Exam 2016, solution sketch

**SQUEEZE**

1. See Section 5.3.1 in textbook.
2. See Section 5.3.2 in textbook.
3. <http://www.uio.no/studier/emner/matnat/ifi/INF3800/v15/slides/index_compression_slides.pdf>
4. <http://www.uio.no/studier/emner/matnat/ifi/INF3800/v15/slides/index_compression_slides.pdf>

*[Grading guide: We explicitly ask for examples (assuming that being able to show good, concrete examples correlates with being better acquainted with the matter), so students that provide good examples should score higher than those that don’t or that only provide textual explanations.]*

**LUCY IN THE SKY WITH DIAMONDS**

1. Using the conservative estimate of the length of unioned postings lists, the recommended order is:  
     
   (*kaleidoscope* OR *eyes*) [87K + 213K = 300K] AND   
   (*tangerine* OR *trees*) [46K + 316K = 362K] AND   
   (*marmalade* OR *skies*) [107K + 271K = 378K]   
     
   However, depending on the actual distribution of postings, (*tangerine* OR *trees*) may well be longer than (*marmalade* OR *skies*) because the two components of the former are more asymmetric. For example, the union of 11 and 9990 is expected to be longer than the union of 5000 and 5000 even though the conservative estimate predicts otherwise.
2. No. For OR-queries we visit all postings in the posting lists of either terms, thus killing the need for skip pointers. Skip pointers are good for AND-queries, but we assume that we don’t generate skip pointer data structures for the intermediate results of the OR-queries.

*[Grading guide: This is a pretty easy question, but really good students might be able to dazzle us here. Consider reserving the maximum score for those that provide exceptional and well-thought-out answers, if any.]*

**NEW YORK DOLLS**

1. See Section 5.1.1 in textbook.
2. Assuming the rate of creation remains the same, we can assume the number of tokens doubles. Thus, there are 800M tokens. Using the equation to model the original corpus of 400M tokens, with *T* = 400M and *M* = 1M, we solve for *k*, yielding k = 50. Using this value for *k*, and *T* = 800M, we estimate that there are roughly = 1.41M tokens.

*[Grading guide: Students should clearly show how they compute k and don’t just “guess” the right answer.]*

**BAYES CITY ROLLERS**

1. See Section 13.2 in textbook.
2. See Section 13.2 in textbook.
3. See Example 13.1 in textbook. Substitute *Chinese* with *longmuir*, *Beijing* with *nobby*, etc.
4. See Example 13.1 in textbook.

*[Grading guide: Presumably straightforward, either they get it or they don’t. If they clearly fumble, they probably don’t. There was an obligatory programming assignment in the course on naïve Bayes.]*

**BEACH BOYS**

1. See Section 21.2 in textbook.
2. See Exercise 21.6 in textbook. Substitute *1* with *brian*, *2* with *dennis* and *3* with *carl*. We then have *P* as in Equation 21.3. Note that 1/6 + 1/2 = 2/3, and that 1/6 + 1/4 = 5/12.
3. See Section 21.2.2 in textbook.

*[Grading guide: Students should show how they arrive at the answers, and not just magically come up with P. Good students will probably mention both the concept of eigenvectors and iterations on the last one to get the full score.]*