Tracking Semantic Change of Technology Words

Capstone in Data Science

Junshu Feng

December 1, 2025

Abstract

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, Leskovec, and Jurafsky 2016). The technique is especially realizable in R as it offers a package named *text2vec* (Mimno 2018). By training separate word embedding models on technology terms presented in news media from each year between 2016 and 2025, 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 (Chang et al. 2023).

Methodology

Building the Annual Corpora

Training Word Embedding Models

Results

Dashboard Tutorial

Ethical Concerns

References

Blank, Andreas. 1999. "Why Do New Meanings Occur? A Cognitive Typology of the Motivations for Lexical Semantic Change." In *Historical Semantics and Cognition*, edited by Andreas Blank and Peter Koch, 61–90. Berlin: Mouton de Gruyter.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2023. *Shiny: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. https://doi.org/10.18653/v1/P16-1141.

Mimno, David. 2018. "Text2vec: Modern and Efficient Toolkit for Text Analysis and Embedding in r." *Journal of Statistical Software* 85 (1): 1–30. https://doi.org/10.18637/jss.v085.i01.