In this project, we are going to analyze different factors that will make your article more popular on the internet. Of course the decisive factor is always that if your article is interesting or not. But thats not something we can quantify. So instead, in this project we select a large range of features and attributes of an article, such as the length of the article, the catogories it belongs to, how many pictures it has, the publication time and so on. And then we are trying to figure out if these features have something to do with the popularity of the article, which reflected in the number of shares in social networks. like facebook and twitter

For those who ever write something online, which is about millions of people, you better pay attention because, for example, if we found that the number of shares tend to drop when the word count is greater than 2000, then you better make your writings shorter. And there are a lot of interesting features just like that and that's what we are about to discover.

The project will be done in three major steps: First, we are gonna fetch data from the target websites and process it. Second, we select meaningful combination of metrics and calculate statistics, like the correlation between word count and the number of shares. According to the survey, these factors do have an impact on people's choice of reading, and we are gonna prove it using data. Moreover, we will learn from the existing datasets and use machine learning to predict the popularity of new articles. This is kinda like an innovation cuz we don't see a lot of people have done that before. The major risk and challenge in doing that is of course the noise. For example, some famous writers may write 49 pages and people still wanna share their writings. Fortunately, according to survey we learned a bunch of machine learning algorithms that can help fixing issues like overfitting. That is how we gonna do. Finally, we are gonna use a set of tools to visualize data in order to provide the most straightforward results to someone who may have no knowledge in data analysis.

The project will take about 2 month, the mid term will be the first two steps and part of the machine learning and visualization is gonna be the second half. and the costs are a lot of pizza.