**Information Retrieval**

**29 June 2015**

**Ex 1 [ranks 4+3+3]** Given the string S = alalabarda,

* Construct the wavelet tree, by assuming that the symbols in the leaves are ordered alphabetically
* Show how it is executed the operation Rank(a,6) over that WaveletTree
* Show how it can be implemented the operation Select(a,4) over that WaveletTree [which returns the position of the 4th letter ‘a’ in S]

**Ex 2 [points 4]** Given a suffix array SA for a string S, detail the algorithm that establishes whether does it exist in S a substring s whose length is larger than L and repeats more than c times in S.

**Ex 3 [points 3+3]** Consider the Consistent hashing technique:

* Discuss its advantages in crawling with respect to the simple scheme that allocates URL to crawlers based on a static url\_domain->crawler scheme
* Simulate its execution over the urls\_ID = {1, 4, 8, 2, 6, 7, 13, 11} and the crawler\_ID = {1,2,3}, by defining proper hash functions.

**Ex 4 [points 3+4+3]** Given a topic annotator, like TagME:

* Detail its goal
* Define the features: link probability of an anchor, commonness of a wiki-page wrt to an anchor
* Define the ratio behind the Milne-Witten similarity measure and where it is used in TagME