Information Retrieval Project 1 Group 11

Michael Merki, Julien Romero, Markus Greiner

November 14, 2016

Installation

Provided is the zip file `project1_group11.zip`. Unzip it:

```
$ unzip project1_group11.zip
```

This creates the directory `project1_group11`, which contains:

- the sources under `src`
- `tinyir.jar` as part of `lib`
- `build.sbt`
- the directory `labelingtestdocs` which contains the three resulting test results.
- and this `README.md` file.

Note: We had continuing conflicts between the tinyir and the breeze libraries. This is why we decided to build tinyir for Scala 2.11.5 and provide it as jar.

Running

To run:

```
$ cd project1
$ sbt "run-main Main <path-to-data-folder>"
```

The `<path-to-data-folder>` must contain the directories `train`, `test`, and `validation`.

It is also possible to give options to influence the iterations and learning rates of the linear regression classifier.
These options are called:

- `ITERATION=<nof-iterations-integer>`
- `LEARNING=<learning-rate-double>`
- `SKIP=(BAYES|LINREG|SVM)+`

It can be run like this:

```
$ sbt "run-main <path-to-data-folder> ITERATION=10000 LEARNING=0.001 SKIP=BAYES,SVM"
```

Results

Upon running the program, it will

- *Naive Bayes Classifier*: train and generate list of tested documents and their codes
- *Logistic Regression Classifier*: train and generate the list
- *SVM*: train and generate the list.

The result files are called `ir-project-2016-1-11-[nb|lr|lsvm].txt` and they are located under `labelingtestdocs`

The project report is under `ir-2016-1-report-11.pdf`