**Information Retrieval**

**12 June 2017**

**Ex 1 [points 4+3]** Given the binary array B = 0110 1110 1111 1100

* Construct the Rank data structure over B with Z=4 and z=2
* Describe how it is executed Rank1(7)

**Ex 2 [points 2+3+3]** Show the compressed encoding of:

* 7 and 18 with DELTA-code
* the sequence (1, 3, 2, 4, 2, 3, 1, 1, 2, 9) with PForDelta: base=0 and b=2 bits
* Given the previous sequence, how you would turn it into a sequence that could be encoded via Elias-Fano, and then show the result of the encoding.

**Ex 3 [points 3]** Comment on the difference between the Term-based versus Document-based partitioning in Distributed indexing.

**Ex 4 [points 4]** Given the directed graph G consisting of nodes {A, B, C, D} and edges {(A,B), (B,C), (D,C), (C,A)}: Describe how you compute the similarity between node B and node A via Personalized PageRank, and then show the result by iterating it once.

**Ex 5 [points 2+3+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 for which purpose it is used in TagME

**Ex 5 [LAB TEST]** Let us assume that we have built a Lucene index, with a Whitespace Analyzer, over the following three documents:

* d1 = " Con il nome Valle dei Re si indica un'area geografica dell'Egitto"
* d2 = "Con Antico Egitto si intende la civiltà sviluppatasi lungo la Valle del fiume Nilo fino alla foce nel Mar Mediterraneo."
* d3 = " In geomorfologia, la valle è una formazione del paesaggio terrestre che si estende in lunghezza fra due pendici montuose."

and then assume that you execute the following three queries, with a Whitespace Analyzer:

* q1 = "Valle dei Re", q2 = "il fiume scende a valle!", q3 = "luoghi montuosi egitto"

Show, for every query, which documents will be returned and in which order, commenting on your assumptions about the term frequencies.

Finally comment on the difference in the results if Lucene and each query use the Italian Analyzer.