How Scholars Quote from Primary Texts: a Case for Corpus-Adapted Text-Reuse Detection

# # Abstract

Using text-reuse detection, we develop a tool for analyzing quotations from primary texts in a body of scholarly writing. While text-reuse detection has been increasingly used within computational literary studies and adjacent computational fields, less attention has been paid to the particular challenges posed by algorithmic detection of quotations from primary sources. We use our tool to analyze quotations from a source text (George Eliot’s \*Middlemarch\*) within a collection of JSTOR-held scholarly writing in order to demonstrate the need for corpus-adapted text-reuse detection––by which we mean methods of text-reuse detection tuned to the text structures of specific corpora. Our methodology traces not only how and where a text has been quoted, but also enables a more granular analysis of what parts have been quoted and how this has developed over time. Bringing text-reuse detection together with rich bibliographic metadata, we showcase the strengths of this local, more corpus-specific method for identifying quotations.

# # Introduction

Quotations, acts of plagiarism, instances of repeating gene sequences–these represent just a small fraction of the different kinds of objects that a researcher might be after when they look for instances of text reuse. Within the context of computational scholarship, text-reuse detection––or the algorithmic identification of passages that appear in two sets of text––has benefited from its ecumenical approach to texts. The ability to trace the reappearance of a portion of a text in another text is useful for understanding many humanistic domains that cite and use primary sources, from law to historiography, as well as many realms in genomics that deal with linear sequences of DNA. Such methods offer a wide range of disciplines the ability to trace instances of textual borrowing, replication or reuse.

Despite being applicable to a very broad range of objects, text-reuse detection as a field often answers a much narrower kind of research question: is there a possibility that portions of one text appear in another (often for the purposes of determining the possibility of plagiarism or textual borrowing). But this assumption, and the assumption that the same methods can be used for any text or corpus, are not well suited to all forms of text reuse––particularly when studying text reuse for purposes other than detecting covert borrowing or plagiarism. Quotation, and quotation in scholarly writings in particular, is a kind of text reuse that is not secretive or concealed, and has distinctive conventions for signaling and demarcating reused portions of text, as well as norms around what properly counts as quotation in contrast to paraphrase (or plagiarism). Text-reuse detection methods developed to identify everything from genetic similarities to plagiarized computer code will have only limited success in detecting the distinctive text reuse that forms the bedrock of much scholarly writing.

Quotation is particularly central to literary scholarship, where well-established disciplinary practices of close reading and textual exegesis give quotations from primary texts a particular role in literary argumentation. In this case, detecting what counts as a true quotation from a primary text requires a more local and corpus-adapted form of text-reuse detection with a shift in perspective. Instead of asking \*if\* there’s a possibility that one text includes text, our focus shifts to what parts of a text (often a primary text) appear in another. While our case study is literary scholarship, the methods we outline for local and domain-specific text reuse could apply (no doubt with some modifications) to the conventions around quotation of sources in other disciplines across the humanities and social sciences. Our central argument is that using text-reuse detection to study scholarly quotations can be improved through methods that take into account norms and conventions of scholarly writing, and in turn will allow scholars to ask more nuanced questions relevant to their discipline and objects of study. Our preliminary findings, which analyze matches between George Eliot’s \*Middlemarch\* and thousands of critical articles from JSTOR, show the benefits of corpus-adapted methods in linking computational text analysis with disciplinary history.

# # Background

Computational text-reuse detection has benefited from development in several specific domains. Plagiarism detection, a special use-case of text-reuse detection, was primarily developed to solve, as one 1981 work puts it, "the problem of plagiarism in programming assignments by students in computer science courses" [@donaldson1981]. A computer science programming assignment, however, differs greatly from published scholarly writing. For one, whitespace and punctuation are significant. In addition, plagiarism detection typically looks either at whole-document similarity or locally replicated syntax (since variables can be renamed in code, or words can be replaced in a paper using a thesaurus). There have been several useful overviews of the state of the plagiarism detection problem, a long-standing problem in natural language processing [@lukashenko2007].

Text-reuse detection has also been used in genomics and bioinformatics generally. The task of gene sequencing has led to the use of sequence alignment algorithms, which have undergone various developments over the years [@heng2010]. Given two strings of genetic sequences comprising cytosine [C], guanine [G], adenine [A] and/or thymine [T], the algorithm aligns them, such that the common parts of the sequence have identical indices.

One of the most well-known gene sequence alignment algorithms is BLAST (Basic Local Alignment Search Tool) from 1990 [@altschul1990basic]. As with plagiarism detection, this problem differs greatly from that of scholarly quotation, in that the possible elements are completely consistent and few in number (A|C|G|T). With quotation from literary sources, the elements to be sequenced potentially include every word in a language. Moreover, verb and noun forms may be changed in order to fit the scholar’s syntax ("he was 'descend[ing] down the staircase'"), variant spellings may appear (because of British vs American English, because of historical vs modernized spellings) and, since human memory is involved, even misquotations sometimes make it into print.

In the early 2000s, methods for studying text-reuse began to make their way from plagiarism detection and bio-informatics into computational linguistics. Some of the early methods were oriented around tracking the reuse of newswire stories from the AP or Reuters in other newspapers[@CloughMETERMEasuringTExt2002; @SeoLocaltextreuse2008; @BarTextReuseDetection2012] or the study of plagiarism communities[@KhritankovDiscoveringtextreuse2015], while others developed methods for detecting repeated sequences of text incidentally (as part of generating hyperlinks between different parts of the Google Books corpus.[@KolakGeneratingLinksMining2008]

Text-reuse detection has, more recently, been used by scholars to identify communities produced by the reprinting of texts, particularly in newspapers and periodicals. In their work on the *Viral Texts Project*, Ryan Cordell, David Smith, and Elizabeth Maddox Dillon have shown how reprinted poems, ads, and short news articles in 18th-and 19th-century American newspript can illuminate patterns in newspaper syndication.[Project [@SmithInfectioustextsModeling2013; @SmithDetectingModelingLocal2014] Their work has paved the way for other scholarship on reprinting, including newspapers in other languages,[@VesantoApplyingBLASTText2017; and @VesantoSystemIdentifyingExploring2017 both apply the BLAST algorithm to Finnish newspapers in the 18th-19th century] and towards different aims, like tracing the recirculation networks of newswire copy in the corpora of 21st-century news articles from the US and the UK.[@NichollsDetectingTextualReuse2019] In a similar vein as text-reuse in literary studies, computational linguistics have used similar methods in examining how the text of a given policy document is reprinted in other bills and legislative texts. [@LinderTextPolicyMeasuring2020] Here, while occasional attention is paid to the particular texts that “go viral,” the focus is not on particular passages but on the networks of publishers, syndicators, and editors that such acts of reprinting reveal.

Early examples of literary scholarship that used text-reuse detection were modeled after these gene-sequencing and plagiarism detection methods and, as a consequence, text-reuse methods were shaped around a particular problem: identifying any instances of potential similarity between two collections of texts. Scholars were chiefly interested in the potential evidence of literary borrowing, using BLAST algorithms to find evidence that eighteenth-century reference texts borrowed passages from Diderot's and D’Alembert’s encyclopedias,[@OlsenSomethingBorrowedSequence2011] or whether 19th-century French writers like Balzac and Gautier plagiarized from one another.[@GanasciaAutomaticdetectionreuses2014]

When computational studies have focused on quotation, they have typically focused on quotation of passages from canonical source texts (often, texts from Greek and Roman antiquity or the Hebrew or Christian Bible) in other non-scholarly texts such as newspapers or novels.[@GessnerBiblicalintertextualitydigital2013; cite Jonathan’s work on Bible quotations] Lincoln Mullen, in *America’s Public Bible*, focuses on quotations from the Old and New Testaments of the Bible in a corpus of American 18th- and 19th-century newspapers.[@LincolnAmericaPublicBible; @QuotationFinderAmerica2022] Marco Büchler’s TRACER project offers something of a hybrid between the study of particular quotations and their circulation: TRACER focuses on reprinting of Classical Greek authors within a corpus of Greek historian’s reference texts and takes frequent quotation as a metric for a work’s “influence.”[@BuchlerMeasuringInfluenceWork2013; @KokkinakisDetectingReuseBiblical2016; see also Frederik Arnold's “Lotte” (later renamed “Quid”) framework[@ArnoldLottev12022; @ArnoldQuid2022; @ArnoldLotteAnnetteFramework] Detecting quotations within primary texts poses particular challenges: what edition and translation of the Bible to use, or how to deal with slight misquotations[@DuhaimeTextualReuseEighteenth2016; @RoeDiggingECCOIdentifying2016] Recent work has also begun to extend text-reuse methods to non-Latin scripts[@SturgeonUnsupervisedidentificationtext2018; @Budakdirectphonologydphon2021; @Budakdphon2022]

A related but distinct area of scholarship to the study of text reuse is the study of *\*citations\**, which often goes by the name “bibliometrics.” Whereas quotation involves the verbatim replication of some portion of the source text, citation involves only a reference. In the case of scholarly writings, this is usually a bibliographical reference formatted according to a citation style determined by the publisher. For example, to *\*cite\** Michel Foucault’s concept of the “author-function” or the article “What is An Author?” from which it derives is distinct from quoting his specific formulations within the article. Citations have particular significance for scholars within the contemporary university, where they are a widely-used metric that shape career opportunities. Many decades of research have demonstrated that the distribution of citations in academia replicates and reinforces broader inequalities related to social categories like gender and race. [See e.g. @FerberCitationsAreThey1986; @EarhartCitationalPoliticsQuantifying2021; @CiteBlackWomenCollectiveCiteBlackWomen] Focused as they are on the social status of the cited works’ authors, these studies don’t typically examine the smaller scale of which portions of a work are cited. [See @RomanelloExploringCitationNetworks2016 for an exception, which also analyzes the book and line numbers included in citations of classical texts like Vergil’s \*Georgics\*.] Methodologically, all these studies function not by detecting text reuse but by tallying bibliographical references, which in scholarly writing have a relatively consistent form (e.g. in-line citations; footnotes; endnotes; list of works cited). In contrast, our method detects instances of text reuse, whether accompanied by a bibliographical reference or not, and doesn’t detect bare citations without quotation. While bare citation is the norm in certain natural sciences, across the humanities and in literary studies in particular, quoting from and citing a source are both options, whose dynamics we want to understand better. It may be, for example, that who scholars choose to quote, as opposed to merely cite, further intensifies the social inequalities already well-documented in citation patterns. [^01]

[^01]: Literary scholars have also studied quotation practices using various non-digital methods. Studies focusing on the presence of quotations in literary works include @MeyerPoeticsQuotationEuropean2015; @CompagnonSecondeMainOu1979; @PrinsVictorianSappho1999; @BuurmaEpigraphsMatter2012; and @HackReapingSomethingNew2016. Studies of how portions of literary texts have circulated in the form of quotations include @GarberQuotationMarks1999; @PriceAnthologyRiseNovel2000; @DamesNotCloseReading2010a; and @MoleWhatVictoriansMade2017. Closest to our own project, there have been several recent (non-computational) studies examining how *\*literary scholars\** quote: @AuyoungWhatWeMean2020 and @KramnickCriticismTruth2020.

Our text-matcher builds on these existing bodies of work by developing methods geared towards a narrower domain: detecting quotations from primary texts used in scholarly writing. Text-reuse detection in the contexts of plagiarism or networks of reprinting cast a wide net (i.e. set the threshold for detecting reuse lower) in order to capture as many potential instances as possible, which can then be further investigated manually as needed. In contrast, our methodology sets the threshold higher, since scholarly norms presuppose precise replication in quoted texts, with only limited variants permissible. In focusing on the quotation of primary texts within scholarship, we build a tool with the aim of identifying quotations in corpora (scholarly writings) that have more standard conventions for direct quotation. In doing so, we take inspiration from several recent digital projects studying scholarly quotation patterns in aggregate. JSTOR Labs and Derek Miller have analyzed quotations from literary texts in scholarly writings: both projects present interactive web visualizations for exploring what lines have been most quoted within Shakespeare’s corpus.[@HumphreysHowJSTORLabs2017;@MillerQuoteNotQuote] Our text-matcher has been used in the study of other corpora, most notably [Our text matcher has been used in @PiperMeasuringUnreading2020 to study quotations of Goethe’s corpus; related work on this topic was presented by the Stanford Literary Lab at the 2016 Modern Languages Association conference.]

^[While we were conducting these experiments, and presenting our initial findings at Digital Humanities 2017, we learned that the Stanford Literary Lab was working on a strikingly similar problem: text matching between a large corpus of literary texts, and a corpus of historical book reviews and critical writings from the British Periodicals Online collection. The Stanford Literary Lab independently arrived at many of the same parameters we use for text matching, even using the same Python library.]

# # Methods

We begin by pre-processing a text into a sequence of tokens (roughly, words). We strip out punctuation and extraneous whitespace, and lowercase the text. We also concatenate hyphenated words, in order to match against words which have been hyphenated due to coincidental line breaks in the scholarly publication. We also remove the NLTK’s standard stopword list, consisting of the most commonly-occurring words such as “of” and “the.” [^02] From there, we convert the tokens into stems––for example, “photography” and “photographer” both become the stem “photog”–– using the Lancaster stemmer of the NLTK, which uses the Paice-Husk stemming algorithm.[cite:@nltkLancaster;@paice1990] We chose this stemmer, after evaluating several other options, because the stems it produces retain a recognizable word form while collapsing small variations in noun and verb endings. Finally, we group these stems into n-grams---three-token sequences, or trigrams, by default---allowing us to alleviate the computational work required of the SequenceMatcher.

[^02]: Notably, this pre-processing removes quotation marks, which could offer another method for detecting quotations. However, we have not relied on this method since automated parsers are not very accurate at identifying material inside quotation marks. In addition, much of the scholarship in JSTOR’s corpus, and in searchable databases of scholarly writings in general, has been digitized through Optical Character Recognition (OCR), which can both mis-recognize actual quotation marks and erroneously insert them.

The core matching algorithm we use is the SequenceMatcher from the Python library ~difflib~ [cite:@peters\_difflib\_2016], which adapts Ratcliff/Obershelp Pattern Recognition, or "gestalt pattern matching" [cite:@ratcliff1988pattern]. This computes the string similarity $D\_{ro}$ of strings $S\_1$ and $S\_2$ according to their matching tokens $K\_m$.

$$ D\_{ro} = \frac{2K\_m}{|S\_1|+|S\_2|} $$

This computes the initial, or core matches. By default, we search for a core match of length three, `--threshold=3` in the command line interface. This amounts to three *\*overlapping\** trigrams---a total of five words. Thus, we define the minimal quotation as five identical stems. We arrive at this set of defaults after a long period of experimentation: too little of a threshold, we discovered, results in too many false positives, since there are many common three- and four-word sequences which are set phrases of the English language, rather than evidence of quotation. This core matching process must involve identical stems---given the origins of the Ratcliff/Obershelp algorithm in bioinformatics – which gives higher initial performance compared to a fuzzier matching procedure.

However, the initial threshold only constitutes the first step of the matching operation. From there, we perform two additional functions, both of which are much slower, computationally. The first is to extend the match to contiguous words which may also be part of the same quotation. To do that, the algorithm looks [one word/token?] backwards and forwards from the boundaries of our initial match, and compares the words at these boundaries in both the source text and scholarly text. To compare these words, we use Levenschtein distance, or edit distance, which describes the number of insertions, deletions, or substitutions required to edit one string into another. [cite:@levenshtein1966binary] When adjusted for the number of letters in each word, we can approximate a word's morphological similarity to another. So long as two words have an edit ratio of 0.4 or below, we consider them part of the same quotation. This allows us to handle differences in American and British spelling, as well as some OCR errors. Considering these example edit ratios:

| Word A | Word B | Edit ratio |

|---------+----------+------------|

| color | colour | 0.1818 |

| theater | theatre | 0.2857 |

| day | today | 0.5 |

| foobar | foo56bar | 0.2857 |

The edit ratios of the words with divergent British and American spellings are below the threshold and are considered matches; the similar words /day/ and /today/ are above the threshold and aren't considered matches. The word with OCR errors, /foo56bar/, is still considered a match.

Even with this additional step, however, we noticed that the algorithm was breaking off before achieving a complete match, for example when the number of OCR errors surpassed this threshold, or when the scholarly text was interrupted by paratextual features, like running headers or page numbers. To mitigate these issues, we heal neighboring matches. If two matches are within eight tokens of each other, we concatenate them into the same match. This allows us to match strings across page breaks, matching, for instance, a string such as "hearing the grass grow and the squirrel's heart beat" with "hearing the grass GEORGE ELIOT--GEORGE HENRY LEWES STUDIES 54 and the squirrel's heart beat, and we should die of that roar which lies on the other side of silence." Without this healing of neighboring matches, the match on the first page would be smaller than the minimum match size.

Because scholarly writings held in databases such as JSTOR are accompanied by detailed metadata relevant to scholars (e.g. publication date, journal/book title, page ranges), we designed our tool to integrate this metadata. In particular, metadata on publication dates allows us to analyze not only where in the text a particular quotation appeared (determined through a simple text processing to determine the index characters of each chapter in our sample text), but how quotation patterns varied over time within our corpus (in our case, a collection of 20th and 21st century scholarship from JSTOR). Extracting the dates from the metadata, we developed a method for diachronic analysis: analyzing what passages were quoted at different moments over time. This diachronic analysis allows the detection of quotation patterns to open out onto questions of disciplinary history.

# # Results and discussion

What our results show is the importance of domain specific text-reuse detection methods. Our algorithm and workflow is designed for a specific case and tuned with hyperparameters that are sensitive to the nature of a quotation in a scholarly text, offering a more nuanced method for humanities-specific applications.

The modifications of our algorithm are designed to capture particular ways that scholarly texts quote their sources. Take, for instance, this quotation that our healing of neighboring matches was able to correctly identify:

> PASSAGE IN MIDDLEMARCH: Bulstrode's sickly body, shattered by the agitations he had gone through since the last evening

> PASSAGE IN SCHOLARLY ARTICLE: Bulstrode's "sickly body" is "shattered by the agitations he had gone through since the last evening"[@CarpenterMEDICALCOSMOPOLITANISMMIDDLEMARCH2010, 521]

Using healing, we’re able to account for some of the ways that scholars weave direct quotations from other texts into their own, modifying the direct quotation to fit the syntax, style, mood, and tense of their own sentences.[@KramnickCriticismTruth2020, 223] In other instances, our fuzzy matching parameters and our healing were able to heal across instance of hyphenation (“ac-commodate” is correctly identified as one word).[@MilletUnionMissBrooke1980, 40] The tool even detects text that was not presented as part the quotation but had been naturalized into the scholar’s own language:

> PASSAGE IN MIDDLEMARCH: "growing good of the world is partly dependent on unhistoric acts; and that things are not so ill with you and me as they might have been, is half owing to the number who lived faithfully a hidden life, and rest in unvisited tombs.

> PASSAGE IN SCHOLARLY ARTICLE: The narrator asserts that "the growing good of the world is partly dependent on unhistoric acts." If things are "not so ill ... as they might have been," Eliot concludes, it is "half owing to the number who lived faithfully a hidden life, and rest in unvisited tombs" [@MECKIERThatArduousInvention1987, 228, text reuse outside quotation marks underlined.]

Because our text-matching algorithm has been designed to be customizable, all of the hyperparameters can be changed and customized depending on the given corpus, providing more opportunity for fine-tuning. In our case study, we’ve set our parameters to be fairly strict. Our text-matching algorithm makes intentionally conservative matches in order to avoid mis-identifying a commonly-used string of text as a quotation. This means that we have extremely high precision (100% of the quotations our text matcher identified in a random, human-verified sample were in fact quotations from our source text). But setting the precision so high also comes with some tradeoffs: recall was only 30%, meaning the tool failed to detect 70% of quotations that are 5+ words in length. This is in part due to the hypersensitivity of our parameters: because our matcher is only set to determine a match if three matching overlapping tri-grams are found, in practice, the smallest sequence of text that we can identify is a string of 5 words long. We’ve set this parameter high because even if we were to match shorter strings, there would be no guarantee that the matched text is, in fact, a string from *Middlemarch*, and not a commonly occurring 1-2-, 3-, or 4-gram. Even removing stopwords would not be a solution, since some of the quotations that critics quote from the novel, like the use of the phrase “you and me” in an epigraph––contain only stopwords. Our analysis shows that detecting quotations under 5 words in length is a more complicated task than simply solving it might imply. Instead, as this method advocates, detecting smaller n-grams, that cannot be solved by a “better” matching algorithm, since such an algorithm presumes the uniqueness of the quotation being matched––-the fact that that quote would appear only in that text––rather than acknowledging that commonly recurring 1- 2-, 3-, 4-grams constitute part of the text of a work that a literary critic might quote in the process of close reading.^[Within our sample dataset for verification, we have a number of fewer than 5 word quotations (the number of which are requotation). Creating probability tables to determine the probability of a given word in a general corpus could help capture some of the shorter n-grams, but would miss highly common n-grams like “you and me” which is a quotation in one article from \*Middlemarch\*] Other researchers––depending on their corpus and research questions–– may be interested in lowering the precision score in order to achieve higher recall.

As previously stated, the experiment for which we developed text-matcher was an analysis of critical quotations of George Eliot's novel \*Middlemarch\*. We were interested in knowing how \*Middlemarch\* was quoted over time, and whether there were patterns to these quotations. With the gracious assistance of our friends at JSTOR Labs, we were able to obtain full texts of over 4,000 critical articles which contain the word "Middlemarch." (The specificity of this novel's title helped us greatly in this query.)

Since text-matcher keeps track of the locations of matches, we are able to answer a number of useful questions about trends in the critical quotation of \*Middlemarch\*. One of our early motivating research questions was, which parts of the novel are quoted the most, and which are quoted the least? After running the text-matcher on our corpus of JSTOR critical articles, we were able to determine the most frequently quoted passages, and categorize these according to chapter, the most meaningful narrative unit for our study. @Fig:byChapter shows these numbers of quotations, according to chapter. The most quoted chapter, in number of quotations, is Chapter 20. This holds true as measured by number of quoted words, as well.

![Number of Quotations of \*Middlemarch\* by chapter](byChapter.jpg){#fig:byChapter}

Next, we set out to discover the histories of these quotations, by comparing their numbers with the years in which the articles were published. @Fig:byChapterDiachronic shows the same breakdown of quotations per chapter, but shown across seven decades. What we found surprised us—with the exception of Chapter 20, which remains the perennial favorite among critics, many quoted passages came in and out of fashion in 20- to 30–year cycles. Certain chapters become more quoted for a time, and then gradually less quoted, until they are ignored altogether. One chapter in particular, the second-to-last, is cited hotly in the 1980s, and then never again.

![Number of Quotations of \*Middlemarch\* by chapter, 1950–2020](byChapter.jpg){#fig:byChapterDiachronic}

We discovered that Chapter 20 is the most-quoted chapter, due to large numbers of quotations of its fifth and sixth paragraphs. This led us to develop an even more granular visualization of \*Middlemarch\*'s critical quotations: we created an annotated edition of the novel, in which its passages are colored according to the relative number of critical quotations: darker colored passages are less-quoted, and lighter colored texts are more quoted. @Fig:annotated shows a screenshot from this annotated edition, showing the most-quoted passage in the novel: the sentence, "if we had a keen vision and feeling of all ordinary human life, it would be like hearing the grass grow and the squirrel's heart beat, and we should die of that roar which lies on the other side of silence."

^[The annotated edition is available here:

https://lit-mod-viz.github.io/middlemarch-critical-histories/annotated.html ]

![Annotated edition of \*Middlemarch\*, by numbers of quotations](byChapter.jpg){#fig:annotated}

Our inspiration for this annotated edition comes from JSTOR's \*Understanding\* series. At the outset of our project, JSTOR had provided editions of Shakespeare plays and the US Constitution, annotated according to the number of JSTOR articles. Now, there are many more texts available for exploration, even \*Middlemarch\*. But while JSTOR's editions provide paragraph-level counts of quotations, our analyses are accurate to the character level. This enabled us to find, for example, that of this most-quoted sentence above, its beginning, "if we had a," is quoted much less than the rest of the sentence.

[Keywords]

These findings led us to investigate whether there were any patterns to quoted phrases: is it just their poetic lyricism? We thus tallied all the words that appear in quotations, weighted them according to numbers of quotations, and compared these word frequencies to the words that never appear in quotations. The result is a list of words that are quotable: if these words appear in /Middlemarch/, they are likely to be quoted by critics.

Quotable words include, in order of quotability: \*life, like, woman, dorothea, love, world, soul, consciousness, little, sort, deep, live, nature,\* and \*history\*. This list contains one character's name, Dorothea, which correlates with frequent quotations; the rest are chiefly abstractions. This leads us to posit whether critics are drawn to passages that contain claims---especially claims which deal with abstractions such as \*life, love,\* or \*soul\*---since they provide an easier entry into literary argumentation than, say, a bare description. We also wonder whether readers seek out abstractions as a way of making sense of the rest of the novel.

On the opposite end are non-quotable words: words that make a pass less likely to be quoted: \*say, Mr., Fred, Bulstrode, Lydgate, Mary, Garth, Celia, James,\* and \*not\*. The presence of the lemma \*say\* in this list indicates to us that dialogue indicators, like "said Bulstrode," are seldom quoted, but rather paraphrased. Otherwise, this reads like a list of minor characters, whom critics seem less interested in discussing.

It turns out that both of these lists of words are corroborated by the work of the Stanford Literary Lab, who find the same patterns with the British Periodicals Online corpus.

# # Conclusion

In this article, we’ve shown the importance of domain-specific studies of text reuse, using our case study in literary criticism of a single primary text (George Eliot’s Middlemarch) and JSTOR. Our goal has been to provide a concrete pipeline and workflow for others to study quotations in primary sources and other corpora of secondary criticism, and to provide a clear pipeline for others using which is available for use on our GitHub repository. While we continue to refine our matching algorithm––in particular, with the aim of matching more 5-word but less than 12-word strings, our goal has been to illustrate some of the parameter tuning that comes with detecting quotations within a humanities field where quotations compose a significant portion of literary texts. While our primary case study has been in literary studies, our aim has been to demonstrate the value of local and context-specific methods for studying what parts of texts are quoted, and to offer pathways for further work on different corpora.

# # Acknowledgements

Alex Gil, Dennis Tenen, Xpmethods group, other contributors to text-matcher