COURSEWORK ASSIGNMENT **UNIVERSITY OF EAST ANGLIA**
School of Computing Sciences

MODULE:	**: CMP-5036A Information Retrieval**	
ASSIGNMENT TITLE:	**: Experimental project: crawl and index**	
DATE SET	**: Thursday 19 October 2017 (Week 4)**	
HAND-IN	**: Code submission Wednesday 15 November (Week 8)** **Demonstration Thursday 16 November (Week 8)**	
MARKS AVAILABLE	**: Week 12**	
ASSIGNMENT VALUE	**: 10%**	
SET BY	**: D. J. Smith**	**SIGNED:**
CHECKED BY	**: S. Taylor**	**SIGNED:**

Overview

The aim of this assignment is to construct a simple search engine. You are expected to program this in Python, although you will not be penalised for the use of another language.

Aims

- To give students practical experience of information retrieval systems, experimental work and evaluation
-

Learning outcomes

- Improved program design and coding skills.
- Improved research and communications skills.

Assessment criteria

You will submit code and present your system in a short demonstration.
Marks will be awarded for:
- the design and functionality of the system,
- the organization of the demonstration,
- the quality of your code, which you will submit.

The breakdown of marks is given on the marking sheet (attached). *Note that the distribution of marks is indicative only and may change*.

Description of the assignment

The assignment is to produce a search engine that works over the https://www.uea.ac.uk/computing domain. Use the NIST crawler as the starting point.
The task is to crawl and index the domain. You will have to strip out the formatting and make decisions about how to handle non-content material (menus, banners, …).

Required:

1. Demonstration and questions
The demonstration should show:
- your system working by crawling a small domain,
- the output from crawling the uea.ac.uk/computing domain.

You should be able to answer questions on the operation of your system.

2. Code
You must submit a copy of all the code used in your experiments as a single zipped file, via Blackboard; the filename should be of the form: *yourstudentid*_IRcoursework.zip.
Marks will be awarded for clear, well-structured code with appropriate and informative comments. Systems that are mostly built from code reused from third parties, cluttered with redundant fragments, have an inconsistent layout, are uncommented, ... will attract very few marks. Note that the unacknowledged use of code from third parties is a form of plagiarism.

CMP-5036A Marking Sheet: Crawler and indexer

Student name		**No.**	
Marker name			
Crawling 6%	*Demonstration of a crawl over a sample of the uea.ac.uk/computing domain, stripping out HTML formatting and any other non-content material using domain-independent code.*		
Indexing 5%	*Demonstration of an inverted index containing docids, postings and vocabulary tables from a crawl over the uea.ac.uk/computing domain, stored in a form suitable for subsequent retrieval.*		
Design and code 6%	*Evidence of a structured approach to design of code, use of appropriate programming conventions and good comments.*		
Demonstration 3%	*Organisation and conduct of the demonstration.*		

Additional comments

Mark	**/20**
	%