

# Assignment 1

## Information Retrieval and Web Search

Ahmed Rekik - 790063 & Tim Henning - 789242

October 31, 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 data accessible from the Internet, while general information retrieval systems may also include non-digital data such as books in their corpus. While web search almost always uses a list of keywords as its query language, other information retrieval systems could also use a more complex query language to get a more specific 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.

Storage of converted document text may be also seen as a time consuming operation and slow the crawling task down.

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 webpages 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.

CSV file generated for the following article: [Trump: Total termination of Iran nuclear deal is a 'very real possibility'](#) (RT US News, October 16th, 2017)

```
id , url , author , text , timestamp , parent_id , up_votes , down_votes
sp_6phY2k0C_406881_c_3AfH17 , http://www.rt.com/usa/406881-trump-
says-total-termination-of-iran-deal-possible / , GoldStrawberry
, The world must stand up to US capitalist terror and
barbarism . , 1508170712 , , 132 , 2
sp_6phY2k0C_406881_c_3AfH17_r-IkuQVq , http://www.rt.com/usa
/406881-trump-says-total-termination-of-iran-deal-possible / ,
PurpleTrumpet , But who has the bravery to stand up to the
bully ? , 1508171553 , sp_6phY2k0C_406881_c_3AfH17 , 30 , 2
sp_6phY2k0C_406881_c_3AfH17_r-Bc2llt , http://www.rt.com/usa
/406881-trump-says-total-termination-of-iran-deal-possible / ,
whatislove , Erdogan . , 1508172109 , sp_6phY2k0C_406881_c_3AfH17
, 22 , 6
sp_6phY2k0C_406881_c_3AfH17_r-CKFDcy , http://www.rt.com/usa
/406881-trump-says-total-termination-of-iran-deal-possible / ,
MrUniverse81 , Our little bubbly bro Mr.K , 1508178527 ,
sp_6phY2k0C_406881_c_3AfH17 , 21 , 0
sp_6phY2k0C_406881_c_UHkOKx , http://www.rt.com/usa/406881-trump-
says-total-termination-of-iran-deal-possible / ,
BlackwateropGla , " Nothing will shock me nor surprise me from
the part of USSA!!! They deserve the trophy for " " The most
hypocrite and the biggest criminal entity-nation in the
history of man kind!! " " " , 1508171292 , , 71 , 1
sp_6phY2k0C_406881_c_UHkOKx_r-VRIAyq , http://www.rt.com/usa
/406881-trump-says-total-termination-of-iran-deal-possible / ,
OliveUmbrella , Our puppets had better cut the strings
. , 1508171956 , sp_6phY2k0C_406881_c_UHkOKx , 14 , 1
sp_6phY2k0C_406881_c_UHkOKx_r-pRsptl , http://www.rt.com/usa
/406881-trump-says-total-termination-of-iran-deal-possible / ,
BlackwateropGla , " US and the elite are extremely desperate ,
the time is running out , their financial system is falling
apart at a rapid space and the end of the petrol-dollar as
world currency is near his end . They need a major war or a
hand made natural catastrophe to blame the financial crisis
on!! If they pick a war with Iran or North Korea I need to
affirm with a lot of confidence that USA will not be the
same as we new it . " , 1508172479 , sp_6phY2k0C_406881_c_UHkOKx
```

,32,2  
 sp\_6phY2k0C\_406881\_c\_UHkOKx\_r\_zWBvWh,http://www.rt.com/usa  
 /406881-trump-says-total-termination-of-iran-deal-possible/,  
 SirLeo,"Yes, they are desperate. I also think they don't  
 know whom to attack. Worse, they are incapable of grasping  
 what the rest of the world is thinking.",1508172571,  
 sp\_6phY2k0C\_406881\_c\_UHkOKx,16,0  
 sp\_6phY2k0C\_406881\_c\_ep58BF,http://www.rt.com/usa/406881-trump-  
 says-total-termination-of-iran-deal-possible/,PurpleSeahorse  
 ,this is nuclear fuel being given to us...political fuel to  
 power political reactors for a thousand years...the more the  
 us/israel talks the deeper they sink,1508170375,,66,1  
 sp\_6phY2k0C\_406881\_c\_ep58BF\_r\_9DHMyO,http://www.rt.com/usa  
 /406881-trump-says-total-termination-of-iran-deal-possible/,  
 OliveSushi,fusion reactors that will be a strategically  
 unbeatable defense against these spawn of darkness  
 ,1508170719,sp\_6phY2k0C\_406881\_c\_ep58BF,20,0  
 sp\_6phY2k0C\_406881\_c\_ep58BF\_r\_ydexxA,http://www.rt.com/usa  
 /406881-trump-says-total-termination-of-iran-deal-possible/,  
 BlueBike,three quarters of a TRILLION per year who cares  
 about exports use correct statistics,1508179895,  
 sp\_6phY2k0C\_406881\_c\_ep58BF,2,0  
 sp\_6phY2k0C\_406881\_c\_ep58BF\_r\_z8IT9g,http://www.rt.com/usa  
 /406881-trump-says-total-termination-of-iran-deal-possible/,  
 GoldTree,Thats very true their military business is the only  
 thing that keeps their economy going and this blowhard is  
 doing his best as a used car salesman to drum up new  
 business. The reality is that if anything other than these  
 proxy wars happen as in a big regional war Israel will  
 vanish during the process and they know it.,1508170837,  
 sp\_6phY2k0C\_406881\_c\_ep58BF,63,0