NaiveBayesParseFilter

Naive Bayes Parse Filter for Nutch

A new plugin for Nutch-1x that provides the capability for focused crawling.

Summary

The plugin uses a Naive Bayes classifier https://en.wikipedia.org/wiki/Naive_Bayes_classifier (needless to say it is the most famous classifier and works magic with good training data). In simple terms a classifier takes some (the more the merrier) positive and negative examples of text of features as a training set. The classifier learns what is positive and what is negative and creates a model. You can say this model is the brain that contains everything that the model learned from the training. Now the model will be used to classify new text as positive or negative.

Implementation

The plugin is implemented in Nutch as a Parse Filter that will filter out the outlinks in two stages. Classify the parsed text and decide if the page is relevant. If relevant then don't filter the outlinks. If irrelevant then go thru each outlink and see if the url contains any of the important words from a list (needs to be provided by the user). If it does then let it pass. You can follow the JIRA issue to gain some more insights into its design

Use

Well, you need Nutch 1x for sure. Set some properties in Nutch in the nutch-site.xml. Create training file (see the description in the related nutch property below: parsefilter.naivebayes.trainfile) Create the wordlist file (see the description in the related nutch property below: parsefilter.naivebayes.wordlist)

```
<property>
<name>plugin.includes</name>
<value>parsefilter-naivebayes</value>
</property>
<property>
  <name>parsefilter.naivebayes.trainfile</name>
  <value></value>
  <description>Set the name of the file to be used for Naive Bayes training. The format will be:
Each line contains two tab seperated parts
There are two columns/parts:
1. "1" or "0", "1" for relevant and "0" for irrelevant document.
3. Text (text that will be used for training)

Each row will be considered a new "document" for the classifier.
CAUTION: Set the parser.timeout to -1 or a bigger value than 30, when using this classifier.

  </description>
</property>
```

```
<property>
  <name>parsefilter.naivebayes.wordlist</name>
  <value></value>
  <description>Put the name of the file you want to be used as a list of
  important words to be matched in the url for the model filter. The format should be one word per line.
  </description>
</property>
```

Some Trivia

Created to be workable in a distributed environment (All because it uses Apache Mahout library).

Questions

For any questions, concerns and new ideas you might want to implement inside this plugin or something similar outside it, contact Asitang Mishra.