Michael Znidarsic

IST 736

Assignment #3

**Cicero Salutem**

Analysis of text to determine its authorship is an application of text mining that dates to the study’s genesis and is probably the most quintessential study in the field, yielding a long backlog of scholastic achievements and methodological advances. But what if there is another person or persons implicated in the text besides the author? Instead of the age-old authorship question, one might instead wonder as to the identity of the recipient. In the case of letters and other personal correspondence, is it possible for data scientists to determine who an epistle was written *to* using only its contents? This is a question asked less often, and worse answered.

A familiar figure in history provides us with an opportunity to test this concept. Marcus Tullius Cicero is one of the most important writers of antiquity. Like other essential historical characters, he is so widely known and read simply because he wrote a huge amount of text, and that text happens to survive. But unlike other giants of his era of history (Herodotus, Thucydides, Polybius, *et al*.) he did not primarily confine his influence to act on posterity; in fact, he achieved massive fame and political preeminence for many years of his life. He was known during his time and in the eras succeeding as the greatest orator, lawyer, and legislator of the late Roman Republic. Despite coming from a middle-class background, he was elected Consul for the year 63BC, during which he foiled the infamous Cataline Conspiracy. For the rest of his life he would serve as a stalwart defender of the republic in the face of the growing power of the triumvirates. One triumvir known to modern readers as Marc Antony would eventually have Cicero assassinated in 43BC, nailing to the Rostra of the Forum the orator’s hands and head, which had respectively written and delivered so many a damning speech against him.

Cicero’s power and connections lent themselves to interesting correspondence, much of which survives. The two people to whom he wrote most often and at the greatest length were his brother, Quintus, and his best friend, Atticus. These letters offer a great opportunity to test whether modern text mining can determine who a letter is addressed to simply through text analysis of the letter’s contents.

This ancient source originally written in Latin offers yet another novel challenge, however. Determining who a letter was addressed to implies a sort of change in the authorial “voice” depending on who the intended recipient was, so the nature of this sort of linguistic forensics begs the question: are distinct characteristics and changes in authorial “voice” lost when a source is translated to another language? This question, if answerable in a universal sense independent of individual translators and languages involved, has critical implications on text mining’s limitations and more essentially on our understanding of how human beings communicate. This study will pose this question in a more testable format: are standard text mining tools better able to classify a letter’s addressee in the original language than in a translation thereof?

***Cui dono lepidum novum libellum arida modo pumice expolitum?***

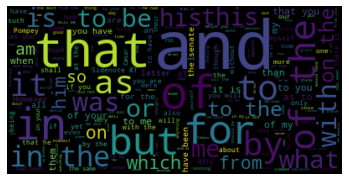
**Catullus**

**About the Data**

There were two separate datasets used for the study. The first was a subset from a collection of Cicero’s early letters translated into English by Evelyn S. Shuckburgh; this data was accessed via Project Gutenberg and contained letters to Quintus, Atticus, and others during the early years of Cicero’s extant correspondence (approximately the mid-60’s BC to the mid-50’s BC). Only letters written by Cicero to either Quintus or Atticus were taken from the source for inclusion in the dataset, allowing for binary classification. It was important to ensure that this dataset, being in English, was all translated by the same person to avoid additional variance in style beyond translation itself.

The second was a subset of a collection of Cicero’s letters in the original Latin access via an online repository called “The Latin Library”. To avoid the potential confounding factor of differences in authorial voice based on age of the author or topics and events in the author’s life, the letters taken from this collection were almost the same letters as those taken from the first source, making both datasets confined to the same approximate time in Cicero’s life. This dataset contained 26 letters by Quintus and 89 letters by Atticus.

When comparing the most common words in both datasets by the two wordclouds below, several trends become apparent. Critically, the relative word frequencies are nowhere near one-to-one. While “est” is clearly visible in the Latin set, its translation “is” is much smaller in the English set. The same is true for “mihi” and “I”, “nihil” and “nothing”, “non” and “not”, and countless others. Although these are mainly basic parts of speech, such discrepancies demonstrate both the change in authorial voice upon translation and the inherent differences in structure between Latin and English.





Each of these datasets were vectorized into their own separate matrix of raw word counts. Numeric (e.g. “15”) tokens were removed, as were words that would act against the integrity of the models, such as the words “Quintus” and “Atticus”. However, the sporadically used Greek words in both sets were left in, as well as those few words in the English set that were left in their original Latin. Raw word counts were converted into TF-IDF normalization, as both would be tried with the models.

Latin is a language in which words have many different conjugations that change their meaning or tense. As such, stemming was attempted on both English and Latin datasets (the CLTK Python package is highly recommended to any readers looking to perform text mining on ancient languages). Stemming failed to improve the results in either case, so it will not be discussed at length in this study.

**Analysis**

**Model #1**

The first model sought use SKLearn’s DecisionTreeClassifier on the translated English dataset to classify a letter’s addressee. The English dataset consisted of 27 letters to Quintus and 92 to Atticus. In order to utilize the maximum acceptable amount of the small sample of Quintus papers for training, the training set was randomly selected to contain a balanced 19 Quintus letters and 19 Atticus letters. A balanced set of the 8 remaining Quintus letters and 8 more randomly selected Atticus letters was created for testing. The remaining 65 Atticus letters would be reserved for a further round of validation testing.

**Model #2**

The first model sought use SKLearn’s MulinomialNB classifier on the translated English dataset to classify a letter’s addressee. The English dataset consisted of 27 letters to Quintus and 92 to Atticus. In order to utilize the maximum acceptable amount of the small sample of Quintus papers for training, the training set was randomly selected to contain a balanced 19 Quintus letters and 19 Atticus letters. A balanced set of the 8 remaining Quintus letters and 8 more randomly selected Atticus letters was created for testing. The remaining 65 Atticus letters would be reserved for a further round of validation testing.

**Model #3**

The third model sought use SKLearn’s DecisionTreeClassifier on the original Latin dataset to classify a letter’s addressee. The Latin dataset consisted of 26 letters to Quintus and 89 to Atticus. In order to utilize the maximum acceptable amount of the small sample of Quintus papers for training, the training set was randomly selected to contain a balanced 18 Quintus letters and 18 Atticus letters. A balanced set of the 8 remaining Quintus letters and 8 more randomly selected Atticus letters was created for testing. The remaining 63 Atticus letters would be reserved for a further round of validation testing.

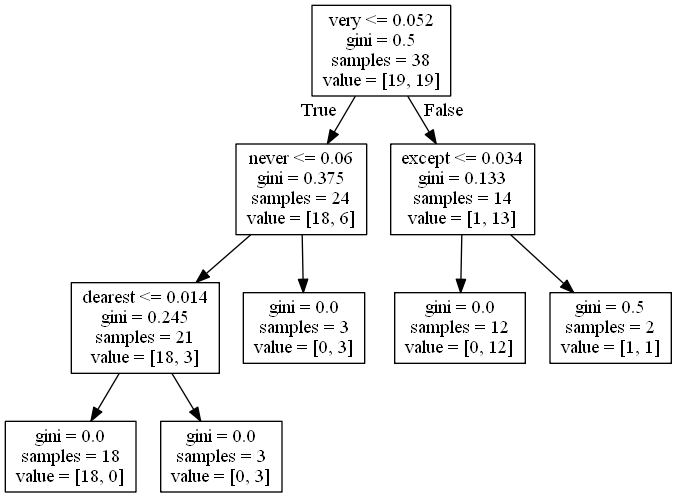
**Model #4**

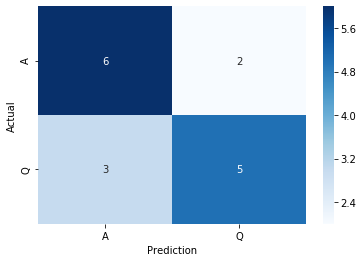
The fourth model sought use SKLearn’s MulinomialNB classifier on the original Latin dataset to classify a letter’s addressee. The Latin dataset consisted of 26 letters to Quintus and 89 to Atticus. In order to utilize the maximum acceptable amount of the small sample of Quintus papers for training, the training set was randomly selected to contain a balanced 18 Quintus letters and 18 Atticus letters. A balanced set of the 8 remaining Quintus letters and 8 more randomly selected Atticus letters was created for testing. The remaining 63 Atticus letters would be reserved for a further round of validation testing.

**Results**

**Model #1**

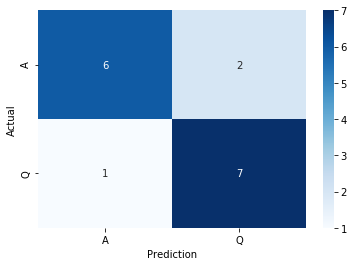
The English set’s decision tree model was completely ineffective with both word count and TF-IDF. Sometimes it would seem to simply guess randomly; still other iterations would overpredict to the point of guessing either Atticus or Quintus exclusively. One of the better performances on a balanced test set of 16 letters is pictured below. While exhibiting a 68.75% testing accuracy on a 97.368% training accuracy, this model was even worse than it appeared. When tested on the remained 65 papers from Atticus, it performed extremely poorly at only 52.308% accuracy. This model lacked all practical use as a classifier.





**Model #2**

The Naïve Bayes model conducted on the English set very rarely demonstrated even a decent accuracy, and even when it did, that accuracy did not carry over to the extra Atticus papers not used in the test and train sets. The model below achieved the best testing accuracy of 81.25% after a 97.368% training accuracy on TF-IDF normalization. This demonstrates significant overfitting, which is difficult to avoid in high-dimensional Naïve Bayes models, particularly with sparse matrices. When this model classified the 65 remaining Atticus papers, it was only 66.154% accurate. This fickleness suggested that the model’s decent performance on the balanced testing set was more of a fluke, due to a favorable selection of the testing set rather than any real robustness. As such, this model was not seen as reliable or effective.



**Model #3**

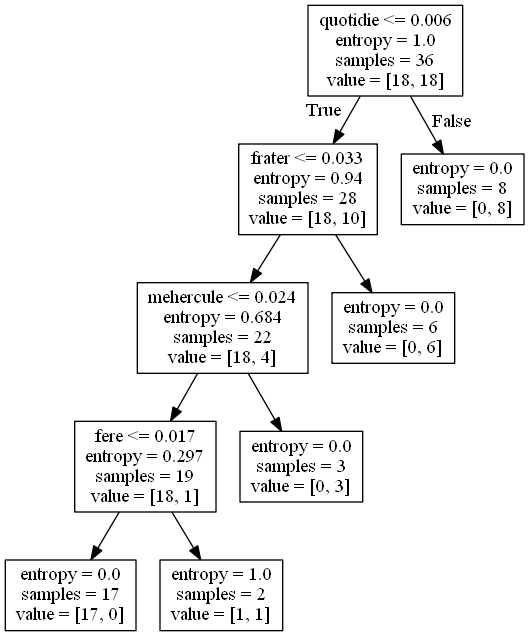
Using a decision tree on the Latin dataset proved very effective, yielding high accuracies with a fair degree of consistency. Initial tests with raw word count showed some promise, but it was the introduction of TF-IDF normalization that drastically increased its accuracy. Both the “gini” and “entropy” criteria were tried with similar results, but slightly more consistency from entropy. To avoid overfitting with a narrow and fragile model type and highly dimensional data, minimum samples per split was set to 3 and minimum leaf samples was set to 2; a maximum depth was not necessary to set due to the aforementioned parameters and the small size of the training set.

Decision trees contain an element of randomization, so there was some fluctuation on the effectiveness for each iteration of the model due to different word choice for the nodes. The unfortunately small size of the dataset once balanced led to some drawbacks: certain splits between test and train seemed to hinder the classifier’s ability to find important words. However, in instances of the model where certain words appeared, accuracy would follow. The three of these words noted in this study for their effective classification were “frater”, “quotidie”, and “mehercule”.

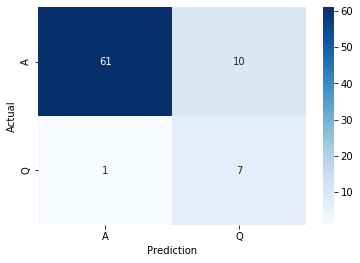
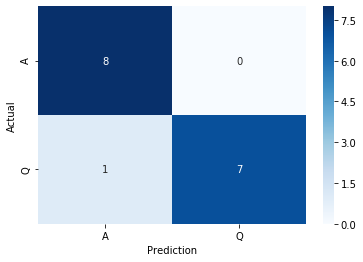
Of these three words, “frater” (translating to “brother”) proved the least clear in its significance. Although Cicero’s use of the word varies based on who he is writing to, it appears somewhat frequently in letters to both Quintus (in which it is used to address him) and Atticus (in which it is used to refer to Quintus). The word generally did not appear at the top of a tree, and perhaps the trees were effective at picking out the lower nodes at which frater becomes important to distinguish between Quintus and Atticus. All these findings defy the intuitive supposition that the word “frater” would be a dead giveaway as to the addressee of a letter when one of the addressees is, in fact, the author’s *frater*. Instead, this word appears to have played a supporting role in the models, and high accuracies were often achieved without it.

“Quotidie” (translating to “usually” or “ordinarily”), by contrast, appears 13 times in the 26 Quintus letters, but not a single time in the 89 Atticus letters. The reason for this is unclear, but the role that these two men played in Cicero’s life may shed some light. Quintus was an active partner in Cicero’s political efforts and spent much of his life living near his brother in Rome (this explains why there are fewer letters to him than Atticus), while Atticus spent most of his life in Greece and did not participate much in politics. Perhaps the somewhat practical use of the word “quotidie” was not necessary in the types of abstract, philosophic correspondence Cicero carried on with Atticus. It could also be that “quotidie” served as a sort of filler word, which did not lend itself to discourse with Atticus. This word appeared at the top of trees most frequently and showed itself to be the term most commonly found in trees yielding high accuracy.

Even more interesting, however, is the word “mehercule” (a compound word used as an exclamation and translating to “by Hercules!”). This word occurs 15 times in the 26 letters to Quintus, but only twice in the 89 letters to Atticus. Perhaps this disparity demonstrates a reluctance on the part of Cicero to resort to language that would be considered vulgar or ineloquent while conversing with the reserved and scholarly Atticus.



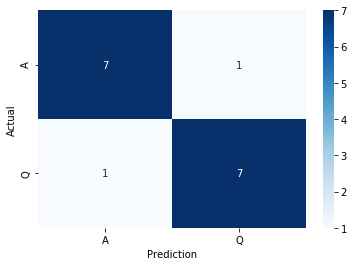
The tree model shown above was trained on a balanced set of 36 letters with a 97.222% training accuracy. The model achieved a 93.75% testing accuracy on the balanced test set of 16 documents. However, to ensure that the model’s accuracy was not simply luck involving the random selection of the testing set, the remaining 63 Atticus letters that went unused in the balanced train and test sets were added were classified using the model for an accuracy of 84.127%. This set of initially unused papers combined with the balanced test set for an 86.076% imbalanced testing accuracy. Considering the imbalance present in this augmented testing set, this result is impressive. It also showed minimal bias, with 85.915% accuracy on Atticus letters and 87.5% accuracy on Quintus letters. The confusion matrices of the original test set and the combined test set augmented with the initially unused Atticus letters are both shown below.



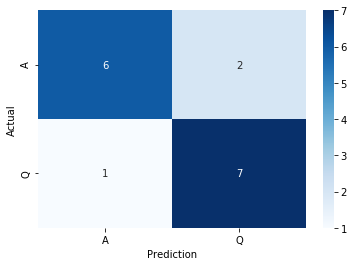
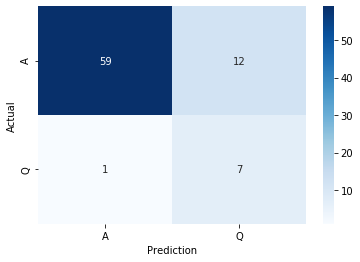
This model showed great promise, and due to its use of only 4 words it was even able to demonstrate that this level of accuracy was obtainable without any reliance on words that may have been confounding to the model’s integrity (e.g. if the model took one of the subject’s names, which would be considered a label, as an input). It proved the most effective, consistent, and parsimonious model of those tested.

**Model #4**

The Naïve Bayes model trained on the Latin dataset also performed better than the models trained on the English dataset but had some flaws. Performing terribly on the raw word frequency counts, this model also showed improvement upon introduction of TF-IDF normalization. Although less consistent than the decision tree model, producing often biased prediction outcomes, and not benefiting from the decision tree’s guaranteed security against confounding inputs, this model was capable of good accuracies on the balanced testing result. However, even on many iterations of the model with promising testing accuracies, the model performed very poorly on the 63 Atticus papers not included in the balanced testing or training sets. The version of the model shown below exhibited a 100% training accuracy on a balanced set of 36 letters and achieved an 87.5% testing accuracy on a balanced set of 16 letters, but only managed a dismal 52.38% accuracy on the 63 unused Atticus papers. Combined into an augmented testing set, the total testing accuracy was an unacceptable 59.49%.



Another iteration of the model obtained a lesser testing accuracy of 81.25% on a training accuracy of 100%. This model, however, managed an 84.13% accuracy when tested on the 63 unused Atticus papers. Combined into an augmented testing set, the total testing accuracy was an encouraging 83.544%. Accuracy on Atticus papers was 83.099% and accuracy on Quintus papers was 87.5% for a small bias. The confusion matrices below are of the balanced and augmented test set respectively.

Although the Naïve Bayes model did not quite reach the accuracy, consistency, or parsimoniousness of the Latin decision tree model, it vastly outperformed both models trained on the English dataset. Significantly, this result suggests that the type of model was not the primary factor that enabled the accurate prediction of an addressee; the more important requirement for creating a successful classifier was to train it on the language in which the text was initially written.

**Conclusions**

This study sought answers on two core questions. First, can modern text mining approaches determine to whom the author was addressing a given letter based on its text? Second, does the translation of a corpus into another language mask the voice of the original author, manifested by a decreased effectiveness of text mining in determining a letter’s addressee?

Predicting the addressee of a letter based on the contents showed very promising results and it was clear that, in Cicero’s writings at least, there was enough of a change in his word choice based on who he was addressing to be identifiable through text mining. What’s more, significant insight was gained as to *how* Cicero’s language choices changed when corresponding with either Quintus or Atticus. It appeared that when writing to Quintus, Cicero took a tone that was less eloquent and more practical, but when writing to Atticus he avoided profanity and discussion of more granular details of his political life.

However, none of the characteristics of Cicero’s prose that made his addressee identifiable seemed to survive translation into English. In their English translation’s, letters to Quintus and Atticus were in no way reliably distinguishable using the models tested. While this result may vary somewhat based on translator and does not serve as a *holistic* measurement of how much of an author’s “voice” survives translation, it serves as a solid piece of evidence for the limitations of text mining’s efficacy on translated works. While this is certainly not the first study to explore this concept, there remains much more research to be done on the effects of translation on the subtle and granular information buried in written works.

**References**

<http://www.gutenberg.org/ebooks/21200>

<http://www.thelatinlibrary.com/cic.html>