# Exercise 1

**Information Retrieval and Web Search**

Ahmed Rekik and Tim Henning

October 26, 2017

## 2. Information Retrieval Introduction

### 2.a. Explain the terms Web search and information retrieval. How do they differ from each other?

As an abstract definition, the term "Information retrieval" refers to the activity of searching and finding relevant information from an unstructured (collection of) document(s) in terms of a needed query. The aimed information are usually unstructured and may be quotes from a full text, papers, a book from a library, numeric data, or even images.

In the other hand, Web search term is a type and a practical application of information retrieval based on indexed data from the world wide web. Using engines specialized in web search one can search for information using queries that express this need. Through search and ranking algorithms the web search engine returns a list of matched information in terms of the existing words in the query.

*The difference between the two terms is that the web search performs only on structured data updated in the Internet, while the informational retrieval performs as well on unstructured data such as books. The relevance may as well differ by both of terms: While the information retrieval take in consideration the exact aims of search, performs the web search using with an automatic system only on the provided keywords in the query and results to a well-defined result set.*

### 2.b. Explain the relevance notion as defined in information retrieval. What makes a document relevant or not relevant to a particular query and user?

The term relevance denotes the suitability of the retrieved information in terms of the query need and the user preferences. Generally expressed in a binary scale (relevant or not relevant) but it could also be represented in a graded scale ( Very relevant, relevant, less relevant, not at all relevant).

A document is relevant to a specific user if it satisfies its information need. This means that retrieving relevant information may not be an easy task to achieve, especially when the user has to describe its often imprecise information need as a list of keywords.

As the information retrieval systems work with indexed data and by matching patterns, some confusions may occur. Example: Querying "Jaguar" to look for cars and getting information about the animal jaguar instead. Thereby, the user has to be very precise in his query by expressing more specification in order to retrieve relevant information.

# 3. Web Crawling

## 3.a. What is the advantage of using HEAD requests instead of GET requests during crawling? When would a crawler use a GET request instead of a HEAD request?

During crawling, the crawler uses the HTTP request `HEAD` in order to easily fetch specific information about the crawled page such as the last modification time (or the age of the page). This attribute may participate in the analysis of targeted page, which occupy a fix URL address, by comparing it with previous update times to adjust the frequency of crawling in terms of the *Age* metric in order to avoid superfluous crawling in the future.

While the `HEAD` request is much faster because it returns less data, a `GET` request is required to fetch the whole page including its content, for example if it has changed since the last visit.

## 3.b. What are the obstacles that a crawler faces when attempting to fetch web pages? Give examples for challenges concerning the semantic information retrieved and the efficiency of crawling.

Typical obstacles for crawlers are for example the downloading speed, politeness rules, information noise like advertisements or navigation elements and the large number of website duplicates.

To access the information on a webpage the crawler has to perform a DNS lookup to get the IP address and must download the corresponding HTML and sometimes JavaScript files. A solution could be the parallel execution of the requests, but to follow the so called politeness rules the crawler has to wait a few seconds before performing another request on the same server. This makes a queue of requests necessary.

Another difficult task is to separate the relevant content in an HTML file from noise like ads and navigation elements. Beside that the crawler also has to keep track of the wepages it already knows (i.e. by a fingerprint algorithm) to not index duplicates of it again.

In the end, the robots.txt file, that was intended to help the crawler, could also be an obstacle if it is not well-formated or restricts the access to relevant parts of the website.

# 4. (Programming) Web Crawling

## 4.a. Print the csv file for one newspaper article of the 16th October, 2017. Choose an article with at least 5 comments.

See the attached file.