Information Retrieval Assignment 1
Group 5 (Dávid Frank, Ferenc Galkó, Zalán Borsos)

Distinct urls:	5780
Exact duplicates:	1407
Near duplicates:	547
English pages:	2530
Student frequency:	2506

Methodology

We distinguish between the entire textual content of the page (named *full text*) and the content within the `#content` element (named *content*). If the amount of text of the *content* retrieved this way is not sufficient (the length of the extracted text is < 10 characters), we use the *full text* as the *content*.

For the calculation of exact duplicates, English pages and student frequency, we use the *full text*, for near duplicate detection we use the *content*. We consider two documents as near duplicates if their *similarity hashes* differ in 0 or 1 bit.

In addition to the above statistics we have made the following observations:

1. There are 1300 login pages, which are all exact duplicates.
2. The very high frequency of *student* can be explained by the header which contains the expression *Student portal* (except for a small number of pages).

The application was compiled against scala version 2.11.7. If you have incompatibility issues, try to run the application using:

```
java -jar ir-2015-crawler-5.jar
```