# WS 2015 Project 1:
# Web Size
# Due by: 20 February 2015, 23h55

**This project counts towards 30% of your final grade for this course.**

## 1 Project description

You will be given:

1. An **article** by Lawrence and Giles (1998) entitled: *Searching the World Wide Web* [1]. You must carefully read this article in order to understand and carry out this project. This article is attached at the end of this document.

2. A **template** file that you should use for writing your report (pdf + latex sources). The pdf of the template is attached at the end of this document. The latex sources of the template are on Absalon (/Projects/Project1/).

3. A list of **500 queries** that you must use to carry out this project.

**New project definition:** This project asks you to carry out a size estimation of the indexable Web using the approach presented by Lawrence and Giles in [1]. Lawrence and Giles [1] estimated the size of the indexable Web in 1998 using the results of 6 different search engine in response to 575 queries. You must replicate their approach in order to estimate the size of the indexable web today. For this, you must use search results from 3 search engines: Bing, Baidu and Yandex, in response to a list of 500 queries (available on Absalon). You must follow the methodology of Lawrence and Giles [1], i.e. apply the same constraints and overlap analysis. Use these estimations of web coverage: 740 million webpages by Baidu, 14 billion webpages by Bing, and 15 billion webpages by Yandex. You will not be able to retrieve as much data as described in [1] you should therefore just try to retrieve as much as possible. In [1] they state that they download each page and checks that the query term occurs in the document, you may omit this step and instead comment of the effect this change has on your results. Another modification to the approach presented by [1] is that you should not compare the linked documents but just the URLs, this is due to the fact that most pages nowadays contains dynamic content and therefore

changes constantly. Since not all webpages from each query are collected you are only calculating the overlaps on a small subset, and you should therefore try to extrapolate your results by using the total amount of results each query produces. This new project definition contains a lot of changes compared to the approach described in [1], you should therefore comment on the effects these changes will have on your estimates compared to the results obtained by [1]. To evaluate the stability of your estimate you should repeat the calculations using the following three approaches, this will tell you something about the robustness of your results:

- Repeat the experiments five times each time using a random subset of 300 queries out of the 500.
- Use the 300 queries which resulted in the most search results.
- Use the 300 queries which resulted in the fewest search results.

**Old project definition:** *This project asks you to carry out a size estimation of the indexable Web using the approach presented by Lawrence and Giles in [1]. Lawrence and Giles [1] estimated the size of the indexable Web in 1998 using the results of 6 different search engine in response to 575 queries. You must replicate their approach in order to estimate the size of the indexable web today. For this, you must use search results from these 3 search engines: Google, Bing, and Baidu, in response to a list of 500 queries (available on Absalon). You must follow the methodology of Lawrence and Giles [1], i.e. apply the same constraints and overlap analysis. Use these estimations of web coverage: 48 billion webpages by Google, 14 billion webpages by Bing, and 740 million webpages by Baidu. You may not be able to retrieve all results from every query. You should therefore retrieve as much data as possible from each search engine and perform the experiments using the following three approaches:*

Report your findings by presenting equivalent tables to Tables 1-2, equivalent figures to Figures 1-3 in [1], and a corresponding discussion. Your report should include the sections specified in the template file (introduction, methodology, findings, and conclusions). You can use any tools you want, including your own programs, commercial or public domain tools like Unix commands.

# 2 Submission

## 2.1 What to submit

You must submit **a single tar.gz file** that contains:

1. your report in pdf (not the latex sources), formatted according to the template, and
2. the source code that you used to perform the relevance feedback (i.e. your programs, commands etc.).

**Everything in your submission must be anonymous** (i.e., do not write your name in the report or your code). There are no length restrictions for the report, however it must contain **all the sections in the template**, plus any more sections you wish to add.

### 2.2 How to submit

- You must upload your submission to Absalon **by 20 February 2015, 23h55, at the latest**.
- If you are unable to submit via Absalon for some reason, you must send your submission by e-mail to ingemar.cox@di.ku.dk **with cc to** brian.brost@di.ku.dk and nhansen@di.ku.dk **by 20 February 2015, 23h55, at the latest**.
- Submissions received after the deadline without prior approval for e.g. medical reasons or similar, by I. Cox, will not take part in the peer-assessment. This results in an immediate -20% reduction of your final portfolio grade (15% for the peer-assessment you will not make + 5% for your missing amendment list - see the *Portfolio Guidelines* for details).

## References

[1] S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.