Information Technology Approaches to Literature Text Analysis

Ayse Tarhan, Mustafa İlkan, Mohammad Karimzadeh

Eastern Mediterranean University in Cyprus

Abstract. Science was considered as part of philosophy in ancient Greece. By the nineteenth century, it was understood that philosophy was very inclusive and that social and human sciences such as literature, history, and psychology should be separated and perceived as an autonomous branch of science. The computer was also first seen as a tool of mathematical science. Over time, computer science has grown by encompassing every area in which technology exists and its growth compelled the division of computer science into different disciplines, just as philosophy had been divided into different branches of science. Now there is almost no branch of science in which computers are not used. One of the newer autonomous disciplines of computer science is digital humanities and one of the areas of digital humanities is literature. The material of literature is words, and, thanks to the software tools created using computer programming languages, data that a literature researcher would need months to complete can be achieved quickly and objectively.

In this article, three different tools that literary researchers can use in their work will be introduced. These studies were created with the computer programming languages Python and R and brought to the world of literature. The purpose of introducing the aforementioned studies is to set an example for the development of special tools or programs on Ottoman language and literature in the future and to support such initiatives. The first example to be introduced is the Stylometry tool developed with the R language. The other is The Metrical Tool, which is used to measure data in poems and was developed with Python. The final literature analysis tool in this article is Voyant Tools, which is a multifunctional and easy-to-use tool.

Keywords: DH, Literature, Information Technologies, Stylometry, The Metrical Tool, Voyant Tools

Introduction

Every point where technology is used shows that it is actually the field of computer science. Therefore, information technology has a close relationship and interaction with the branches of science in five (5) different classifications given below. Following Ehrlich, 1991, p. 316), these are:

1. the physical sciences (computer science, mathematics, and physics);
2. the humanities (literature, foreign languages, linguistics, philosophy, history, and the fine arts);
3. the social sciences (psychology, cognitive science, sociology, anthropology, and economics);
4. professional and vocational programs (library and information science, education, journalism, the media);
5. interdisciplinary areas (technology and society, history of science, American studies, women's studies, third world studies, environmental studies).

The relationship between literature and computer science is close in terms of helping the computer to understand and research literature. Although many computer programming

languages help literature research, the most used programming languages are perhaps R and Python.

Python (http://www.python.org/) is a scripting language created in 1980 and was first released in 1991. The name “Python” was given after British comedy troupe Monty Python influenced Guido van Rossum of The National Research Institute for Mathematics and Computer Science (CWI). Python is the most preferred language because it is understandable and effective: Python uses statements and expressions; is a dynamic, object-oriented programming language; and has a useful structure. It offers a rich standard library, and its working structure has developed links to most of the popular libraries. Python's biggest advantage is its very expressive and readable syntax; as a result, it can achieve a lot with a very short code and fewer bugs (Grabar, 2009, pp. 80-145).

Advanced statistical programming languages such as R are used to do the stylometric analysis. The R programming language makes the work much easier and enables the analysis of very large texts. R is an open-access, free programming language for computation and graphics built for statistics and data analysis. It compiles and runs on UNIX, Windows, and MacOS. Anyone can run and develop the programming language. Therefore, the R programming language is constantly updated, and new features are added for all kinds of applications related to data processing. It has been developed for a wide range of applications ranging from simple statistical calculations such as mean, correlation and frequency analysis to multi-level modeling and artificial intelligence applications.

The study methods of this age, which we can call the computer age, are also changing rapidly in parallel with technology. The approach to literature with Information Technology methods has begun to adapt its own methodology to these fields, while encompassing all human sciences. The use of computer technology in literature and language studies has changed the filing methods that have been applied for years, and the data obtained as a result of long studies has been provided to the researcher in a few minutes. In the study, it is aimed to introduce the tools and programs created through the use of computer technologies as an example for the projects to be produced in the future.

Some Sample Literature Analysis Tools Created With IT Approaches

Stylometry with R, A Package for Computational Text Analysis

“Stylometry” is a portmanteau of the English words “style” and “metry” and is a tool that refers to the “metry of style.” English style means “style, type, form, format, fashion, elegance, variety, spindle, substance, technique, pen, pen tip.” Likewise, "metry" is used to mean “measure, measurement” in English. This tool is created using R, a programming language, to provide high-level analysis of the writing style. Hence the name “Stylometry with R.” Stylometry deals with the quantitative study of writing style. This tool can be used in historical research by verifying the authorship of copies, as well as an application with significant potential for examining evidence in forensic contexts or as proof of plagiarism. Stylometry with R was built in 2015 by Maciej Eder, Jan Rybicki, Mike Kestemont.

The purpose of the tool is to identify the authors of works whose author is unknown or attributed to someone else. If such a study can be adapted to Turkish or Ottoman, the original authors of the poems who have common pseudonyms and are attributed to each other can be determined. The creators of the tool tried to answer the debates about the attribution of a

pseudonymous work to Rowling, the well-known author of the *Harry Potter* series, and the publication of the original version of Harper Lee's *To Kill a Mockingbird*, and that its editor may have played a role here (see Eder, Rybicki, and Kestemont, 2016).

Prof. Dr. Fazlı Can (2015), on the Turkish novel, used this program in his work titled "The Change of Writing Style Over Time," and one of the works he cited there is "A Short Non-Quantitative Presentation in the Remembrance Meeting of Ali Teoman's 'The Writer Who Writes his Unwritten Writings.'" He tries to answer the claims about the author of a novel using this program.

In literary work analysis, stylometry does not start from a direct reading of the text; instead, it tries to explore large collections of text using computational techniques and visualizations. Typically, stylometry analysis is a complex, multi-stage system consisting of pre-processing, feature extraction, statistical data identification, and the presentation of results through visualization.

A Metrical Tool for Greek and Latin Poetry

A dynamic programming language applied to the social sciences is required to analyze and develop texts or poems. Python represents one of the most advanced programming languages. Its relevance consists of versatile applications with libraries that cover and support many programming needs, not only to provide a collection of features that cover and support many programming needs and to demonstrate a very simple and consistent syntax, but also to provide a wide range of features. It has libraries that enable data processing for social sciences. Such libraries have grown significantly in recent years, with e.g. libraries such as NumPy, Jupiter, Matplotlib, Pygame, PySAL, Rpy and Python 3 for handling computation of large and multidimensional numerical data. The Classical Language Toolkit (CLTK) is a Python library that offers natural language processing (NLP) for pre-modern Eurasian languages and is built on nineteen languages.

The Metrical Tool uses the Python 3 library to analyze the stress of Ancient Greek and Latin words. This tool, developed in Python, can algorithmically scan Greek and Latin poetry and allows the user to search for metric patterns in a poem of more than 200,000 lines. The Metrical Tool enables researchers to find any given metrical poem pattern, as well as to explore the structure of metrics. The metric tool used a rules-based algorithm. It showed ninety-eight percent (98%) success in teaching measurement units to the computer. The algorithm assigned percentages to poetic syllables in Greek and Latin: first, when two syllables were joined, and second, whether a particular syllable was counted long or short. This brought the code to perfect (100%) accuracy for the two most prominent counters, at least based on repeated spot checks. The application methods of the tool are as follows:

1. It represents – the long syllables in the poem. Short syllables also meet the ∨ sign.
2. There are common patterns in poetry, and the tool gives the researcher patterns that may be suitable for his poetry: Here is an example: –∨ ∨ |– –
3. " ⁀ " should be chosen to avoid a break between syllables and "," to pause.

Poems written in Ottoman Turkish also have the *aruz* unit of measure, which comes from the Arabic tradition. The smallest unit of the poem is the couplet, and it consists of two lines. The verses are written with the units of measurement formed by the combination of short and long syllables, just like in Greek poetry. Like Greek poetry, Ottoman poetry has two structures: short

syllables and long syllables. One of them is “Fâ‘ilâtün / Fâ‘ilâtün / Fâ‘ilâtün / Fâ‘ilün” pattern, and the following signs represent the pattern in terms of short and long syllables:

— ∨ — — | — ∨ — — | — ∨ — — | — ∨ —

Voyant Tools

Voyant Tools has been created in accordance with the English grammar structure to understand the style of Jane Austen and William Shakespeare. Voyant Tools is open access and provides information on how to use it for researchers. Voyant Tools is a web-based text reading and analysis program. Text is a project designed to make it easier for you to work on your texts in a variety of formats, including HTML, XML, PDF, RTF, and MS Word, and to facilitate reading and interpretation practices for social science students and scholars, as well as the general public.

It is the easiest-to-use tool for learning how computer-aided analysis works. One can analyze online collections, magazines, blogs or websites, as well as analyze one’s own texts, and registering an account is not necessary to use Voyant Tools. As they tell their users, Voyant Tools can be used for the following tasks:

- You can use it to learn how computer aided analysis works.
- You can use it to add functionality to your online collections, magazines, blogs or websites so others can see behind your texts with analytical tools.
- You can use it to add interactive evidence to your articles that you publish online. You can add interactive panels directly to your research papers so your readers can summarize your results.
- You can use our functionality and code to develop your own tools.

Voyant Tools’ code is available via GitHub. The code is under a GPL3 license, and the content of the web app is under a Creative Commons by Attribution license. Describing their work as a labor of love, the project team states that the ancestors of the tool are HyperPo and Taporware, and more distantly, TACT. The coordinators of the project, which started in 2016, are Stéfan Sinclair from McGill University and Geoffrey Rockwell from the University of Alberta.

Voyant Tools' user interface is available in several languages. By default, Voyant Tools will detect the browser's language preferences and present the interface in the first available language (or English if no other language is available). Language options on the vehicle include Arabic, Bosnian, Croatian, Czech, English, French, Hebrew, Italian, Japanese, and Serbian, unfortunately Turkish is not available.

Voyant Tools' review visualizations are available under the following headings: Bubblelines, Bubbles, Cirrus, Collocates Graph, Corpus Collocates, Correlations, Document Terms, Documents, Knots, Mandala, Microsearch, Phrases, Reader, Scatterplot, StreamGraph, Summary, Terms Radio, TextualArc, Topics, Trends, Veliza, WordTree.

Conclusion

The claims that many novels or poems belong to famous writers or poets, such as the attribution of a work to the famous *Harry Potter* novelist Rowling, have tired the minds of literary researchers for centuries, and they have carried out studies on it. However, since these studies

are not based on objective data, they have taken their place in the scientific world with question marks. Stylo, also known as Stylometry, is a software produced with R, one of the computer programming languages, and it is a research tool that can calculate who the book or poem belongs to, and show the probability with its data, which has been tiring their minds for a long time. Therefore, the R programming language is useful in producing software that can facilitate the operations of literature researchers.

Python is a very useful tool that can eliminate the problems arising from literature researchers with its software. It is easier to present numerical data and make visualizations with Python. In this context, an exemplary tool created with Python in this study is The Metrical Tool, which can find the measures of Greek and Latin poems. In the article, both Stylo and The Metrical Tool were introduced and provided an example for tools and programs that could be made on Turkish and Ottoman Turkish in the future.

Voyant Tools is a computer aided language and literature review tool. Researchers can access many numerical data on the texts they have examined through Voyant Tools, as well as create visuals with these numerical data. Perhaps it is the best tool that Turkish researchers can use it for the stylistic analysis. The best results in Turkish or Ottoman can be reached through this tool.

References

'An Interdisciplinary Bibliography for Computers and the Humanities Courses'. *Computers and the Humanities* 25, no. 5 (October 1991): 315–26. https://doi.org/10.1007/BF00120968.

Brooker, Phillip D. *Programming with Python for Social Scientists*. SAGE, 2019. https://books.google.com.cy/books?id=Frq9DwAAQBAJ

Can, Fazlı, "Yazı Üslubunun Zaman İçinde Değişimi". 15.11.2021 http://www.cs.bilkent.edu.tr/~canf/libraryTalk2015.pdf

Can, Fazlı, "A short non-quantitative presentation in the remembrance meeting of Ali Teoman 'The Writer who Writes his Unwritten Writings'", 15.11.2021 http://www.cs.bilkent.edu.tr/~canf/libraryTalk2015.pdf

Classical Language Toolkit (CLTK) 21.11.2021. http://cltk.org/

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 'Stylometry with R: A Package for Computational Text Analysis'. *The R Journal* 8, no. 1 (2016): 107. https://doi.org/10.32614/RJ-2016-007.

Grabar, Darko. '07. 1957-2007: 50 Years of Higher Order Programming Languages' 33, no. 1 (2009): 72.

Mailund, Thomas. *Beginning Data Science in R*. Berkeley, CA: Apress, 2017. https://doi.org/10.1007/978-1-4842-2671-1.

Mason, Winter, Jennifer Wortman Vaughan, and Hanna Wallach. 'Computational Social Science and Social Computing'. *Machine Learning* 95, no. 3 (June 2014): 257–60. https://doi.org/10.1007/s10994-013-5426-8.

Python. 15.11.2021. (http://www.python.org/)

The Metrical Tool. 15.11.2021. http://206.207.50.59/about

Voyant Tools. 11.11.2021. https://voyant-tools.org