White paper on the DSI Capstone Project 'Operation: Debunker.'

Introduction and Problem Statement

Spurious quotations represent aa fascinating stylometric author analysis challenge to the NLP engineer. A quotation is typically very short, far from the hundreds of thousands of words found in the published works of an author. (For reference, the average novel is about 30,000 words, or 150,000 characters. A single tweet is just 142 characters.) Given an extant corpus, is it possible to predict the validity of a quotation that has been removed from its context?

Research and Sources

To identify worthwhile parameters for this project, I read quite a few papers, most of which are in the same repository as this white paper. In general, it appears that author analysis is most effective at the n-gram level, or, rather, that the character n-gram level has the highest degree of predictive power.

The training corpora were taken from Project Gutenberg (http://www.gutenberg.org). Examples of modern writing were taken from McSweeney's Internet Tendency (http://www.mcsweeneys.net), as well as from Project Gutenberg's works of James Allen (a precursor to Napoleon Hill and Dale Carnegie) and from the blog Thought Catalog (http://www.thoughtcatalog.com).

Validating quotations is a research task unto itself, as it was necessary to have at least a few verified and debunked quotations for each author in order to have any kind of representative test set.

Data Collection and Cleaning

Three authors were selected as test subjects for this project: Abraham Lincoln, Mark Twain, and Oscar Wilde. Among them, their lives span the whole of the long nineteenth century, from the Napoleonic Wars to the eve of World War I. Anecdotally, these three seem to have the most quotations attributed to them. Several factors account for the volume, which make them excellent test subjects for this project.

Following his assassination, Abraham Lincoln was mythologized almost overnight from a divisive figure into the Martyr of the Republic. Add to that his reputation as a folksy story-teller, which has been well-attested, and any quotation that sounds either saintly or folksy seems to get automatic validity. Lincoln's written style, anecdotally, reflects his training as a lawyer, more than his image as a backwoods farmer.

Mark Twain cut his teeth as a riverboat pilot on the Mississippi and as a reporter in the boom towns of the mountain west. He made a fortune serializing his novels and giving talks around the country – which accounts for his enormous volume of written work. Twain was known for irreverence and a clear-eyed cynicism, which makes him very popular to quote. A running joke in the spurious quote community – yes, that's a thing – is that if you don't know who said it, attribute it to Twain.

Oscar Wilde was the founding father of the aesthetic movement of the late 19th century. His mantra was 'Art for art's sake' and he shared Twain's gift for observation as he moved through the high echelons of London society. He made his reputation as a wit and aesthete, which makes it even more of a challenge to verify quotations – it's likely, as James Whistler quipped, that if Wilde didn't say it, he soon would.

Fortunately, all three of these writers have been dead long enough that their works are in the public domain, and their complete or collected works are on Project Gutenberg.

The bulk of this project went into cleaning the data. All of the documents came with editor's marks, prefaces, and introductions, none of which are intrinsic to the writer in question. Because there was no clear, repeatable pattern, I decided to clean by hand. I removed all of Gutenberg's copyright information and licensing, all introductions and prefatory remarks. That was the easy part.

At 3 million words, Mark Twain's is the largest corpus in the project. Mark Twain was very fond of quoting people in his work, sometimes his own work, and sometimes parodies that he made up. In those cases, I have attempted to research what was Twain's and what was someone else's. Those elements that were not Twain's were removed.

The Wilde corpus was by far the cleanest. His early editors were his disciples, and they made almost no notes or revisions to his works in the public domain versions currently available. Unlike Twain and Lincoln, Wilde's corpus does not include personal papers such as letters and telegrams, only his published writings. The Wilde corpus also excludes one of his most famous works, the play *Salome*. Gutenberg does have a translation from the original French by Alfred Lord Douglas, and although if anyone could have captured Wilde's style it would have been Bosey, I thought it best to limit myself to the writer's words as he wrote them.

The collected papers of Abraham Lincoln, published in 1953, includes most of the papers Robert Todd Lincoln collected after his father's death. The editor makes quite a few annotations, even going so far as to excise large swathes of text for 'irrelevance.'

This has the effect of creating the impression of hagiography, rather than biography, which may indicate that we do not have a 'clean' sample of Lincoln's writing. His early career as a lawyer include a lot of papers that are highly formalized, including legal jargon and presidential pronouncement language that is not necessarily native to Lincoln's own style. Those have been removed to the best of my ability.

Cleaning Lincoln by hand also revealed an interesting feature: his wartime telegrams. These follow a terse idiom, different in feel from the rest of the corpus. A book has been written about them, *Lincoln's T-Mails*. Given their similarity to modern text messaging, it might be worthwhile to examine the telegrams and only the telegrams for author analysis or stylometry.

Content from McSweeney's Internet Tendencey was pulled using a BeautifulSoup-based webscraper. The modern corpus is the smallest of the three, with only 67,000 words.

Author	Character Count	Word Count	Sentence Count
Mark Twain	14,712,397	3,143,540	136,199
Oscar Wilde	2,856,299	595,302	38,528
Abraham Lincoln	2,646,227	531,241	17,649
Modernity	3,941,00	692,367	39,657

On the whole, over 18 hours were spent removing nearly 100,000 words. In future, and on further consideration, leaving the text as-is apart from things that are clearly editor's marks – usually set off by brackets – would expedite the process. After all, I don't know what makes for a writer's style apart from 'feel' and my editing by hand may have introduced my own biases into the process.

The final step of preprocessing – getting data into a computer-readable form – was far and away the longest and most challenging. I decided that it made sense to have a variety of documents for each corpus. In part this was because the original plan called for using a TF-IDF vectorizer, which requires multiple documents. To simulate multiple documents, I tokenized each writer at the character, word, and sentence level, and broke those into pieces that returned 1,000, 500, and 100 observations each. Thus, Twain has a thousand entries of 14,712 characters, 500 at 6,287 words, and so forth.

<u>Modeling</u>

From the beginning, a Naïve Bayes multinomial classifier was going to be the model used. This is because Naïve Bayes assumes the independence of features. That assumption is a bit complex as it both makes sense and does not make sense. If we approach letters, words, and sentences as pieces of a whole, their interdependence is quite clear: letters make up words, words sentences, sentences speeches. Displacing a letter here, a word there, results in mounting confusion. But if the written word is only so much digital scratching on a non-existent pad, if it is in fact purely symbolic then their arbitrariness and their independence is assured. For my part, I'm inclined to think that the truth is somewhere in the middle, but for purposes of stylometry and for seeing patterns that humans cannot see because of the fog of simulacra, we are assuming an independence of features.

In addition, since the point of this whole exercise is to minimize the number of false positives – that is, to debunk quotations that do not properly belong to the speaker – the preferred metric is precision – true positives out of the total true values – rather than accuracy.

Having tested several variations of both count and tfidf vectorization, at both the word n-gram and the character n-gram, I have found that the count vectorizer with n-grams at the 1-4 range perform best on the training data, returning a mean precision of .98 on pair 7 (100 observations at the character level for all classes).

Wild data was scraped from successories.com, a website that sells engraved merchandise with 'inpsirational' quotations – exactly the sort of thing I'm interested in debunking. Approximate 1100 quotations from Twain, 500 from Wilde, and 500 from Lincoln were collected. Another thirty-odd were sourced and validated or invalidated by hand using twainquotes.com and wikiquote.

Results

The original results were disappointing, with only 8 of the 90 known quotations correctly attributed. The quotation from Lincoln that started the project was incorrectly attributed to Lincoln.

Much of the problem resulted from the significantly imbalanced classes. Once the "modern" corpus was brought up to over 600,000 words – more in line with its comparators – the results improved significantly, with approximately 30 of the known quotations attributed or debunked correctly.

There is more work to be done, of course. I know that the model works and that I have an idea worth pursuing at greater length. There are additional authors to add,

and there are other concepts to try, but these initial results are promising and I have reason to believe that building a universal debunker is possible, if not entirely feasible.