**Information Retrieval**

**15 February 2018**

**Name: Surname: Matricola:**

**Ex 1 [points 7]** Given the four texts:

* 1. T1=”a beautiful cat”
  2. T2=”cat after cat”
  3. T3=”after a beautiful girl a cat”
  4. T4=”girl after girl”

Show the inverted list built over them by using gamma-coding over the docID gaps.

Compute the TF-IDF vectors of the four texts above (logs are in base two).

Find the most similar text to T3 in the vector space model (do not apply any normalization to the cosine score).

**Ex 2 [points 5]**

Describe the LSI technique and comment on how a query is projected into a k-dimensional space. Comment also the value and significance of k.

**Ex 3 [points 2+5]**

Describe the algorithm for extracting bigrams (collocations) based on statistical information on the constituting words and PoS tagging (noun, verb, adjective, etc.).

Describe the RAKE algorithm for extracting long keywords from texts, and comment its pros/cons.

**Ex 4 [points 6]**

Show how to synchronize via zsync the the old file “acaddbdabb” (on the client) by means of new file “bacaddabb” (on the server), setting blocks of 3 chars each.

**Ex 5 [points 5]** Given the binary vectors A = 10111, B = 10010, C = 00010, D = 00000, E = 01100; apply LSH for approximating the hamming distance by setting k=2, L=2 and projecting sets I1 = {0,3} and I2 = {3,4}.