Ben Plotke & Renato Stoco

What is the problem or task you propose to solve?

The task proposed to our project is to develop a system that will crawl over the internet categorizing webpages as if they are interesting for the user in a scale that varies from 1 to 4.

What is interesting about this problem from an NLP perspective?

The heart of the problem lays on how are we going to select the proper features and how we can use supervised learning in a way that aids us to mitigate any misclassified webpage.

What technical method or approach will you use?

We will be using Naïve Bayes and Maximum entropy models to calculate the probabilities of a class given the webpage. In addition, we will create two categorizers functions: one dedicated to build our training data based on our selected topic labels, and the second system will be the one dedicated to categorize the webpages based on their interestingness. The key on this approach is to try to mitigate any misclassified result from the first categorizer pass.

On what data will you run your system?

For our topic labels categorizer we will be using training data from the 20Newsgroup provided by Jason Rennies page:
http://people.csail.mit.edu/u/j/jrennie/public_html/20Newsgroups/ [1]

And, for the interestingness categorizer we will be generating our training data based on the webpages that we classify.

How will you evaluate the performance of your system?

For the topic labels categorizer we have the labeled test data so we can run the 10-fold validation to evaluate our system. For the interestingness categorizer, since interestingness is subjective we will evaluate those by judging if the webpages correspond to our expectations.

What NLP-related difficulties and challenges do you anticipate?

We expect that tuning two system categorizer tasks will be challenging because we have to think what key features (aside from bag of words) can better play a role in order to classify web pages. Also, we will be using some NLP analysis (Language Models) to better prune the webpages that are not a good source of information.

[1*] Ana Cardoso Compiled the data from Jason Rennies into various datasets, these datasets includes the ones that we are using for training the data and the testing data. Webpage: http://ana.cachopo.org/datasets-for-single-label-text-categorization