Using Machine Learning to Detect Plagiarism in Written Works

(Technical Paper)

Minimizing Gentrification in Tech Hubs

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree

Bachelor of Science, School of Engineering

Wonyoung Choi

Spring, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature ___________Wonyoung Choi__________________ Date __5/2/2022__

Wonyoung Choi

Approved ____ Date __________

Rosanne Vrugtman, Department of Computer Science

Approved ____ Date __________

Hannah Rogers, Department of Engineering and Society

Introduction

This prospectus will consist of two different research topics: one technical topic which consists of a machine learning algorithm used to detect plagiarism and one STS research topic which focuses on the effects of big technology companies on gentrification.

The technical project will be Plagiarism Detection using Machine Learning. Plagiarism is defined as “ a form of cheating that disregards academic values” (Barnes, 2014). Universities in the United States have measures for dealing with cheating and plagiarism however, plagiarism can also be viewed as an intentional or unintentional act. A simple example of this is the difference between a student copying text from another source without citations or quotes and a student unintentionally appropriating ideas from another source. For this reason, specific cases of plagiarism are difficult for universities to assess (Camara, 2016). This technical project entails creating a machine learning model that will be able to discern between plagiarized works and legitimate ones. This tool will allow professors and universities to determine which students have committed plagiarism.

The STS research paper will explore the effects of big technology companies on gentrification and how these large companies and cities take steps to minimize the effects. For the context of this paper, big technology companies are defined as companies that focus on technology development and are prosperous in today’s market such as Amazon, Facebook, Microsoft, Google, and Apple. In addition, “Gentrification is a process through which lower-income neighborhoods experience capital investments and an influx of wealthier residents” (Mujahid, 2019). In particular a large majority of big technology companies choose to create their headquarters in similar locations which results in the increasing of real estate value and rent which drives out the local population that originally lived there. This research paper will go more in depth about the development of gentrification in major cities and the measures that can be taken to minimize its effects on the local demographics.

Technical Topic: Plagiarism Detection using Machine Learning

The potency of plagiarism has only increased since the inception of the World Wide Web which provides students with a plethora of written works that can easily be accessible. With so much reference material online, it has become increasingly difficult for university professors to identify cases of plagiarism in their students' work (Batane, 2010). In order to combat the growing issue of plagiarism, many universities have adopted software that detects instances of plagiarism in students' papers. One example of this was the Turnitin software which was used at the University of Botswana. After the introduction of this software, the rate of plagiarism dropped by 4.3% within the first use (Batane, 2010). It appears that universities are making a conscious effort to develop tools to mitigate plagiarism but it is important to note that there are cases where works of plagiarism go undetected even by the most advanced technologies. This can be caused by a variety of reasons such as when students change the wording of plagiarized texts in order to bypass checks in the algorithm.

The proposed solution is a plagiarism detection tool that utilizes machine learning. Machine learning is a computational process that can "automatically alter or adapt their architecture through repetition (i.e., experience) so that they become better and better at achieving the desired task" (Naqa, 2015). Generally, machine learning algorithms create models that are trained by consuming large amounts of data which allows them to determine certain characteristics about new inputs that are passed in. In this case, the algorithm would first be presented with a large amount of written works from students, some of which are legitimate and some of which will be plagiarized, the algorithm will then find similarities in the data and train a model that can accurately determine the validity of new inputs in the future.

When considering plagiarism detection, there are 2 different types of plagiarism from a detection standpoint. The first type is known as extrinsic plagiarism where an algorithm simply compares a written work to another document and checks for similarities to draw conclusions about the legitimacy of the work. This method is very effective if the plagiarized work's original source can be found in digital form on the internet. However, if the plagiarizer copied text from a physical book or an unpublished source such as another student, the plagiarized work would not be considered extrinsic plagiarism. The other type of plagiarism is known as intrinsic plagiarism which examines the work in question in a vacuum without considering other sources. The method involves identifying changes in writing style to detect illegitimate works (Lalmas, 2006). Of course, if a student were to copy the entirety of an online text or a large amount of it, intrinsic analysis of the document would not work as there would be no changes in writing style. For these reasons, it is important to account for both styles of plagiarism detection to best counteract this growing problem. The proposed solution will utilize both methods in plagiarization detection in order to provide an accurate and reliable tool against plagiarism.

STS Research Topic: Big Tech Companies and Gentrification

Throughout the last 2 decades, the arrival of large technology companies such as Facebook, Apple, and Google in areas such as San Francisco has resulted in "tech booms" which have overtime displaced previous residents in the area due to increasing housing costs. More specifically, this is caused by an "influx of tech capital and wealthy tech workers" which has "precipitated spectacular increases in rental and housing prices throughout the region" (Hou, 2017) resulting in the previous inhabitants being unable to afford the new prices and having to relocate.

The research question that this paper will be trying to discuss is how can the effects of gentrification be minimized through the cooperation of the city and the companies at play?

A potential Science, Technology, and Society theory that can be used to analyze this research question is Actor Network Theory. Actor Network Theory was first developed by John Law, Michel Callon, and Bruno Latour and explains how "humans and nonhumans (including tools, technologies, texts, and the material world) come together (Crawford, 2020)" in a system called a network. This theory emphasizes the importance of relationships between the actors in the network and demonstrates the importance of each individual actor on how the system functions. The theory explains that a system cannot be fully diagnosed without considering all the actors in play as they all play a role in the system that is being analyzed. This can be applied to this research question concerning gentrification as there are a lot of different actors at play including, big companies, the city, tech workers, original inhabitants, and the landlords. Each of these actors are crucial to understanding the problem and examining each actor's viewpoints and their relationships with each other are integral to developing potential solutions. I believe that this is a suitable method source for this research question as I believe that the causes of gentrification do not come from one source and in reality there are a lot of different actors that impact and create the network.

Conclusion

This portfolio will have 2 different projects. The technical project will explore plagiarism detection using machine learning. The project will have a focus on analyzing both intrinsic and extrinsic plagiarism to validate student works. The STS works Actor Network Theory and Do

Artifacts Have Politics will serve as methods in analyzing the STS research topic of Big Tech Companies and Gentrification and potentially provide solutions to the ongoing problem such as increasing accessibility for remote work or for companies to disperse their headquarter locations rather than centralizing them in a specific location. The Plagiarism Detector and STS research paper are on track to meet their deadlines.

References

Mujahid, M. S., Sohn, E. K., Izenberg, J., Gao, X., Tulier, M. E., Lee, M. M., & Yen, I. H. (2019). Gentrification and Displacement in the San Francisco Bay Area: A Comparison of Measurement Approaches. *International Journal of Environmental Research and Public Health*, *16*(12), 2246. https://doi.org/10.3390/ijerph16122246

Camara, S. K., Eng-Ziskin, S., Wimberley, L., Dabbour, K. S., & Lee, C. M. (2016). Predicting Students' Intention to Plagiarize: an Ethical Theoretical Framework. *Journal of Academic Ethics*, *15*(1), 43–58. https://doi.org/10.1007/s10805-016-9269-3

Barnes, B. D. (2014). Plagiarism, morality, and metaphor. (Doctoral dissertation). Retrieved from Proquest Dissertation and Theses (Accession order number: 3668689).

Batane, T. B. (2010). International Forum of Educational Technology & Society. *Turning to Turnitin to Fight Plagiarism among University Students*. Published.

Naqa, E. I., Li, R., & Murphy, M. J. (2015). *Machine Learning in Radiation Oncology: Theory and Applications* (2015th ed.). Springer.

Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., & Yavlinsky, A. (2006). *Advances in Information Retrieval*. Springer Publishing.

Hou, J., & Knierbein, S. (2017). *City Unsilenced: Urban Resistance and Public Space in the Age of Shrinking Democracy* (1st ed.). Routledge.

Opillard, F. (2015). Resisting the Politics of Displacement in the San Francisco Bay Area: Anti-gentrification Activism in the Tech Boom 2.0. *European Journal of American Studies*, *10*(3). https://doi.org/10.4000/ejas.11322

Crawford, T. H. (2020). Actor-Network Theory. *Oxford Research Encyclopedia of Literature*. Published. https://doi.org/10.1093/acrefore/9780190201098.013.965

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136. http://www.jstor.org/stable/20024652