Argument

Digital Humanties, as term used to signify a field of academic research, might trigger the impression that it is simply a hip buzzword, or merely a passing trend. The trend definetly isn’t going to fade out any time soon, since the digital age itself has ramained unrelenting since it’s dawn in the late 80’s / early 90’s. As for the other part of our binomial, the Humanities, to attempt to compress its rich history and tradition in paper’s argument would be mean either naivety or arrogance. What I can humbly observe is the self-evident truth that the Humanties cannot be possible divorced by the tehnical progress that occurs within societies. The classic, widely known example: Gutenberb’s press; it revolutionized at its time the capacity for text production and circulation. Pursuing this self-evident line of thought, we arrive at the conclusion that it fundamentaly changed how humanists structured their activity, since the text was always their fundamental work object.

It is also true that humanties evolved to be more specialized and branched out, to suit the needs of the increasing body of knowledge that it produced, also the demands for accuracy and veracity (a classic example is how ancient and medieval historians/chroniclers depicted events vs. the modus operandis of contemporary historians). This shift is itself indicative of a change of paradigm, from the generalist, Rennaisance scholar to the modern, highly specialized in his niche modern researcher. In this context, I’m stating that the Digital Humanities is itself a field of subfields, an umbrella term, similarly to how the Humanities, in their traditional understanding hold the same significance. Therefore, the adjective “digital” is a consequence of the Humanities continuing to exist in the current paradigm. It isn’t, by any means, a micro-niche made up by some pretentious researchers as a result of their unconventional experimentation (although such practices have their merits and have lead to unexpected breakthroughs across all fields of research).

As I’ve already stated, the text has always been the vehicle of thought within the humanist tradition. For other domains, text is also serves a paramount role, but it is secondary, an “escort” of knowledge and results. In STEM domains, the aim of research are real life objects and entities, the role of text is to provide theoretical frameworks and attest the results of empirical tests. In the humanities, the text itself is the main object of study. There are also connected domains that mainly study real life, palpable objects, such as archeology or numismatics, or human behavior, such as anthropology, neurosciences or social studies. Regardless of that, the humanist tradition lies on fundamental texts: The Bible, Homer’s Illiad and Odyssey, Plato’s Republic, Aristotle’s Poetica and so on.

Nowadays, information is generated and circulated at overwhelming rates. The internet is an ever-shifting environment where different media coexist and combine. Patrik Svenson’s provides insights on this phenomenon:

One important and apparent consequence of increased digitalization and, in particular, the web, is highly increased access to and availability of different types of content and media. Some of this content is born-analogue and much of it is born-digital. Increasingly, but not necessarily, these expressions are media rich, polytextual and mixed. [Schnapp & Shanks 2009, 147] discuss "fungibility" — the gathering of many types of content (moving image, text, music, 3D-design, database, graphical detail, virtual walk-through etc.) into a single environment — as the core of digital mediation. Content can accordingly be infinitely manipulated and remobilized without loss.

[Svenson, Patrik. The Landscape of Digital Humanities in the Digital Humanties Quaterly, Providence Vol.4. 2010]

As shown by Svenson, content produced in the digital age is heterogeneous by nature. In contrast, established literary works, fiction or non-fiction, are just blocks of texts, and sometimes images. Them existing in digital format doesn’t mean a fundamental transformation at the content level. However, this format opens up the usage of digital tools and, therefore, new angles of viewing established works. Svenson (2010) provides his take on what the general aim of digital humanties should be:

The digital humanities comprise the study of what happens at the intersection of computing tools with cultural artefacts of all kinds. This study begins where basic familiarity with standard software ends. It probes how these common tools may be used to make new knowledge from our cultural inheritance and from the contemporary world. [Ibidem]

What I focus on in this paper are textual cultural aretfacts: literary texts that I am familiar with. Although it was theorized by various scholars and given multiple defintions, distant reading, as a general, collective term, encompasses the usage of various digital tools applied to literary corpora. My research aims to use distant reading with aim of tracing and comparing the progression of established authors in the Sci-fi genre.

Introduction

1.1 BookNLP

The digital tool that I am employing is a Python-based NLP pipeline suggestively named BookNLP. It was devoloped by David Bamman. In his own words:

BookNLP is a natural language processing pipeline that scales to books and other long documents (in English), including:

* Part-of-speech tagging
* Dependency parsing
* Entity recognition
* Character name clustering (e.g., "Tom", "Tom Sawyer", "Mr. Sawyer", "Thomas Sawyer" -> TOM\_SAWYER) and coreference resolution
* Quotation speaker identification
* Supersense tagging (e.g., "animal", "artifact", "body", "cognition", etc.)
* Event tagging
* Referential gender inference (TOM\_SAWYER -> he/him/his)

[Bamman, David, BookNLP Github repository, 2022 <https://github.com/booknlp/booknlp>]

Citations

BookNLP is a new Python library created by David Bamman. It was originally created as a Java library in 2014 under the same name, BookNLP by David Bamman, Ted Underwood, and Noah Smith (see, David Bamman, Ted Underwood and Noah Smith, “A Bayesian Mixed Effects Model of Literary Character,” ACL 2014). While Java is a powerful coding language, both in speed and ease-of-use, not many digital humanists code in Java primarily. I suspect (I want to emphasize I could be wrong) the reason for the Python library was to address the larger Python-coding community both in general and specifically within the digital humanities.

[Mattingly, William. *Introduction to BookNLP*, 2022. [booknlp.pythonhumanities.com](https://booknlp.pythonhumanities.com/booknlp.pythonhumanities.com)]

Both the documentation and this textbook emphasize the word *large* here. The reason? Because most language models do not perform well with larger documents. Old RNN-based language models had a hard time remembering earlier words and while newer transformer-based models, such as BERT, have a larger memory and can look forwards and backwards, the size of the input they can take in is only 512 words. For larger documents, therefore, different solutions (and libraries) should be considered. This is where BookNLP comes in. It also addresses several problems associated with books and larger documents, such as:

* Characters (and people) are referenced by different names. BookNLP solves this problem with name clustering and coreference resolution. This is a task in NLP where we try and find all uses a name and correctly assign them to the same identifier, such as Harry, Harry Potter, and Mr. Harry Potter all being the same person, Harry Potter.
* An adjacent problem is referential gender inferencing. Like coreference resolution, often times in a book or larger document, a person will be referred to as a pronoun. This is where referential gender inferencing comes in. This allows a user to correctly assign the antecedent or postcedent to the correct pronoun. When done successfully, this also allows you to make decisions about the gender of the character or person based on how they are referenced in the text. Because this task is so delicate, given the delicate nature of assigning gender, BookNLP fortunately gives users the data with each pronoun used to reference a character and also includes non-binary pronouns.
* Another issue is quotation speaker identification. This is when we need to understand who is speaking, so that we can correctly link characters to their dialogues. It is possible to do this with spaCy, but it is extremely difficult to do well. BookNLP does a remarkable job of handling this problem and it does it with a fair degree of accuracy, from what I have seen.
* Event tagging is another key issue with longer documents and books. There are machine learning models that find events and you can easily cultivate a list of domain-specific events to improve a pipeline, but for BookNLP event is defined more broadly. From my experience, it is more based around key actions, rather than named events (as it is in named entity recognition). This has a tangential benefit known as triple extraction. In my opinion, it might be a bit better to view BookNLP events through this lens. Triple extraction is when we try and extract three pieces of information, such as (Actor, Action, Recipient) or (Actor, IS, Something). With these types of tuples, we can construct a knowledge tree about a corpus fairly easily. This a very challenging problem in NLP because triple extraction can be very domain-specific. BookNLP provides a great starting place for triple extraction with its events.

[Mattingly, William. *Introduction to BookNLP*, 2022. [booknlp.pythonhumanities.com](https://booknlp.pythonhumanities.com/booknlp.pythonhumanities.com)]

Treating literary works themselves as networks, however, poses distinct computational challanges. While research into information propagation in social media tends to presume access to explicit networks, the character networks represented in novels are implicit.

[Sims, Matthew; Bamman, David, Measuring Information Propagation in Literary Social Networks].

Our goal in this work is to investigate the behaviour of information propagation in literary texts. In order to identify acts of propagation in this context, we need to determine the underlying network structure of a novel, including the nodes (by infering characters) and the edges (by inferring some interactions between them).