Section Draft

Max Feng

Introduction

Linguistic systems are not always static. Although they are often well-balanced, language still evolves sometimes due to newly developed objects or ideas (Blank, 1999). For instance, cloud develops a new meaning “a vast online storage space” in the recent years due to the need for a name of that object. This process, known as semantic change, is one of the central topics of historical linguisitics. In the digital era, certain words such as bug, web, mouse, and cloud come to develop new meanings to accommodate the invention of new technological devices. They are now more often recognized as technology terms. Interested in the semantic change of technology terms, I will explore this idea as my data science capstone project.

Semantic change has been investigated historically through qualitative analyses of texts over time (Blank, 1999). However, this traditional method is time-consuming and can be subjective. In the modern era, the vast storage capability of computers allows us to possess an unprecedented amount of data presented in media, which makes the quantification of semantic change possible. Therefore, in the present project, I choose to look at the technology terms presented in news media. News articles are not only readily accessible but also time-sensitive so that I can track the semantic change over time.

Fortunately, advancements in natural language processing (machine learning to help computers understand human language) have equipped us with an approach to quantify semantic change over time. This approach, known as diachronic word embeddings, is a technique that represents words as high-dimensional vectors and creates a quantitative coordinate map of meaning over time (Hamilton et al., 2016). By training separate word embedding models on technology terms presented in news media from each year between 2014 and 2024, I create a temporal series of semantic snapshots. The core of the analysis involves aligning these snapshots from each year to one single map and then measuring the cosine distance between a word’s vector representation across different years. This distance serves as a numerical index of semantic change, which allows me to identify which technology words have shifted most significantly over the past decade.

After observing the semantic change of certain technology terms, I will perform an analysis to each word that explains why it undergoes a significant amount of semantic change at that time point. I will incorporate real-world events in the analysis to make sense of the semantic change. Finally, I will create a Shiny dashboard that lists the technology terms as well as their magnitude of semantic change and the analysis of that change.