LanguageIdentifier

Architecture

TODO

NGram profile format

TODO

Generating some NGrams profiles

Generating a new language profile in Nutch is really easy. Simply launch the following command:

```
java org.apache.nutch.analysis.lang.NGramProfile -create <profile-name> <filename> <encoding>
```

where

- **profile-name** is the ISO-639 2-letter codes of the new language.
- **filename** is the name of the file used to build the new language profile (the biggest it is, and the most it contains different sources and subjects the better the profile will be).
- **encoding** is the encoding of the file used to build the new profile (**filename**).

Open Issues

- *Labs* tests are quite good (LanguageIdentifierBenchs), but in *real life*, they are not. In fact, in its actual version, the NewLanguageIdentifier expects that the provided text to analyze is UTF-8 encoded. However, it is not the case for a lot of fetched documents. So, the NewLanguageIdentifier needs to refer to a `content-encoding` meta-data. This data must be provided by a (todo) EncodingDetectorPlugin (see NUTCH-25 issue).