LAB 4: Exercises

1. Construct the binary representation of documents D1-D4 and construct the inverted index.

D1 = New solution for lung cancer
D2 = Treatment for lung cancer
D3 = New cancer treatment bring hopes
D4 = Results of cancer treatment research

Binary representation (columns)				
Term/Doc	D1	D2	D3	D4
brings				
cancer				
for				
hopes				
lung				
new				
of				
solution				
research				
results				
treatment				

Inverted index				
brings				
cancer				
for				
hopes				
lung				
new				
of				
solution				
research				
results				
treatment				

Find the relevant documents for:

- **q = „lung AND cancer:**
- **„q = cancer OR lung”**:

2. Build a suffix tree of a word “DODO$” using naïve algorithm.

3. Build a suffix tree of a word “DODO$” using Ukkonen’s algorithm.

4. Build a suffix array for a word "DODO$" using Qsufsort algorithm.

	1	2	3	4	5
S	D	O	D	O	$
I[i]					
V[i]					
V'[i] = V[I[i] + 1]					
I[i]					
V[i]					
V'[i] = V[I[i] + 2]					
I[i]					

5. There are four documents D1-D4 in the collection and one query Q. Firstly, compute a Jaccard index between every two documents (and Q). Then, Compute the cosine similarity using TF and TF-IDF models. For this purpose, compute: the number of occurrences of each term in each document, IDF coefficients, and length of each vector for TF and TF-IDF representations.

Documents and the query

Q	=	I	R		
D1	=	I	R	I	R
D2	=	R	R	R	R
D3	=	A	I	R	A
D4	=	R	A	R	A

IDF coefficients

	IDF
R	
A	
I	

Jaccard index

	Q	D1	D2	D3	D4
Q					
D1					
D2					
D3					
D4					

Cosine similarity (TF and TF-IDF)

		Term occurrences			TF			$\|v\|$	Cosine similarity	TF-IDF			$\|v\|$	Cosine similarity
		R	A	I	R	A	I			R	A	I		
Q	=													
D1	=													
D2	=													
D3	=													
D4	=													

6. Find unique 3-grams (3-shingles) for D1, D2, and D3. Then, compute a similarity (Jaccard index) between each two documents.

			3-shingles
D1	=	to be or not to be or not to be	
D2	=	to be or maybe to be or not	
D3	=	maybe to be but surely not	

Jaccard index

	D1	D2	D3
D1			
D2			
D3			