

## Web page language identification based on URLs

**Eda Baykan**

**EPFL, Switzerland**

<http://people.epfl.ch/eda.baykan>

Abstract:

Given only the URL of a web page, can we identify its language? A language classifier based on URLs is, for example, useful for crawlers of web search engines, which frequently try to satisfy certain language quotas. To determine the language of uncrawled web pages, they have to download the page, which might be wasteful, if the page is not in the desired language. With URL-based language classifiers these redundant downloads can be avoided. Our best methods achieve an F-measure, averaged over all languages, of around 90% for both a random sample of 1,260 web page from a 100 million web crawl and for 25k pages from the ODP directory.