ID2223- Project proposal

Clickbait titles generator

Aurélien Blicq

I- Project description

In the recent years, some news websites have been called out for producing so called “clickbait” content. This content is generally low in quality, easy to produce in large volume and is wrongfully advertised by a title and a thumbnail that are sensationalists and misleading. This is mainly done in order to attract traffic and generate ad revenue.

For instance the website BuzzFeed is renowned to use such practices and is responsible for articles like “21 Images That Capture How Scary It Is Being In Sydney Today” or “Only An Actual Teacher Can Correctly Answer 6/8 Of These Grammar Questions”.

This trend has been so much exploited by some websites that it has become a meme on the internet, and is oftentime frowned upon.

For this project, I propose to implement a clickbait titles generator using RNN and LSTM.

II- Tools

For this project, I will mainly be using the language python and the library TensorFlow to help me implement these neural networks.

III- Data

The Data used to trained the models comes in part from a dataset of 16 000 clickbait titles used by a research team to implement a SVM classifier that would discriminate clickbait titles and regular titles [1], and in part from the titles that I have been able to extract from BuzzFeed through their RSS feed.

IV- Methodology

First, we need to preprocess the data. For that, we need to embed the words of the titles into sequences of vectors using tokenization and word2vec.

Then, I want to compare the performances of a RNN and a LSTM to generate such short sequences of words. I will thus train each of the models to generate sequences of words.

To assess the performance of the model, we have to take into account the speed of convergence of the training, and the quality of the generated titles (syntactic and semantic correctness, similarity to the training data).

V- References

[1] https://github.com/bhargaviparanjape/clickbait