plans and A Study in Scarlet of Conan Doyle. This is the output from the first run on a Study in Scarlet:

=== Run information ===

Scheme: weka.classifiers.misc.InputMappedClassifier -I -trim -W weka.classifiers.bayes.NaiveBayes
Relation: author_detection
Instances: 8
Attributes: 4
lexical_diversity
unique_words
word_count
author
Test mode: user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

InputMappedClassifier:

Naive Bayes Classifier

	Class		
Attribute	Virginia Woolf (0.36)	James Joyce (0.36)	Arthur Conan Doyle (0.27)
lexical_diversity			
mean	0.1496	0.1654	0.2126
std. dev.	0.0445	0.0039	0.0709
weight sum	3	3	2
precision	0.0236	0.0236	0.0236
unique_words			
mean	15907.8095	23861.7143	5965.4286
std. dev.	2812.13	16872.78	994.2381
weight sum	3	3	2
precision	5965.4286	5965.4286	5965.4286
word_count			
mean	133174.9048	133174.9048	36320.4286
std. dev.	45299.5333	85608.0711	36320.4286
weight sum	3	3	2
precision	36320.4286	36320.4286	36320.4286

Attribute mappings:

```
Model attributes                      Incoming attributes
----------------------------     ----------------
(numeric) lexical_diversity      --> 1 (numeric) lexical_diversity
(numeric) unique_words           --> 2 (numeric) unique_words
(numeric) word_count             --> 3 (numeric) word_count
(nominal) author                 --> 4 (nominal) author


Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances           1               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0008
Root mean squared error                  0.0009
Relative absolute error                  0.1668 %
Root relative squared error              0.1752 %
Total Number of Instances                1

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
Class
               ?        0.000    ?          ?        ?          ?        ?         ?          Virginia Woolf
               ?        0.000    ?          ?        ?          ?        ?         ?          James Joyce
               1.000    ?        1.000      1.000    1.000      ?        ?         1.000      Arthur Conan Doyle
Weighted Avg.  1.000    ?        1.000      1.000    1.000      ?        ?         1.000

=== Confusion Matrix ===

 a b c   <-- classified as
 0 0 0 | a = Virginia Woolf
 0 0 0 | b = James Joyce
 0 0 1 | c = Arthur Conan Doyle
```

And this is the output from running the second classifier on a Study in Scarlet:

```
=== Run information ===

Scheme:       weka.classifiers.misc.InputMappedClassifier -I -trim -W
weka.classifiers.bayes.NaiveBayes
Relation:     author_detection
Instances:    8
Attributes:   7
              avSentenceLength
              avWordLength
              lexical_diversity
              numPronouns
              unique_words
              word_count
              author
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

InputMappedClassifier:

Naive Bayes Classifier

                          Class
Attribute             Virginia Woolf       James Joyce Arthur Conan Doyle
                        (0.36)          (0.36)          (0.27)
==========================================================================
=====
avSentenceLength
  mean                  83.4066         70.7143          70.7143
  std. dev.             11.1772          8.8828          5.4396
  weight sum                  3               3               2
  precision              5.4396          5.4396          5.4396

avWordLength
  mean                   4.7194          4.5965          4.4965
  std. dev.              0.0784           0.115          0.0231
  weight sum                  3               3               2
```

precision	0.0461	0.0461	0.0461
lexical_diversity			
mean	0.1496	0.1654	0.2126
std. dev.	0.0445	0.0039	0.0709
weight sum	3	3	2
precision	0.0236	0.0236	0.0236
numPronouns			
mean	266349.8095	266349.8095	72640.8571
std. dev.	90599.0666	171216.1423	72640.8571
weight sum	3	3	2
precision	72640.8571	72640.8571	72640.8571
unique_words			
mean	15907.8095	23861.7143	5965.4286
std. dev.	2812.13	16872.78	994.2381
weight sum	3	3	2
precision	5965.4286	5965.4286	5965.4286
word_count			
mean	133174.9048	133174.9048	36320.4286
std. dev.	45299.5333	85608.0711	36320.4286
weight sum	3	3	2
precision	36320.4286	36320.4286	36320.4286

Attribute mappings:

Model attributes	Incoming attributes
(numeric) avSentenceLength	--> 1 (numeric) avSentenceLength
(numeric) avWordLength	--> 2 (numeric) avWordLength
(numeric) lexical_diversity	--> 3 (numeric) lexical_diversity
(numeric) numPronouns	--> 4 (numeric) numPronouns
(numeric) unique_words	--> 5 (numeric) unique_words
(numeric) word_count	--> 6 (numeric) word_count
(nominal) author	--> 7 (nominal) author

Time taken to build model: 0 seconds

=== Evaluation on test set ===

```
Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances           0               0      %
Incorrectly Classified Instances         1             100      %
Kappa statistic                          0
Mean absolute error                      0.6077
Root mean squared error                  0.6732
Relative absolute error                125.3407 %
Root relative squared error            130.9132 %
Total Number of Instances                1

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               ?        1.000    0.000      ?        ?          ?        ?         ?         Virginia Woolf
               ?        0.000    ?          ?        ?          ?        ?         ?         James Joyce
               0.000    ?        ?          0.000    ?          ?        ?         1.000     Arthur Conan Doyle
Weighted Avg.  0.000    ?        ?          0.000    ?          ?        ?         1.000

=== Confusion Matrix ===

 a b c   <-- classified as
 0 0 0 | a = Virginia Woolf
 0 0 0 | b = James Joyce
 1 0 0 | c = Arthur Conan Doyle
```

Strangely, increasing the amount of features has caused more failure in the classifier. Looking at the data, this could be the result of human error. The average sentence length seems high; all authors appear to average over seventy words per sentence which does not make sense. This could be due to how that feature was programmed. I split the text on periods, then on exclamation points, then on question marks, as those are the punctuation marks that commonly end sentences. This does not account for some things, however, such as shortened names, or

certain abbreviations. The method of splitting could also be counting white space as a word, or it could be counting every character as a word itself. These would explain the high number. That particular algorithm could be reworked. One method that could isolate the problem would be to run the classifier again with only one of the new features at a time to see if it falters on its classification. Whatever the shortcomings of the program, I feel I can claim that the results show that more consistent authorship and style leads to being able to predict authorship. Doyle is definitely the most consistent and least experimental, which lines up with the results of the first classifier. I believe that the reason the second classifier failed to classify two of Doyle's works was due to my own error, though if I were to not be at fault, my position becomes more shaky. All of Woolf's works across both classifiers were predicted, and while she was an experimental writer, her works were more consistently so than Joyce, who basically walked up a parabola of complexity over the course of his life. For the core, of whether Dubliners or Portrait is more stylistically similar to Ulysses, I will say that neither are. As neither predicted that it was a Joyce text with either classifier, alone or together, There is no reason to believe that either share creative or stylistic similarities to Ulysses over the other. I would conclude that this is due to the the evolution of Joyce's style; it is just that steep. So, while style may be predictive of authorship, it is not for Joyce.

For the future, more work could be done on the features. Average Sentence Length could be cleaned up, and more than seven features would also be welcome. It is also worth considering expanding the corpora. Finnegan's Wake is the most natural next step; maybe Joyce's most experimental work would have something in common with Ulysses? Woolf and Doyle also have more texts, but expanding the authors could be worthwhile itself. One could find authors who

went through similar stylistic self-revolutions as Joyce who might mirror the results here, which would confirm that stylistic evolution makes authorship hard to predict. Or, they could be predictive, which would mean that Joyce specifically is hard to predict. Programming refinements and data refinements could shed more light on this analysis of authorship and style.

Works Cited

https://www.statslife.org.uk/culture/1572

https://analoguehumanist.wordpress.com/2018/04/18/a-statistical-analysis-of-the-narrators-of-ulysses-or-why-ulysses-isnt-wisdom-literature/