INKUBATE
Solutions for Discovering Successful Content

JUOLA & ASSOCIATES

Writing Style Analysis Applications

Published by Inkubate and Joula Associates, Spring 2016

Introduction

One of the most important questions for a new manuscript finding its way to market is its close neighbors, or put simply, "who does this person write like?" Being able to answer this question can help a new author inform readers about his or her novel, and being able to answer this question can also help the publisher find the new author they want -- "Bring me someone who writes like J.K. Rowling!" While these tasks are currently undertaken informally, we present a method based on the science of stylometry that can provide a robust, objective, informative, and cost-effective method to address these challenges.

What is Stylometry?

Stylometry is, in broad terms, the scientific study and measurement of writing style. More specifically, it generally refers to the computational analysis of a document in order to determine something about the document's author. It has a wide range of uses, including in history, law, medicine, and journalism, or anywhere we are interested in learning more about the person behind a piece of writing.

At its base, stylometry hinges on the theory that all language is a set of choices, and that people are free to choose among several different ways of expressing themselves. Each person makes his or her individual choices, but these choices are typically habitual --- a person will normally make the same choice in a similar context, because it "sounds right" to them. By tracking these choices, analysts can determine which person's choices a given piece of writing reflects, and by extension, who wrote the document in question (a very useful determination, for example, when the passage in question is a ransom note or a possibly-forged will).

In roughly the same way, "similar" people will often make "similar" choices (for example, US-educated writers will normally choose standard US constructions such as "honor" and "in the hospital," while Commonwealth-educated writers will choose "honour" and "in hospital"), and the same techniques can be used to identify group membership. This kind of "profiling" can extend to many different attributes such as gender, education level, age, and even personality type or self-esteem.[1]

Are "authors who sell well" another well-defined group of "similar" people? This point, despite its obvious commercial interest, has not received a lot of research attention, and there is no clear-cut answer. One major confounding factor is the simple fact that a lot more goes into making a best-seller than the quality of a manuscript, and even the best story may falter in the hands of the wrong publisher. On the other hand, we also all understand intuitively the appeal that "if you like Author X, you should like Author Y, because s/he writes very similarly."

---

[1] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, February 2009; M. Corney, O. de Vel, A. Anderson, and G. Mohay. Gender-preferential text mining of e-mail discourse. In *Proceedings of Computer Security Applications Conference*, 2002, pages 282–289, 2002; John Noecker Jr, Michael Ryan, and Patrick Juola. Psychological profiling through textual analysis. *LLC*, 28(3):382–387, 2013.; Elizabeth DeCarlo. Inferring authorship through Myers-Briggs Type Inventory. In *Proceedings of DHCS* 2013, Chicago, 2013; Patrick Juola and John I. Noecker Jr. Inferring self-esteem from keyboard behavior. In *Proceedings of DHCS 2014*, 2014; John I. Noecker Jr. and Patrick Juola. Stylometric identification of manic-depressive illness. In *Proceedings of DHCS 2014*, 2014.

Writing Similarity

We have developed a computer program to address "writing similarity" in a way that is objective, accurate, and transparently understandable. Using the same underlying technology that has been used in court[2] to address questions of identity, we identify four major axes of literary style and assess manuscripts along all four axes together. Using these, we can find both the closest match(es) to any given author along a single dimension, or we can fold these dimensions together to get an overall best match. Furthermore, these axes make sense, so an author/reader/reviewer can understand how a single story can be similar in some ways to Danielle Steel and in other ways to John Grisham (thus satisfying the over-demanding editor who wants "Steel meets Grisham!")

The core of our technology[3] is a method of breaking novels down into indicative "features" such as words, parts-of-speech, and similar tiny chunks, then assessing each novel for closeness using high-powered machine learning technology. For example, a modern, fast-paced action/adventure novel would be high on verbs and nouns, describing the events in the novel and the things involved in the events; a more flowery Victorian domestic novel would use more adjectives, adverbs, and perhaps prepositions. An introspective novel would involve many words relating to emotions and feelings, while a detective potboiler might not. And, of course, every author has an individual writing style: think of Hemingway, his direct and unadorned style, and his avoidance of Victorian descriptive tropes. (As Hemingway said of his own work, "What many another writer would be content to leave in massive proportions, I polish into a tiny gem.")

While in theory we could analyze any of thousands of these dimensions[4], our proof-of-concept focuses on four specific elements of writing style:

- Authorial Vocabulary (Words): Words are, of course, what the work is fundamentally all about. A crime novel is usually about a dead body and how people deal with the problem it poses; a romance novel is about a small group of people and their feelings for each other. Authorial vocabulary is also one of the best ways to tell individual writers apart, by looking at the choices they make, not only in the concepts they try to express, but the specific words they use to create their own individual expression.

- Expressive Complexity (Word lengths): One key attribute of authors is, on the one hand, their complexity, and on the other, their readability. A precise author who uses exactly the specific word to every event -- "that's not a car, that's a Cadillac; that's not a cat, but a tabby" -- will more or less be forced to use rarer words. These rarer words, by their very nature, are longer[5]. A large and complex vocabulary will naturally be reflected in longer words, producing a very distinctive style of writing. By tracking the distribution of word lengths, we can assess the expressive complexity of a given author.

---

[2] Patrick Juola. Stylometry and Immigration: A case study. *Journal of Law and Policy*, XXI(2):287–298, 2013; Patrick Juola. The Rowling case: A proposed standard protocol for authorship attribution. *Digital Humanities*, 2016.

[3] Patrick Juola. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3), 2006.

[4] Patrick Juola. 20,000 ways not to do authorship attribution and a few that work. In *Proceedings of 2009 Biennial Conference of the International Association of Forensic Linguists* (IAFL-09), Amsterdam, 2009; Patrick Juola. Large-scale experiments in authorship attribution. *English Studies*, 93(3):275–283, May 2012.

[5] Augustus de Morgan. Letter to Rev. Heald 18/08/1851. In Sophia Elizabeth De Morgan (Ed.) *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters*, 1851/ 1882; Zipf, George Kingsley. *Human behaviour and the principal of least effort.* Cambridge, MA: Addison-Wesley. 1949.

- Grammar (Part of Speech n-grams): Just as an author chooses the words to write about the events in the manuscript, so does the author choose the grammar of the sentences. This grammar is reflected in the parts-of-speech chosen, which in turn reflect the relationships described in the text. Simple action relationships (X did something to Y) can be reflected in simple transitive sentences, but complex mental structures (A believed that X did something to Y) require more complex grammar such as prepositional phrases, phrasal complements, and other complexities. To capture these relationships, we look at groups (formally speaking, "n-grams") of these parts-of-speech to assess the preferred grammar of the author, and by extension, another important aspect of the writing style.

- Tonal Quality (Function words): One of the most telling aspects of an individual writer is their use of function words[6], the simple, short, common, and almost meaningless words that form a substantial fraction of English writing. (To understand how these words lack meaning, consider that the three most common words in English are "the," "a/an", and "of." What would you offer as a dictionary definition of "the"?) These words are called "function words" because they do not carry meaning of their own, but instead describe the functions of the other words in the sentence in relation to each other. These words thus provide a good indication of the tone of the writing and the specific types of relationships expressed throughout the manuscript.

As discussed above, our system assesses each manuscript against a customizable database of manuscripts by established authors, in order to address each individual axis and find the closest match. We can thus find the single closest, the top 3, 5, 10, or even rank the entire database in terms of similarity. This allows you to look at as much or as little information as you need to make an informed decision.

Putting It All Together

To get an overall score, we calculate the total rank sum of each author along all four dimensions. For example, if Pegasus (by Danielle Steel) were the single closest book in terms of authorial vocabulary, vocabulary complexity, and tone, but the third closest book (behind two others) in grammar and the tenth (behind nine others) in grammatical complexity, the overall rank sum score for Pegasus would be 1 + 1 + 1 + 10 or 13. This is very likely to be the lowest overall score, but if another book had scored second place on all four axes, that second book would score 2 + 2 + 2 + 2 or 8. Of course, what this really means is that (in our hypothetical) both Pegasus and the unnamed book are very close to our manuscript.

So how did Pegasus do, sales-wise? Because our new book is very close in all aspects of style to this book, we expect our new book to have a similar level of potential sales, assuming a similar level and type of marketing effort on the part of the publisher.

How does this help the author? Writers are constantly revising their work and seeking advice from friends and colleagues. This feedback contributes to the process of self-editing and drives many decisions that writers make as they perfect their works. One of the most common pieces of feedback writers get are opinions that compare their works to those of others. "Your work reminds of Great Expectations" or "You write a lot like Dan Brown," are just subjective opinions; but now, with our tools, analysis is available that helps to drill down to the essence of writing style that can actually support rendering these opinions *objectively*.

Which means that writers can now ask the question privately: "Who do I write like?" This is perhaps the most basic question that authors ask themselves as they seek to compare what they have written to

[6] J. F. Burrows. 'An Ocean Where Each Kind…': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5):309–21, 1989.

the commercial marketplace; and, that question, slightly turned, is one of the most common writing skill exercises typically assigned of high school and college literature students: e.g. “Write an essay in the style of Mark Twain” or some other author.

How does this help the publisher? It provides an easy counterpart to the “gut check” that is so important to an experienced reader and may provide another, objective, piece of information to the decision. To understand this, let’s imagine an acquisition editor who is on the fence about a particular book by an unknown author -- yes, it’s well-written, but it’s not clear whether it will sell well enough to justify picking it up.

Our system can tell the editor the closest books in our extensive library -- and let’s assume that these are five books (by four different authors), all of which sold more than 100,000 copies. This may provide evidence, that, with proper handling by the publisher, this new manuscript can also sell in the six figure range. This piece of information by itself may be enough to justify “green-lighting” the manuscript. Perhaps just knowing that the single closest book sold 250,000 copies is enough. If this is all the information wanted, the publisher need not read further.

On the other hand, sometimes one needs to know more, for example, to address skeptical colleagues at the meeting. This system can name as many similar books as desired, as well as the types of similarity. If it’s not enough to know that this manuscript is “like” a best-seller, this system tells you what kinds of similarity are present.

Conclusions

Assessing the potential of a manuscript is an important task for authors, agents, editors, and publishers. Our tools can provide these people with key information to address this task accurately, insightfully, quickly, and with confidence. By comparing a new manuscript along several different dimensions with our large library of books, the user can quickly learn what kind of manuscript it is as well as how other, similar, manuscripts have done in the market. More importantly, the user will also learn why the specific comparison book(s) were chosen. This will help not only with manuscript selection but with market selection and discovery as well.

Contact Information

If you are an author and have questions or comments, please email Stacy Clark: stacy@inkubate.com. If you are interested in partnership opportunities to offer Inkubate's analytic tools to authors or publishers, please email Don Seitz: don@inkubate.com.