A Comparative Study of Word Embeddings for Scene-Based Genre Classification of Movies and Genre Co-occurrence

Gaoussou Youssouf Kebe[1], Izzet Fatih Şentürk[1]

[1] Bursa Technical University

Abstract

Processing natural language such as English has always been one of the most sought after research goals of artificial intelligence and computer science in general. Text classification in particular is one of the most prominent subfields of natural language processing. There are numerous projects based on classification algorithms of books, blogs and other textual data generally by genre. However, there is very few examples of movie script genre classification. Even rarer is the usage of such algorithms to partially classify the different scenes composing a movie, despite the amount of knowledge we could obtain from such a task. Presented solution addresses the problem of classifying movie scenes solely based on the full script of the movie. Movie scripts, obtained from the Internet Movie Script Database (IMSDb) are broken down with scrutiny into individual scenes. Features from each of these individual scenes are then extracted using three state-of-the-art word embedding methods (Word2vec, Glove and FastText). An efficient similarity search library (NMSLIB) is used to classify the scenes using a K-Nearest Neighbors classifier. Obtaining such data would allow a better understanding of the timeline of the movie and can potentially help provide more insight into decision support systems such as recommender systems.

Keywords: *Natural Language Processing, Machine Learning, Word Embeddings, Classification*