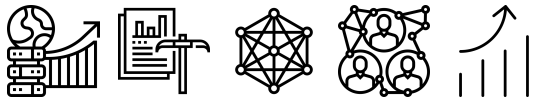


CSC 5741
Lecture 1: Administrivia and Course Overview

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia

Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
- Part V: About Next Week

Lecture Series Outline

- Part I: Administrivia
  - Personal Introductions
  - Learning Outcomes
  - Course Structure
  - Prescribed Books
  - Tools and Services
  - Course Grading, Academic Dishonesty and Course Management
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
- Part V: About Next Week

Personal Introductions

- Your full names and preferred reference (first name, Mrs./Ms.Mr. X)
  - **I prefer that you refer to me using my first name: Lighton**
- Your formal education background
- What you are presently upto (Think what you do for a living)
- What you hope to get from CSC 5741

CSC 5741 Learning Outcomes

- Identify the key processes of data mining, data warehousing and knowledge discovery process
- Describe the basic principles and algorithms used in practical data mining and understand their strengths and weaknesses
- Apply data mining techniques to solve problems in other disciplines in a mathematical way
- Apply data mining methodologies with information systems and generate results which can be immediately used for decision making in well-defined business problems

CSC 5741 Desired Outcome

- Desired outcome, for me, is to ensure we are all in a position to successfully undertake a Data-driven Research Project.
- [...]
- Data Mining "Research" Project
- Practical Knowledge
- Experimentation
- Evaluation Strategies
- Ethics and Bias
- [...]
- We will need to read and discuss what others have done

Course Structure (1/9)

- CSC 5741 is a half course
- CSC 5741 will be run using a seminar session
  - One three hour-long lecture session per week
    - One seminar every fortnight.
    - Paper reading sessions every fortnight.
    - One class theory tests

Course Structure (2/9)

- Tentative Lecture series and session structure
  - Lecture session (120 minutes)
  - Paper discussion (30 minutes)
  - Seminar session (30 minutes)
- We will tentatively spend two weeks on each CSC 5741 theme

TERM I	
Monday 28th January, 2019 to Sunday 10th March, 2019	On-line Registration
Monday 4th February, 2019 to Friday 8th February, 2019	Orientation of First Year students
Sunday 10th February, 2019	Arrival of Returning Regular Students
Monday 18th February, 2019 to Friday 31st May, 2019	Lectures for Regular Students (15 weeks)
Monday 4th March, 2019 to Thursday 7th March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11th March, 2019 to Sunday 24th March, 2019	Late Registration Period (2 Weeks)
Friday 26th April, 2019	Social and Cultural Day
Monday 13th May, 2019 to Friday 26th July, 2019	IDE Students School Experience (11 Weeks)
Monday 20th May, 2019 to Thursday 24th May, 2019	Graduation Week (Second Graduation Ceremony)
Monday 3rd June, 2019 to Friday 7th June, 2019	Study Break and Post Graduate Seminar Week

Course Structure (3/9)

- **Lecture sessions**
  - Basic introduction to core concepts
  - Practical walkthroughs

TERM I	
Monday 28th January, 2019 to Sunday 10th March, 2019	On-line Registration
Monday 4th February, 2019 to Friday 8th February, 2019	Orientation of First Year students
Sunday 10th February, 2019	Arrival of Returning Regular Students
Monday 18th February, 2019 to Friday 31st May, 2019	Lectures for Regular Students (15 weeks)
Monday 4th March, 2019 to Thursday 7th March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11th March, 2019 to Sunday 24th March, 2019	Late Registration Period (2 Weeks)
Friday 26th April, 2019	Social and Cultural Day
Monday 13th May, 2019 to Friday 26th July, 2019	IDE Students School Experience (11 Weeks)
Monday 20th May, 2019 to Thursday 24th May, 2019	Graduation Week (Second Graduation Ceremony)
Monday 3rd June, 2019 to Friday 7th June, 2019	Study Break and Post Graduate Seminar Week

http://www.unza.zm

Course Structure (4/9)

- **Paper discussions**
  - Explore problems tackled by other researchers
  - Implicitly look at aspects that will not be explicitly discussed, e.g. ethics and experimentation

TERM I	
Monday 28th January, 2019 to Sunday 10th March, 2019	On-line Registration
Monday 4th February, 2019 to Friday 8th February, 2019	Orientation of First Year students
Sunday 10th February, 2019	Arrival of Returning Regular Students
Monday 18th February, 2019 to Friday 31st May, 2019	Lectures for Regular Students (15 weeks)
Monday 4th March, 2019 to Thursday 7th March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11th March, 2019 to Sunday 24th March, 2019	Late Registration Period (2 Weeks)
Friday 26th April, 2019	Social and Cultural Day
Monday 13th May, 2019 to Friday 26th July, 2019	IDE Students School Experience (11 Weeks)
Monday 20th May, 2019 to Thursday 24th May, 2019	Graduation Week (Second Graduation Ceremony)
Monday 3rd June, 2019 to Friday 7th June, 2019	Study Break and Post Graduate Seminar Week

http://www.unza.zm

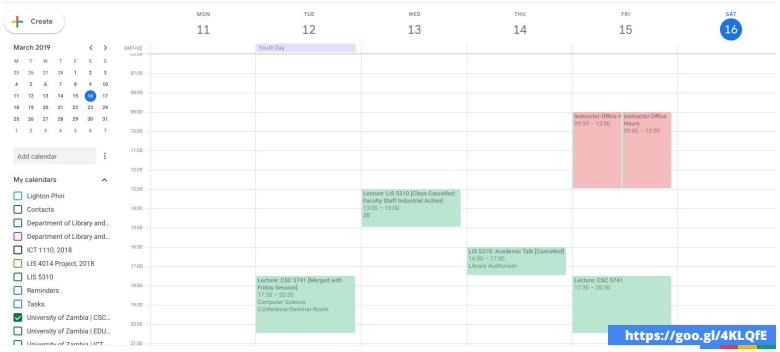
Course Structure (5/9)

- **Seminars**
  - Academic talks by current and former students
  - Industry talks from entities that employ data mining techniques

TERM I	
Monday 28th January, 2019 to Sunday 10th March, 2019	On-line Registration
Monday 4th February, 2019 to Friday 8th February, 2019	Orientation of First Year students
Sunday 10th February, 2019	Arrival of Returning Regular Students
Monday 18th February, 2019 to Friday 31st May, 2019	Lectures for Regular Students (15 weeks)
Monday 4th March, 2019 to Thursday 7th March, 2019	Graduation Week (First Graduation Ceremony)
Monday 11th March, 2019 to Sunday 24th March, 2019	Late Registration Period (2 Weeks)
Friday 26th April, 2019	Social and Cultural Day
Monday 13th May, 2019 to Friday 26th July, 2019	IDE Students School Experience (11 Weeks)
Monday 20th May, 2019 to Thursday 24th May, 2019	Graduation Week (Second Graduation Ceremony)
Monday 3rd June, 2019 to Friday 7th June, 2019	Study Break and Post Graduate Seminar Week

http://www.unza.zm

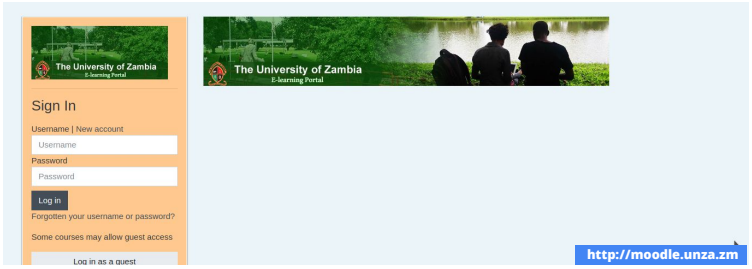
Course Structure (6/9)



https://goo.gl/4KLQfE

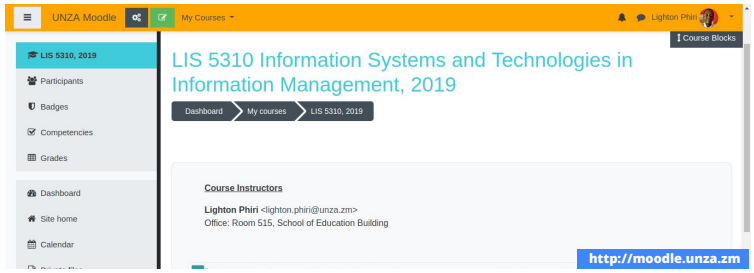
Course Structure (7/9)

- **Course Resources**
  - All course resources will be made available on The Moodle.



Course Structure (8/9)

- **Course Resources**
  - All course resources will be made available on The Moodle.

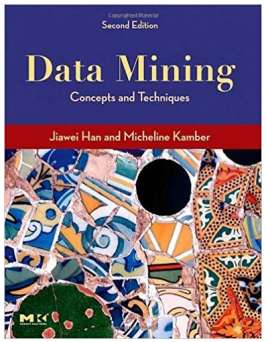


Course Structure (9/9)

- **Additionally, course resources will be disseminated as follows:**
  - Large files such as videos and software tools will be made available on the instructor's profile Website

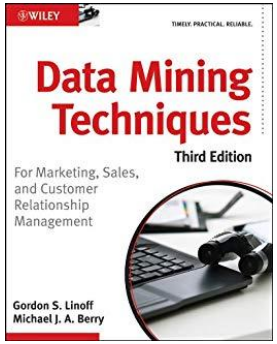
Prescribed and Recommended Textbooks (1/4)

- **Data Mining Concepts and Techniques**
  - J. Han and M. Kamber (2011)



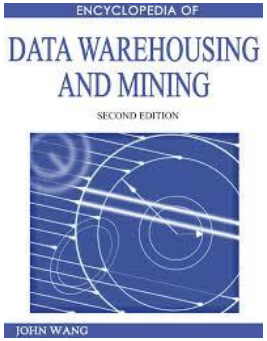
Prescribed and Recommended Textbooks (2/4)

- **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**
  - G. S. Linoff and M. J. Berry (2011)



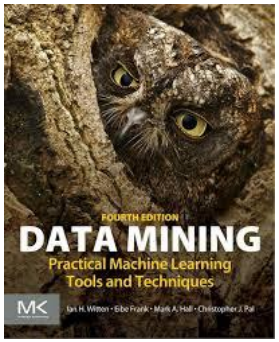
Prescribed and Recommended Textbooks (3/4)

- **Encyclopedia of Data warehousing and Mining**
  - J. Wang (2005)



Prescribed and Recommended Textbooks (4/4)

- **Data Mining: Practical Machine Learning Tools and Techniques**
  - I. H. Witten, E. Frank and M. A. Hall

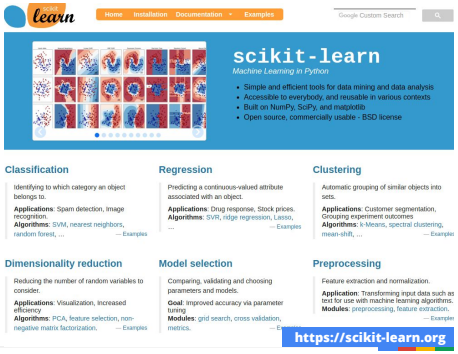


Tools and Services (1/4)

- **Tools and services**
  - VirtualBox for creating virtual environments for running Ubuntu 18.04.
  - Ubuntu 18.04 for running all practical-oriented activities and tasks.
  - Python for scripting
  - TensorFlow
  - Weka

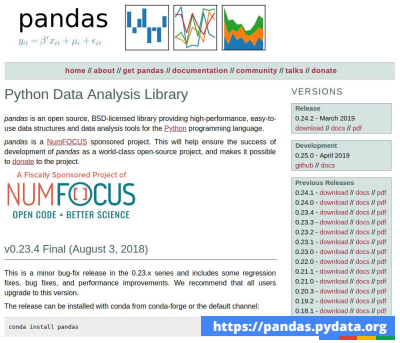
Tools and Services (2/4)

- **scikit-learn**
  - Python machine learning library
  - Implements most of the algorithms we will be exploring



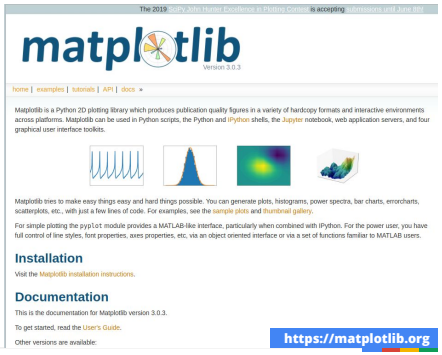
Tools and Services (3/4)

- **pandas**
  - Python data analysis library



Tools and Services (4/4)

- **Matplotlib**
  - Graphical representation during EDM and analysis



Course Grading (1/4)

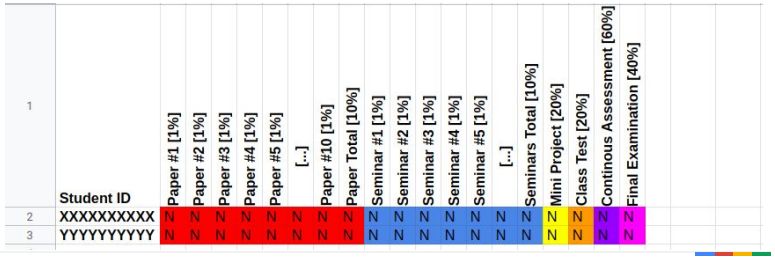
- **10% Paper readings**
  - Paper summaries of peer-reviewed publications.
- **5% Seminar presentations**
  - Questions and discussions during seminars and reading sessions. Marks awarded for participation.
- **5% Class participation**
  - Discussion in class.
- **20% Practical Projects**
  - Hands-on practical project assignments that will involve a project deliverable.

Course Grading (2/4)

- **20% Class Theory Test**
  - One 90 minutes-long class test will be held towards the end of Term #1
- **40% Final Examination**
  - The final examination is based on the entire course outline.

Course Grading (3/4)

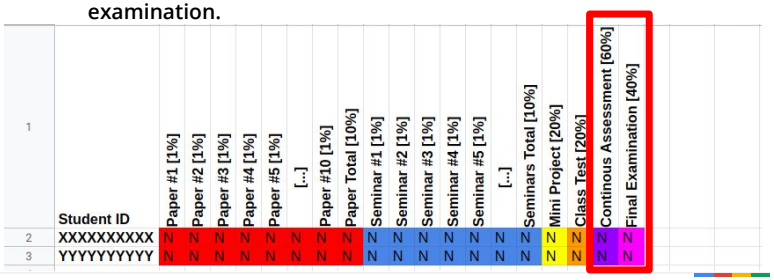
- **Final grading is based on a 60/40 split**
  - You MUST pass both the continuous assessment and examination.



	Student ID	Paper #1 [1%]	Paper #2 [1%]	Paper #3 [1%]	Paper #4 [1%]	Paper #5 [1%]	[...]	Paper #10 [1%]	Paper Total [10%]	Seminar #1 [1%]	Seminar #2 [1%]	Seminar #3 [1%]	Seminar #4 [1%]	Seminar #5 [1%]	[...]	Seminars Total [10%]	Mini Project [20%]	Class Test [20%]	Continous Assessment [60%]	Final Examination [40%]
2	XXXXXXXXXX	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
3	YYYYYYYYYY	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Course Grading (3/4)

- **Final grading is based on a 60/40 split**
  - You MUST pass both the continuous assessment and examination.

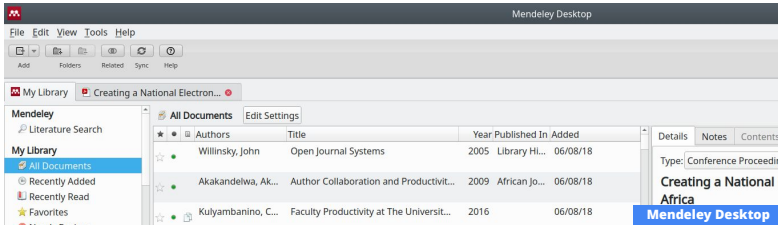


	Student ID	Paper #1 [1%]	Paper #2 [1%]	Paper #3 [1%]	Paper #4 [1%]	Paper #5 [1%]	[...]	Paper #10 [1%]	Paper Total [10%]	Seminar #1 [1%]	Seminar #2 [1%]	Seminar #3 [1%]	Seminar #4 [1%]	Seminar #5 [1%]	[...]	Seminars Total [10%]	Mini Project [20%]	Class Test [20%]	Continous Assessment [60%]	Final Examination [40%]
2	XXXXXXXXXX	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
3	YYYYYYYYYY	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Course Grading (4/4)

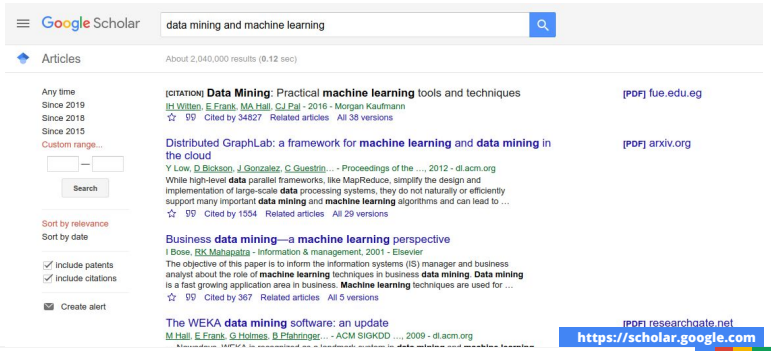
GRADE	DESCRIPTION	SCORE RANGE	GRADE POINT
A+	DISTINCTION	86–100	5
A	DISTINCTION	75–85	4
B+	MERITORIOUS	70–74	3.5
B	CREDIT	65–69	3
C+	CREDIT	55–64	2.37
C	PASS	50–54	1.5
D	FAIL	<49	0

Readings and Paper Summaries (1/5)

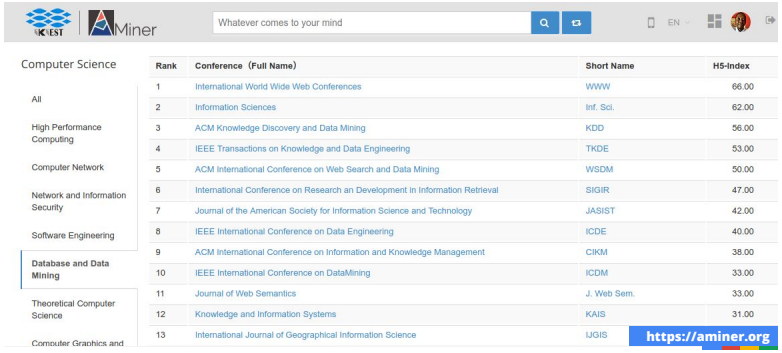


- **You MUST develop the habit of using a bibliographic manager**
  - Mendeley, Zotero, JabRef, RefWorks

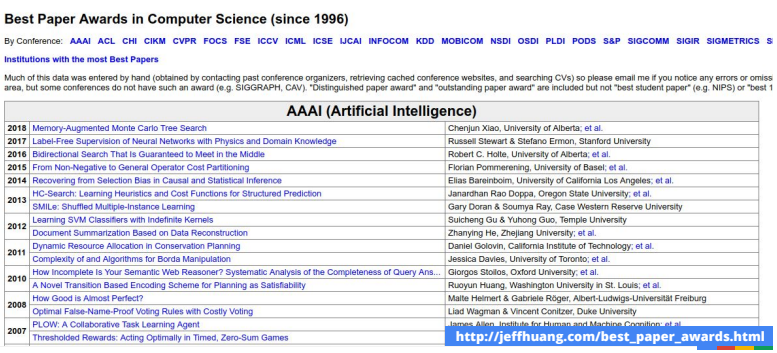
Readings and Paper Summaries (2/5)



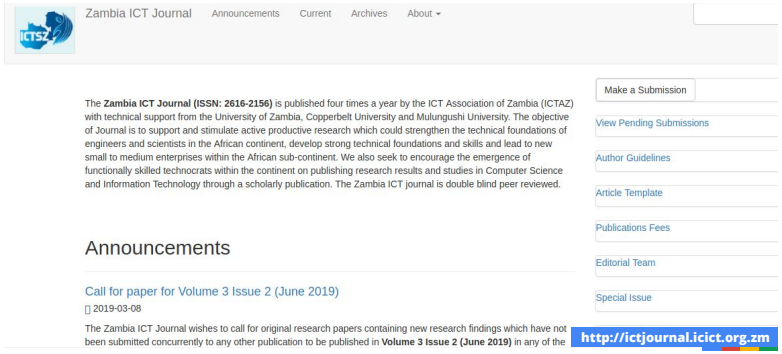
Readings and Paper Summaries (3/5)



Readings and Paper Summaries (4/5)



Readings and Paper Summaries (5/5)



On How to Read a Paper (1/5)

On How to Read a Paper (2/5)

- Title
- Abstract
- Introduction
- Related Work
- Implementation
- Evaluation
- Discussion
- Conclusion
- References

On How to Read a Paper (3/5)

- Keshav's Three Pass Approach is very helpful when initially getting started.
  - Pass #1
    - Title -> Abstract -> Introduction
    - Sections and subsections -> Conclusion -> References
    - Outcome of pass: paper classification, context, correctness, contributions, clarity
  - Pass #2
  - Pass #3

On How to Read a Paper (4/5)

- **Keshav's Three Pass Approach is very helpful when initially getting started.**
  - Pass #1
  - Pass #2
    - Analyse floats
    - Note key references not read
    - Outcome: Firm understanding of paper
  - Pass #3

On How to Read a Paper (5/5)

- **Keshav's Three Pass Approach is very helpful when initially getting started.**
  - Pass #1
  - Pass #2
  - Pass #3
    - Outcome: Identify potential flaws with experimental designs and analyses.

Mini Project Overview (1/2)

- **Ideally, everyone will work on an individual project, however, the themes will be similar**
  - Mini Project descriptions
  - Grade distribution
  - Open date and deadline
- **Deliverables**
  - Properly curated datasets
  - Well-organised scripting code (Github/Bitbucket)
  - Technical report

Mini Project Overview (2/2)

- **The 20% score allocated to the Mini Project will be distributed**
  - Implementation of chosen problem
  - Presentation of implementation
  - Technical report based on implementation
- **Individual problem sets**
- **Open date: March 29, 2019**
- **Due date: May 17, 2019**

ASPECT	SCORE
1. IMPLEMENTATION (data collection, data cleaning, functionality)	8%
2. TECHNICAL REPORT (presentation, layout, style, language, evaluation, analysis)	8%
3. PRESENTATION (Q&A, style, presentation slides)	4%

Academic Dishonesty

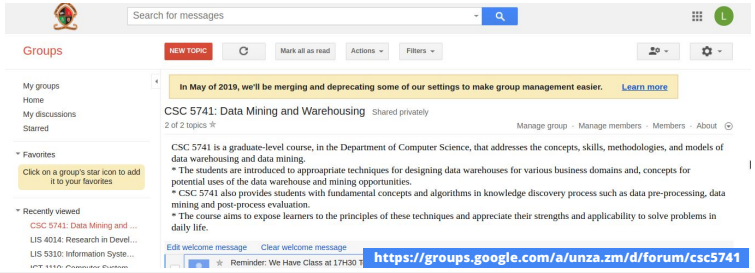
- **Every assessment submitted must be your own work. Academic dishonesty of any form is considered very seriously.**
  - NOTE: Any form of academic dishonesty (plagiarism, copying, cheating etc) will result in a ZERO mark for the entire continuous assessment score.

Course Management (1/2)

- **Instructor: Lighton Phiri and TBA (possible Invited Talks)**
- **Email: lighton.phiri@unza.zm**
- **Office: Room 515, Fifth Floor, School of Education Building**
- **Office hours: Friday 09H00–13H00**
  - Alternatively, schedule an appointment via email after checking free/busy slots on my calendar (https://goo.gl/6kHrnA)

Course Management (2/2)

- **Communication exclusively done electronically**
  - The Moodle, Course Mailing List and Email



Q & A Session

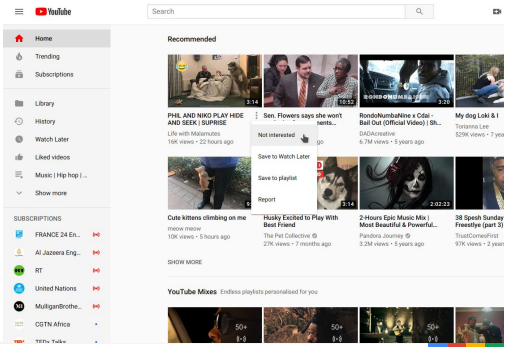
- **Comments, concerns and complaints?**

Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
  - Contextualising Data Mining and Warehousing
  - CSC 5741 Themes and Topics
- Part III: On Academic Activities
- Part IV: Getting Started With Python
- Part V: About Next Week

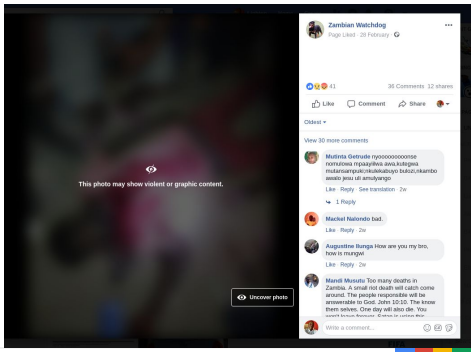
Contextualising Data Mining & Warehousing (1/14)

- With the ever increasing amount of data being generated, application of data mining techniques are increasing.



Contextualising Data Mining & Warehousing (2/14)

- Effective ways are needed to automatically make sense out of digital content.
  - Relevance
  - Recommendation
  - Restricted and obscene materials



Contextualising Data Mining & Warehousing (3/14)

- Past CS@ UNZA Dissertations
  - Lillian Muzyece (2019). Automatic Weather Prediction
  - Soft Mulizwa (2019). Automatic Customer Segmentation for effective Targeted Campaigns
  - Friday Chazanga (2019). Automatic Number Plate Recognition
- Current CS@ UNZA Dissertations
  - Simon Hawatichke Chiwamba (2019—). Machine Learning Automated Image Capture and Identification of Fall Armyworm
  - Francis Chulu (2019—). Automatic identification and early warning and monitoring web based system of fall Armyworm

Contextualising Data Mining & Warehousing (4/14)

- **Automatic classification of scholarly research**
  - Automatic generation of metadata
  - Automatic reclassification of digital objects


```
▼<header>
   <identifier>oai:dspace.cbu.ac.zm:123456789/6
   <datestamp>2011-08-18T08:59:44Z</datestamp>
   <setSpec>hdl_123456789_23</setSpec>
 </header>
▼<metadata>
  ▼<oai_dc:dc xmlns:oai_dc="http://www.openarch
   instance" xsi:schemaLocation="http://www.ope
    ▼<dc:title>
       Customer service management in the retail
     </dc:title>
     <dc:creator>Atanga, Muyenga</dc:creator>
     <dc:subject>Banks</dc:subject>
     <dc:subject>Retail banking</dc:subject>
     <dc:subject>MBA THESIS</dc:subject>
     <dc:subject>Customer service</dc:subject>
     <dc:description>v,116p.</dc:description>
     <dc:description>Copperbelt University, Sch
     <dc:date>2011-07-19T14:32:14Z</dc:date>
     <dc:date>2011-07-19T14:32:14Z</dc:date>
     <dc:date>2011-07-19</dc:date>
     <dc:type>Thesis</dc:type>
```


Contextualising Data Mining & Warehousing (5/14)

- **University of Zambia Ranking Committee Research Report**
  - Mining for scholarly output on the Web


```
Staff ID|Title|URL|Publication Date|||||||Except
007132.csv|The moral injury symptom scale-military version|http://scholar.google.com/https://
E=#018|None|http://scholar.google.com/scholar?cites=5193522035196824996 6as_sdt=2005&sciodt=0,5
sdt=0,5&as_ylo=2018&as_yhi=2019|None|The purpose of this study was to develop a multi-dimensi
asure in intervention studies that target moral injury (MI) in Veterans and Active Duty Milit
007132.csv|Patient-reported outcomes in cirrhosis: A scoping review of the literature|http://
ep.29756|2018|8|5|1.86051127466716E+019|None|http://scholar.google.com/scholar?cites=106051127
ster=10605112746671039020&hl=en&as_sdt=0,5&as_ylo=2018&as_yhi=2019|None|Patients with cirrhos
faction with the care they receive. However, these patient-reported outcomes (PROs) are not r
bers …
007132.csv|Executive summary of the Merck Animal Health Veterinary Wellbeing Study|http://sch
31|2018|1|2|1.2366532437836 7E+019|None|http://scholar.google.com/scholar?cites=12366532437836
12366532437836663220&hl=en&as_sdt=0,5&as_ylo=2018&as_yhi=2019|None|Still, questions remain as
widespread it is, and whether it has risen to crisis levels. To date, no definitive assessme
terinarians …
007132.csv|Screening for Moral Injury: The Moral Injury Symptom Scale-Military Version Short
t/183/11-12/e059/49342292|2018|2|710.9807949189933542E+018|None|http://scholar.google.com/scholar
.com/scholar?cluster=4980794918993354212&hl=en&as_sdt=0,5&as_ylo=2018&as_yhi=2019|None|Introd
ry Symptom Scale-Military Version (MISS-M) to use when screening for moral injury and monitor
A …
007132.csv|A CHAF1B-dependent molecular switch in hematopoiesis and leukemia pathogenesis|htt
510018384707|2018|1|513.612394113417 71E+018|None|http://scholar.google.com/scholar?cites=3612
?cluster=3612394113417710872&hl=en&as_sdt=0,5&as_ylo=2018&as_yhi=2019|None|CHAF1B is the p60
for assembly of histones H3. 1/H4 heterodimers at the replication fork during S phase. Here w
007132.csv|Moral injury, religiosity, and suicide risk in US veterans and active duty militar
/advance-article-abstract/doi/10.1093/milmed/usy148/5038451|2018|1|51.6340320721551 9E+019|No
iodt=0,5&hl=en|http://scholar.google.com/scholar?cluster=16340328721551867141&hl=en&as_sdt=0,
al injury (MI) is related to greater suicide risk among Veterans and Active Duty Military (V/
moderating effect of religiosity on this relationship in …
007132.csv|An interferon-driven oxysterol-based defense against tumor-derived extracellular v
s/pii/S1535610818385555|2019|1|413.9165578243661E+018|None|http://scholar.google.com/scholar
```


Contextualising Data Mining & Warehousing (6/14)

- **LMS Log Mining**
  - Moodle usage logs


```
1###1##Moodle at University of Zambia###moodle##########site###1###3###0###0###0##
01760###2016-11-22 11:33:45
2###25###4940086###VMM 7802 Health Economics, Policy, Monitoring and Evaluation###VMM 7802###V
0##########1479915665###1525439465###0###1###0###1535578236###2016-11-23 17:41:05
3###25###4940089###VMM 7501 Principles of Epidemiology and Biostatistics###VMD 7501###VMD 7501
#######1479916418###1530255320###0###1###0###1535391769###2016-11-23 17:53:38
&###25###4940813###Socio-Anthropology###VMM7412###VM7412#######1###topics###1###2###1479852080
#1###0###1535391769###2016-11-23 17:53:39
5###25###4940081###Applied Environmental Health, Water and Sanitation###VMM 7312###VMM 7312###
##########1479916423###1480861951###0###1###0###1535391769###2016-11-23 17:53:43
7###25###4940082###Applied Food Microbiology and Nutritional Toxicology###VMM 7120###VMM 7120#
#######1479916432###1480860975###0###1###0###1535391769###2016-11-23 17:53:52
9###25###4940010###VMM 8001 Research Methodology###VMM 8901###VMM 8901#######1###topics###1###5
5439517###0###0###0###1535391769###2016-11-23 17:57:44
11###25###4940083###Ethics in Food Safety Practice###VMM 8911###VMM 8911#######1###topics###1##
492893217###0###1###0###1535391769###2016-11-23 17:59:57
14###25###4940085###Food Safety Managemnt###VMM 7501###VMM 7501#######1###topics###1###1###1543
###0###1###0###1535391769###2016-11-23 18:04:53
16###25###4940012###VMM 8201 Risk Analysis and Surveillance###VMM 8201###VMM 8201#######1###top
917386###1525439575###0###0###0###1535550221###2016-11-23 18:09:46
17###220###5120804###VMD 6808 Veterinary Public Health###VMD 6808###VMD 6801#######1###topics##
###1524427683###0###0###0###1535391769###2016-11-23 21:19:13
19###25###4940807###Health Promotion,Education and Communication###VMM8711###VMM8711#######1###
79976494###1480862724###0###1###0###1535391769###2016-11-24 08:54:54
20###25###4940014###Zoonotic Diseases and Infections###VMM 7610#########1###topics###1###2###1
890###0###1###0###1535391769###2016-11-24 10:41:59
1.1###25###4940011###Research Paper###VMM 8908###VM 8908###<p><br></p><p><strong>
\
</strong><span lang="EN">The RESEARCH PROJECT is an important component of these programme and
ded the degree of Master of Science in One Health Food Safety. The project is not only importa
erves as the final test of students' capability to work independently and think critically. It
he sense that researchers attempt to build on and improve upon previous work' </i>(Johnson 199
p;that will result in writing a research paper that will be evaluated and graded by your super
s short duration of the time allocated for research, you will have limited time and therefore
```


Contextualising Data Mining & Warehousing (7/14)

- **Mwabu Tablet Usage Analysis**
  - Android app usage and interaction logs
  - Interaction patterns for learners and educators

Contextualising Data Mining & Warehousing (8/14)

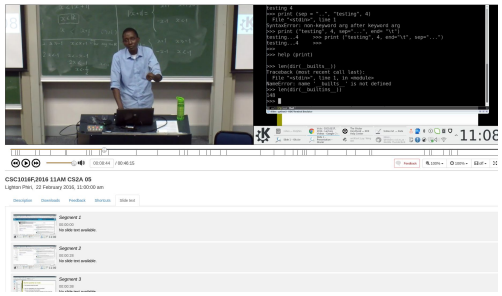
- **Effectiveness of FISP Programme Using 'Tripple Effect' Method**
  - Collaboration with two economists

```
> colnames(dataset_cfs0405_crop)
 [1] "PROV"     "DIST"     "CONST"    "WARD"     "REGION"   "CSA"
 [7] "SEA"      "HHNUM"    "CROP"     "ID009"    "S1AFIELD" "S1ACF01"
[13] "S1ACF02"  "S1ACF03"  "S1ACF04"  "S1ACF05"  "S1ACF06"  "S1ACF07"
[19] "S1ACF08"  "S1ACF09"  "S1ACF10"  "S1ACF11"  "S1ACF12"  "S1ACF13"
[25] "S1ACF14"  "S1ACF15"  "S1ACF16"  "TOTHARV"  "WEIGHT"   "HA_HARV"
[31] "convert"  "HA_PLANT"
> head(dataset_cfs0405_crop)
    PROV    DIST CONST WARD REGION CSA SEA HHNUM                CROP ID009
1 Central Chibombo    1    1    1 14    1    55               Maize     1
2 Central Chibombo    1    3    1  2    2    77               Maize     1
3 Central Chibombo    1    3    1  2    2    33 Other crops (specify)     2
4 Central Chibombo    2   12    1  2    3    96               Maize     3
5 Central Chibombo    2   12    1  2    3    96          Groundnuts     3


> colnames(dataset_cfs2004_2005_weight)
[1] "ID001"  "ID002"  "ID003"  "ID004"  "ID005"  "ID006"  "ID007"  "ID009"
[9] "WEIGHT"
> head(dataset_cfs2004_2005_weight)
    ID001    ID002 ID003 ID004 ID005 ID006 ID007 ID009    WEIGHT
1 Central Chibombo    1    1    1    14    1    1 388.84409
```


Contextualising Data Mining & Warehousing (9/14)

- **Open Matterhorn Video Segmentation Analysis**
  - Seeking to points of interest



Contextualising Data Mining & Warehousing (10/14)

- **Automatic Content Generation**
  - Underrepresentation on platforms like Wikipedia
  - We have VERY few Textbooks!!!

Contextualising Data Mining & Warehousing (11/14)

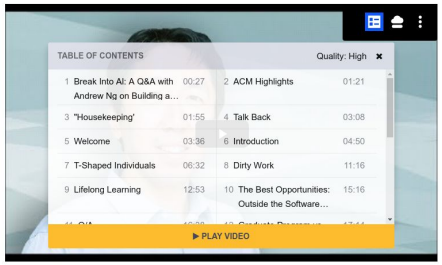
- **There is more out there [...]**
  - Parliament TV? Video and audio analysis
  - Tollgates! Automatic detection of vehicles
  - Automatic prediction of learning outcomes
  - [...]
  - [...]
  - Sentiment analysis: Popular Zambian Facebook pages, Twitter
  - Opinion mining from social media
  - What are people discussing on platforms like WhatsApp?
  - What if we harvested articles written in mainstream newspaper articles

Contextualising Data Mining & Warehousing (12/14)

- **At the rate data is being generated, we will have an endless list of data mining problems to work on.**
  - What problems to work on?
  - [...]
  - [...]

Break Into AI: Building a Career in Machine Learning with Andrew Ng

December 4, 2018



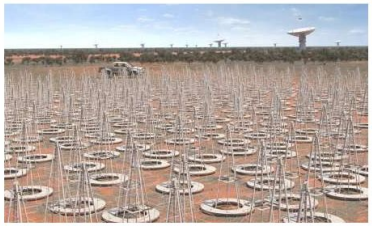
Andrew Ng will share tips and tricks on how to break into AI. He will discuss some of the most valuable

Contextualising Data Mining & Warehousing (13/14)

- **Curiosity-driven research**
  - Puzzles
  - Games

The rise of machine learning in astronomy

September 4, 2018, Particle



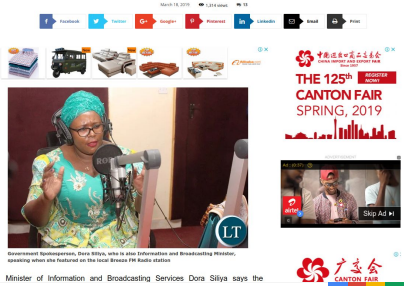
The SKA will have over 2000 radio dishes and 2 million low-frequency antennas once finished. Credit: The Squ

When mapping the universe, it pays to have some smart programming. Experts sh future of astronomy

Contextualising Data Mining & Warehousing (14/14)

- **Impact-driven research/studies**
  - Education
  - Health
  - So-called ICT for development perhaps?

Government wants help with with monitoring content from Radio and TV stations in Zambia-Siliya



Minister of Information and Broadcasting Services Dora Siliya says the

Introduction (1/2)

- **Identify the key processes of data mining, data warehousing and knowledge discovery process**
- **Describe the basic principles and algorithms used in practical data mining and understand their strengths and weaknesses**
- **Apply data mining techniques to solve problems in other disciplines in a mathematical way**
- **Apply data mining methodologies with information systems and generate results which can be immediately used for decision making in well-defined business problems**

Introduction (2/2)

- Data Mining and Data Pre-processing
- Data Warehousing
- Classification
- Associative Rule Mining
- Clustering Analysis

Theme #1: Data Mining and Data Pre-processing

- Data Mining vs. Statistics
- Knowledge discovery process
- Machine learning
- Pattern recognition
- Data cleaning
- Data integration
- Data selection
- Data transformation
- Pattern evaluation
- Knowledge presentation

Theme #2: Data Warehousing

- Decision support system
- Data warehouse architecture
- Online transaction processing
- Online analytical processing
- Star schema, Snowflake schema
- Fact constellation
- Dimension Tables and Fact tables
- Data Granularity
- Data cube
- Pivot, slice and dice, roll-up and drill down

Theme #3: Classification

- Decision Tree; Hunt's Algorithm; C4.5; Tree Induction; Binary split and Multi-way
- split; Measures of Impurity: Gini Index, Entropy and Misclassification error; Rule-
- Based Classifier; Coverage and Accuracy; Mutually exclusive and exhaustive rules;
- Ripper; Rule Pruning; Instance-Based Classifiers; Nearest neighbour classification;
- Probabilistic classifier; Naïve Bayes classifier.

Theme #4: Associative Rule Mining

- Rule Evaluation Metrics: Support and confidence
- Frequent Itemsets, Maximal
- Frequent Itemset, Closed Frequent Itemsets
- Brute-force approach
- Apriori principle
- Frequent-Pattern Tree
- Prefix paths, Conditional FP-Tree
- Rule Generation

Theme #5: Clustering Analysis

- Intra-cluster distances, Inter-cluster distances
- Partitional clustering
- K-means
- Centroid; Sum of Squared Error
- Hierarchical clustering
- Agglomerative and divisive
- Dendrogram
- Single linkage, complete linkage and group average
- Ward's Method.

Closing CSC 5710 Remarks

- Beyond CSC 5741
  - Research focus
  - Vision 2030
- About assessments
  - Ensure all assessments are attempted
- Academic dishonesty
  - NOTE: Any form of academic dishonesty (plagiarism, copying, cheating etc) will result in a ZERO mark for the entire continuous assessment score.

Q & A Session

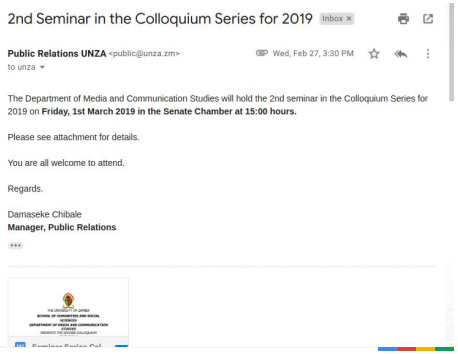
- Comments, concerns and complaints?

Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
  - Public Talks
  - Public Oral Examinations
  - DRGS Organised Events
- Part IV: Getting Started With Python
- Part V: About Next Week

Public Talks

- Make time to attend public academic talks irrespective of whether it is computing related
  - Inspiration for potential topics next year
  - Potential collaboration



2nd Seminar in the Colloquium Series for 2019 Inbox ×

Public Relations UNZA <public@unza.zm> Wed, Feb 27, 3:30 PM

to unza

The Department of Media and Communication Studies will hold the 2nd seminar in the Colloquium Series for 2019 on **Friday, 1st March 2019 in the Senate Chamber at 15:00 hours.**

Please see attachment for details.

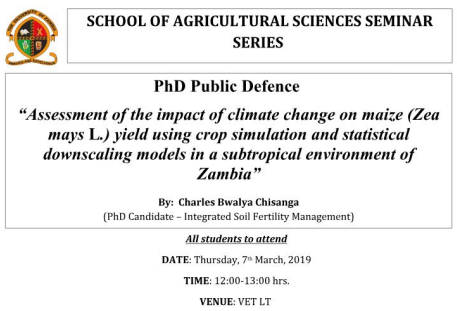
You are all welcome to attend.

Regards.

Damaseke Chibale
Manager, Public Relations

Public Oral Examinations

- Make time to attend public oral examinations so you have an idea what to expect.



SCHOOL OF AGRICULTURAL SCIENCES SEMINAR SERIES

PhD Public Defence

"Assessment of the impact of climate change on maize (Zea mays L.) yield using crop simulation and statistical downscaling models in a subtropical environment of Zambia"

By: Charles Bwalya Chisanga
(PhD Candidate – Integrated Soil Fertility Management)

All students to attend

DATE: Thursday, 7th March, 2019

TIME: 12:00-13:00 hrs.

VENUE: VET LT

DRGS Organised Events

- You want to attend important postgraduate events in order to gain a sense of what is expected
  - Announcements are sent through to your official UNZA-assigned email addresses.

Monday 13th May, 2019 to Friday 26th July, 2019	IDE Students School Experience (11 Weeks)
Monday 20th May, 2019 to Thursday 24th May, 2019	Graduation Week (Second Graduation Ceremony)
Monday 3rd June, 2019 to Friday 7th June, 2019	Study Break and Post Graduate Seminar Week
Wednesday 2nd October, 2019	Senate Curriculum and Examinations Committee
Monday 14th October, 2019 to Friday 18th October, 2019	Study Break and Post Graduate Seminar Week
Friday 15th November, 2019	
Saturday 16th November, 2019	Vacation for Regular Students Starts
Monday 24th November, 2019 to Friday 29th November, 2019	Deferred examination (5 Days)
Friday 29th November, 2019	Senate Examination and Irregularities Committee

Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
  - Installation and Setup
  - Basics
  - Data Structures
  - Flow Control
  - Functions and Modules
- Part V: About Next Week

Getting Started With Python (1/2)

```
Python 3.6.7 (default, Oct 22 2018, 11:32:17)
[GCC 8.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> import this
The Zen of Python, by Tim Peters

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
```


Getting Started With Python (2/2)

- Python is an interpreted language
- Python is a scripting language
- Python is a general purpose language
- Python is an Object Oriented language
- [...]
- [...]
- We recommend using Python 3

Installation and Setup

- [...]

Installation and Setup

- Python statements can be executed directly from the interpreter
- Python scripts can be executed as shell commands

Basics

- No need to specify data types on variable declaration
- Indentation is important

Data Structures

- Tuple
  - var = (1, 2, 3, 4, 5)
- List
  - var = [1, 2, 3, 4, 5]
- Dictionary
  - var = {“one”:1, “two”:2, “three”:3, “four”:4, “five”:5}
- Set
  - var = {1, 2, 3, 4, 5}

Loops

```
for i in [1,2,3]:
    print(i)

while i < 5:
    i += 1
    print(i)
```

- No curly braces or "end for"
- Structure is derived from level of indentation
- One statement per line
- No semicolons required

Loops

```
def csc5741(x, y='Y', z='Z'):
    print(x + ' ' + y)
return 0

csc5741('Xxxx', 'Yyyyy')
```

- All arguments are named
- Naming useful for optional arguments
- Return is optional

Modules (1/2)

- Modules facilitate extensibility and reusability
- [...]
- [...]
- from math import sqrt
- import math

Modules (2/2)

- Modules come with documentation
- Module documentation also accessible via interpreter

Installing pandas, scikit-learn and matplotlib

- pip

Installation and Setup

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
  - Installation and Setup
  - Basics
  - Data Structures
  - Flow Control
  - Functions and Modules
- Part V: About Next Week

Lecture Series Outline

- Part I: Administrivia
- Part II: Course Introduction
- Part III: On Academic Activities
- Part IV: Getting Started With Python
- Part V: About Next Week
  - Data Mining and Data Processing
  - Getting Started: SciKit-learn, Pandas
  - Mini Project Descriptions
  - Paper Reading List [Trial]
  - Academic Talk: L. Phiri [Trial]

Theme #1: Data Mining and Data Processing

- Data Mining vs. Statistics
- Knowledge discovery process
- Machine learning
- Pattern recognition
- Date cleaning
- Data integration
- Data selection
- Data transformation
- Pattern evaluation
- Knowledge presentation

Getting Started with Python, SciKit-learn & Pandas

- Tools installation
- Common commands
- SciKit-learn
- Pandas
- Sample datasets

Paper Reading List [Trial]

[1] S. Keshav (2007) "How to Read a Research Paper" https://doi.org/10.1145/1273445.1273458

[2] P. W. L. Fong (2004) "How to Read a CS Research Paper?" https://doi.org/10.1145/1595453.1595493

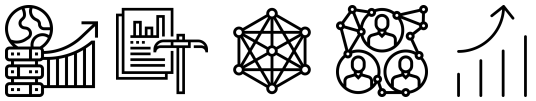
[3] L. Phiri (2018) "Research Visibility in the Global South: Towards Increased Online Visibility of Scholarly Research Output in Zambia" http://lis.unza.zm/~lightonphiri/papers/paper-icict18-online_visibility.pdf

Academic Talk: L. Phiri [Trial]

[1] L. Phiri (2018) "Research Visibility in the Global South: Towards Increased Online Visibility of Scholarly Research Output in Zambia" http://lis.unza.zm/~lightonphiri/papers/paper-icict18-online_visibility.pdf

Bibliography

[1] CSC 5741 Syllabus http://lis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741



CSC 5741
Lecture 1: Administrivia and Course Overview

Lighton Phiri <lighton.phiri@unza.zm>
Department of Library and Information Science
University of Zambia