Name of the student:		**Roll No.**	
Assignment Number:		**Date of Assignment:**	
Relevant CO's	**1.Explain the basics of search engines and search engine optimization**		
Sign here to indicate that you have read all the relevant material provided before attempting this assignment			**Sign:**

Assignment grading using Rubrics

Indicator	**Very Poor**	**Poor**	**Average**	**Good**	**Excellent**
Timeline (2)	More than a session late (0)	NA	NA	NA	Early or on time (2)
Organization of content (2)	N/A	No sense of organization, Paragraphs lack clear ideas (0.5)	Some paragraphs have clear ideas, support from examples may be missing and transitions are weak (1)	Most paragraphs have clear ideas, are supported with some examples and have transitions(1.5)	All paragraphs have clear ideas, are supported with examples and have smooth transitions. (2)
Level of content (4)	N/A	Major points omitted or addressed minimally(1)	Content is sound and solid; ideas are present but not particularly developed or supported; some evidence, but usually of a generalized nature.(2)	Well-presented and argued; ideas are detailed, developed and supported with evidence and details, mostly specific. (3)	Exceptionally well-presented and argued; ideas are detailed, well-developed, supported with specific evidence & facts, as well as examples and specific details (4)
Grammar and Mechanics (2)	N/A	Spelling, punctuation and grammatical errors create distraction, making reading difficult(0.5)	Most spelling, punctuation and grammar correct. Some errors remain(1)	Few spelling, punctuation and grammatical errors allowing reader to follow ideas clearly (1.5)	Assignment is free of distracting spelling, punctuation and grammatical errors(2)

Late submission details (if any)

Reason(s) of late submission	Submission date	Actual submission date	sign of student

Assignment 1
Assignment on Search Engine Optimization

Course title: Advanced Internet Technology
Course term: 2017-2018
Instructor name: Saurabh Kulkarni

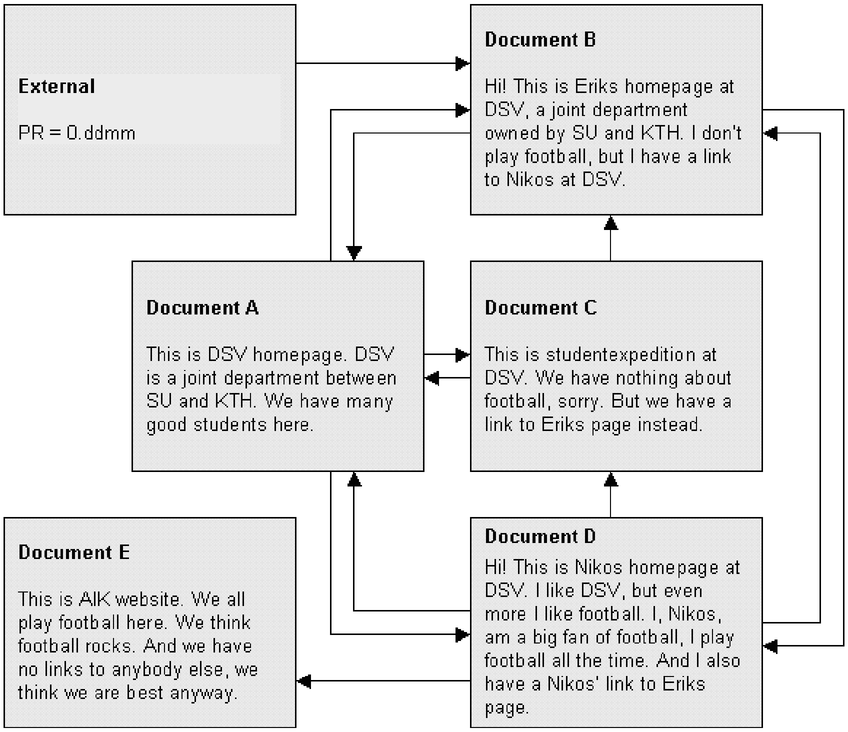

External
PR = 0.ddmm
Document B
Hi! This is Eriks homepage at DSV, a joint department owned by SU and KTH. I don't play football, but I have a link to Nikos at DSV.
Document A
This is DSV homepage. DSV is a joint department between SU and KTH. We have many good students here.
Document C
This is studentexpedition at DSV. We have nothing about football, sorry. But we have a link to Eriks page instead.
Document E
This is AIK website. We all play football here. We think football rocks. And we have no links to anybody else, we think we are best anyway.
Document D
Hi! This is Nikos homepage at DSV. I like DSV, but even more I like football. I, Nikos, am a big fan of football, I play football all the time. And I also have a Nikos' link to Eriks page.

Figure 1: Web graph

Select one of the two queries:

- DSV football
- Nikos DSV

Procedure

1. Text Similarity

Match the selected query to all the documents and rank the documents according to their similarity score. Calculate normalized document similarity, use term frequency as term weights. That is, apply the following similarity formula:

$sim q, d \ \frac{\sum_{i1}^{n} q_i.d_i}{N_d}$

where q and d are the query and a document respectively;

q_i is the i$^{\text{th}}$ term frequency in the query q;

d_i is the i$^{\text{th}}$ term frequency in the document d;

N_d is the number of all words in the document d, including stop-words.

2. Page rank

Calculate the page rank of each of the 5 documents. "0.ddmm" is the page rank value of the external document that is included in the formula as any other page rank value coming from another page. dd stands for your birth day and mm stands for your birth month. For example, if the birthday is November 24, then 0.ddmm = 0.2411. You can use MS excel or write a program to compute page rank of the page. Now remove the external document and recalculate page rank.

Typical mistakes- Be careful when you count the number of outgoing links. Document A has 3 outgoing links, Document B has 2, Document C has 2, Document D has 4, and Document E has no outgoing links. Be careful when you do Excel iterations (if you do). Check several times whether you still have the right formula in all the Excel cells.

3. Combining Text Similarity with Page Rank

Re-rank the documents using a different (but still very simple) formula that makes use of page ranks:

$SIM_1q, d \;\; simq, d \;\; 0.1.PRd$

where sim(q,d) is the old similarity value obtained in (1); PR(d) is the page rank of the document d obtained in (2) calculated using the external page rank value. Observe how the ranks of the documents change.

Once again, re-rank the documents using an equivalent formula with a different constant multiplier:

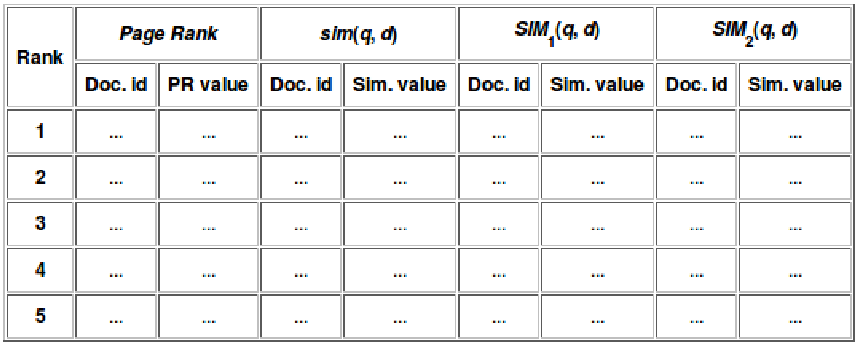
$SIM_2q, d \;\; simq, d \;\; 0.5.PRd$

Observe how the ranks of the documents change, if they change.

Attach a report to these sheets as solution to this assignment which will comprise of:

- the query you have selected;
- the eventual formula you used to calculate textual similarity sim(q, d) between the query and a document;
- the page rank formula for each document; tell how you calculated the page rank values.

In order to show the relevance of the documents to the query applying different formulas, fill in the table below.



Rank	*Page Rank*		*sim(q, d)*		*SIM$_1$(q, d)*		*SIM$_2$(q, d)*	
	Doc. id	PR value	Doc. id	Sim. value	Doc. id	Sim. value	Doc. id	Sim. value
1	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...
3	...	...	...	...	...	...	...	...
4	...	...	...	...	...	...	...	...
5	...	...	...	...	...	...	...	...

Attach the Excel sheets and program code if you have any.

References

1. "Mining of massive datasets- J.Ullman,Anand Rajaraman", website-mmds.org
2. "Art of SEO", O'reilly publication