Analyze News Headlines and Predict Article Success

Sowmith Nallu 50286932
Shwetasree Chowdhury 50296995

Abstract:

We observe that having sensational headlines for an articles intrigues our interest to read. But do the headlines really matter for the article to become popular? We aim to answer this question from this project.

In this project we use word2vec model to create word and title embedding and then visualize the relationship between title sentiment and article popularity, and then we attempt to predict the article popularity from the embeddings and other available features.

Apache Spark is advantageous for text analysis because it provides a platform for scalable, distributed computing and is faster when dealing with JSON objects and provides fast computing using intermediate caching. We collect the data from News api and NY Times API.

In a nutshell, the steps we aim to follow in order to carry out the same:

1. obtain data on a particular topic from NY Times, sort by 'rank', ‘number of email shares’ and ‘views’ and topic.
2. use MLLib (Spark’s machine learning library) - and Word2Vec which creates vector representation of words in a text corpus.
3. Use Sentiment analysis on the above representation and plot the same as a means of visual comparison.