1. Warm up exercise (do by yourself at class)  
**Exercise 1.1**

Consider these documents

Doc 1 - breakthrough drug for diabetes  
Doc 2 - new diabetes drug  
Doc 3 - new approach for treatment of diabetes  
Doc 4 - new hopes for diabetes patients  
  
a. Draw the term document incidence matrix for this document collection.  
b. Draw the inverted index representation for this collection.  
  
**Exercise 1.2**  
For the document collection shown in Exercise 1.1, what are the returned results for these queries?  
a. ‘diabetes’ AND ‘drug’  
b. ‘for’ AND (NOT (‘drug’ OR ‘approach’))

# II. Programming Exercise (do by yourself at home)

**Exercise 1.3**

**Objective:** The main purpose of this assignment is to review some text processing techniques and implement inverted index data structure using C++ / C# / MATLAB. You should choose one of them to be familiar with from now to the end of this semester.

You are provided a set of documents in folder “docs”. Each document has a name corresponding to its ID. Using one of above programming languages to implement the following requirements:

a) **Dictionary gathering**: Read content of all text files in “docs” folder to collect set of words that appear in at least one document. Please write list of words in a text file named “dictionary.txt”, where each word is put in a single line.

b) **Building Inverted Index**: Implement Inverted Index data structure for storing term-document relationship. To test this data structure, list all documents that contain a given term.

c) **Search engine**: using inverted index constructed from *Exercise 1.3.b* to return a list of documents which contain string of terms (input query in string format).