LogicPlum

CASE STUDY

CATEGORIZING LONG-LASTING CONTENT ON THE WEB

StumbleUpon

Company Info

Name: StumbleUpon

Location: Calgary, Canada

Industry: Advertisement Engine

Uncovering hidden gems. Building a classifier to categorize webpages for StumbleUpon.

Do you enjoy certain webpages more than others? Would you prefer to reach content that stays relevant over a long time on an Internet search list? Well, this is exactly what StumbleUpon does. And, in order to provide a great service, the company uses social networking and forefront AI technologies.

Since 2001, when the company started to provide its services, StumbleUpon has worked hard to find those websites that will pass the test of time.

Their classification method divides pages as evergreen or ephemeral. Evergreen pages are those pages whose content has a timeless quality and thus, can be recommended long after having been discovered. Conversely, ephemeral pages are those websites that are only relevant for a short time.

StumbleUpon classifies the webpages through a process known as collaborative filtering that creates virtual communities of users with similar preferences.

Collaborative filtering is an automated process that combines human opinions and machine learning insights into personal preferences.

With this motivation, StumbleUpon collected data from 10,566 URLs and decided to build a model that could evaluate a broad set of URLs and label them as either evergreen or short-term.

“We knew that we were equipped with the latest AI algorithms.“

LogicPlum Uncovers Itself

LogicPlum provides its users with a platform that unveils the best algorithmic solution to a data science problem. It does this in an automated manner by using a combination of AI and machine learning technologies.

According to LogicPlum, "***What this platform brings to machine learning is the possibility of exploring hundreds of different possible solutions in a short time and in an efficient way“.***

So, when LogicPlum's data scientists were faced with the StumbleUpon's problem, they formed a team to tackle it

Uncovering the Data

The data showed the raw content for each URL, as detected by StumbleUpon's crawler. It included a unique identifier, a boilerplate field, alchemy category and score fields, links shared, ratio of tags versus text, number of anchor tags, and several other ratios describing the relationships of text to other elements as images and hyperlinks.

"We wanted to understand the importance of the different features. Thus, we created some visual representations of their values and relationships," explained the team leader. And, he added: ***"The results found were fascinating. First, there was a lot of noise in the data. Second, we found that some features, such as the alchemy category, were much more relevant than others"***.

Uncovering the solution

Once the team understood the data available, they decided to let the platform find the right solution.

The platform tried several hundreds of solutions and classified them by using a ROC curve.

The winning solution showed to be fairly straightforward. It entailed several "classical" models applied separately to the TF-IDF of texts, URLs, and metatags.

Finally, the results were combined with the help of a linear model. The final AUC was 90%, which indicated a very good classification prediction based on given factors for a specific website.

The team was thrilled that they had found an obvious and straightforward solution in a short time. ***"The platform had found the right feature combination and then applied a simple linear classifier to it. It had clearly shown us the importance of good feature engineering!"*** explained the team leader.

Uncovering the Final Solution

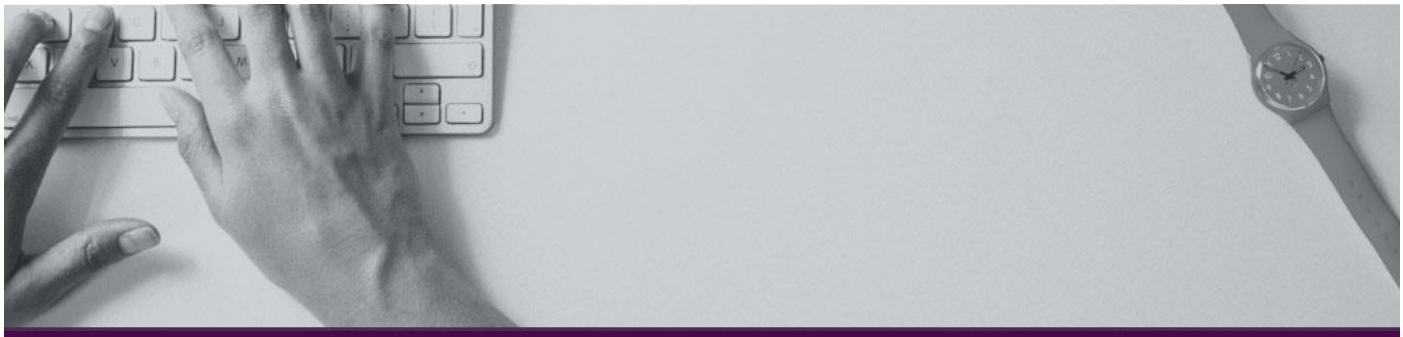
Now that they had the right solution, the team wanted to complete the project by providing a final product that would ensure business adoption. For this, the team had to develop a production version of the algorithm and the necessary documentation.

They first developed a user-friendly interface that connected to the platform via its single point API. This was very important in their eyes, as ***"it would allow for the use and maintenance of the solution by a user who was not proficient in AI,"*** explained the team leader.

Next, the team wanted to create a comprehensive report that described their findings, steps taken, and the final solution. Hence, they used another component of the platform named R.E.A.S.O.N.™, an AI tool that helps users prepare the critical parts of a project report.

"We used this tool because we wanted to document the whole process, from beginning to end, in a highly readable manner. We wanted to provide a chronicle that could be useful to our clients and ourselves alike," summarized the team leader.

The team had once more succeeded and had proved that there was a way for a better website-classification that could find the best of the web according to each person's preferences.



Contact Us

LogicPlum
1550 West
McEwen Drive
Suite 300
Franklin, TN 37067USA

www.logicplum.com
message@logicplum.com

LogicPlum