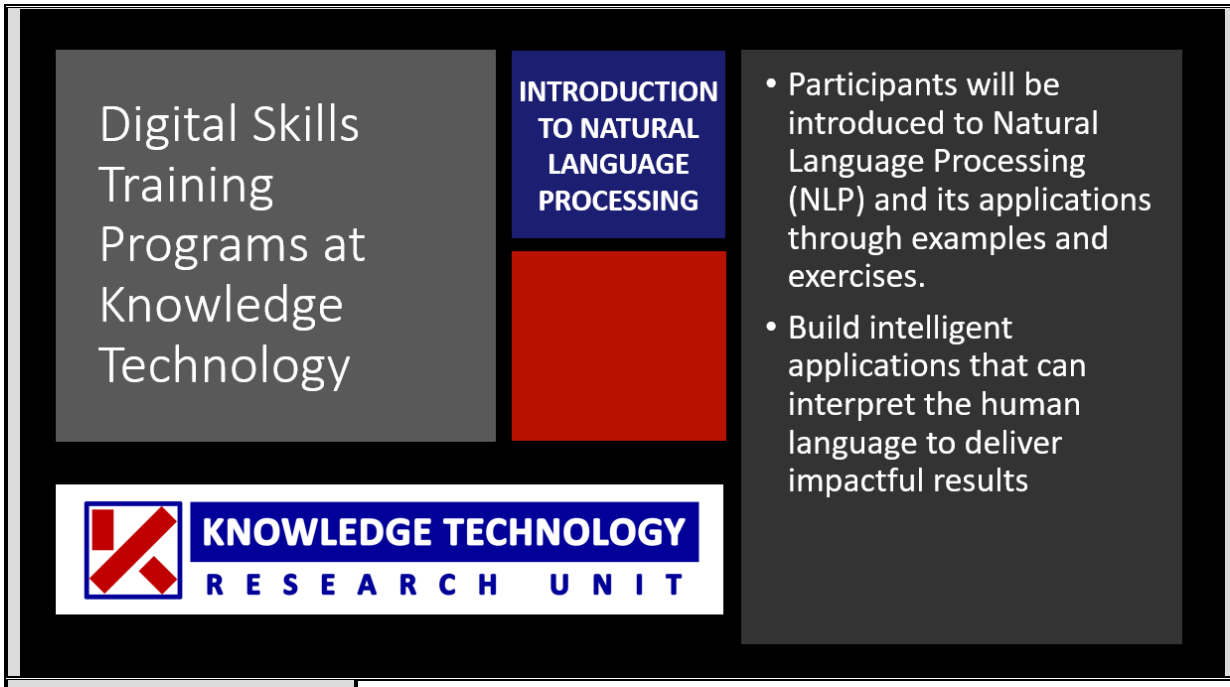


| | |
|---|---|
| **Course Title** | Introduction to Natural Language Processing |
| **Duration** | 10 Days |
| **Trainer** | Assoc. Prof. Dr. Rayner Alfred |
| **Cost** | Email ralfred121@gmail.com or call 013-881-9966 for quotations |
| **Max Participants** | 25 |

## SYNOPSIS

Participants will be introduced to Natural Language Processing (NLP) and its applications through examples and exercises. This will be followed by an introduction to the initial stages of solving a problem, which includes problem definition, getting text data, and preparing text data for modeling. With exposure to concepts such as advanced Natural Language Processing algorithms and visualization techniques, participants will learn how to create applications that can extract information from unstructured data and present it as impactful visuals. Although participants will continue to learn NLP-based techniques, the focus will gradually shift to developing useful applications. In those sections, participants will gain an understanding of how to apply Natural Language Processing techniques to answer questions, as can be used for chatbots.

## LEARNING OUTCOMES

What participants will gain at the end of the course

- Obtain, verify, and clean data before transforming it into a correct format for use
- Perform data analysis and machine learning tasks using Python
- Gain an understanding of the basics of computational linguistics
- Build models for general NLP tasks
- Evaluate the performance of a model with the right metrics
- Visualize, quantify, and perform exploratory analysis from any text data

## JUSTIFICATION TO LEARN NATURAL LANGUAGE PROCESSING

In the past years, the tech world has seen a surge of Natural Language Processing (NLP) applications in various areas, including adtech, publishing, customer service and market

intelligence. According to Gartner's hype cycle, NLP has reached the peak of inflated expectations in 2018. Many businesses see it as a "go-to" solution to generate value from the 80% of business-relevant data that comes in unstructured form. To put it simply—NLP is wildly adopted with wildly variable success, it is a unique NLP component critical for the core business of our company and most important is that there is a large demand to have an internal competence to develop IP-relevant NLP technology.

This Natural Language Processing course is designed for novice and mid-level data scientists and machine learning developers who want to gather and analyze text data to build an NLP-powered product. It'll help participants to have prior experience of coding in Python using data types, writing functions, and importing libraries. Some experience with linguistics and probability is useful but not necessary.

## TOPICS LIST

[1] Introduction to Natural Language Processing
[2] Basic Feature Extraction Methods
[3] Developing a Text classifier
[4] Collecting Text Data from the Web
[5] Topic Modelling
[6] Text Summarization and Text Generation
[7] Vector Representation
[8] Sentiment Analysis

## COURSE SYLLABUS (10 DAYS)

| DAY | TOPICS COVERED | TIME |
|---|---|---|
| One | **MODULE 1: INTRODUCTION TO NATURAL LANGUAGE PROCESSING**<br>➢ History of NLP, Text Analytics and NLP<br>➢ Various Steps in NLP<br>➢ Tokenization, Part of Speech Tagging, Stop Word Removal, Text Normalization, Spelling Correction, Stemming, Lemmatization, Named-Entity Recognition (NER), Word Sense Disambiguation, Sentence Boundary Detection<br>➢ Kick Starting an NLP Project | 8am – 5pm |
| Two | **MODULE 2: BASIC FEATURE EXTRACTION METHODS**<br>➢ Types of Data<br>➢ Categorizing Data Based on Structure<br>➢ Categorization of Data Based on Content<br>➢ Cleaning Text Data<br>➢ Tokenization and Types of Tokenizers, Issues with Tokenization<br>➢ Stemming, RegexpStemmer, The Porter Stemmer<br>➢ Lemmatization<br>➢ Language Translation<br>➢ Stop-Word Removal<br>➢ Feature Extraction from Texts, Extracting General Features from Raw Text<br>➢ Bag of Words, Zipf's Law<br>➢ TF-IDF, Feature Engineering<br>➢ Word Clouds and Other Visualizations | 8am – 5pm |
| Three and Four | **MODULE 3: DEVELOPING A TEXT CLASSIFIER**<br>➢ Machine Learning<br>➢ Unsupervised Learning | 8am – 5pm |

| | | |
|---|---|---|
| | ➢ Hierarchical Clustering<br>➢ K-Means Clustering<br>➢ Supervised Learning<br>➢ Classification<br>➢ Logistic Regression<br>➢ Naive Bayes Classifiers<br>➢ K-Nearest Neighbors<br>➢ Regression<br>➢ Linear Regression<br>➢ Tree Methods<br>➢ Random Forest<br>➢ GBM and XGBoost<br>➢ Sampling<br>➢ Developing a Text Classifier<br>➢ Feature Extraction<br>➢ Feature Engineering<br>➢ Removing Correlated Features<br>➢ Dimensionality Reduction<br>➢ Deciding on a Model Type<br>➢ Evaluating the Performance of a Model<br>➢ Building Pipelines for NLP Projects<br>➢ Saving and Loading Models | |
| Five | **MODULE 4: COLLECTING TEXT DATA FROM THE WEB**<br>➢ Collecting Data by Scraping Web Pages<br>➢ Requesting Content from Web Pages<br>➢ Dealing with Semi-Structured Data<br>➢ JSON<br>➢ XML<br>➢ Using APIs to Retrieve Real-Time Data<br>➢ API Creation<br>➢ Extracting Data from Local Files | 8am – 5pm |
| Six | **MODULE 5: TOPIC MODELING**<br>➢ Topic Discovery<br>➢ Discovering Themes<br>➢ Exploratory Data Analysis<br>➢ Document Clustering<br>➢ Dimensionality Reduction<br>➢ Historical Analysis<br>➢ Bag of Words<br>➢ Topic Modeling Algorithms<br>➢ Latent Semantic Analysis<br>➢ LSA – How It Works<br>➢ Latent Dirichlet Allocation<br>➢ LDA – How It Works<br>➢ Topic Fingerprinting | 8am – 5pm |
| Seven | **MODULE 6: TEXT SUMMARIZATION AND TEXT GENERATION**<br>➢ What is Automated Text Summarization?<br>➢ Benefits of Automated Text Summarization<br>➢ High-Level View of Text Summarization<br>➢ Extractive Text Summarization<br>➢ Abstractive Text Summarization | 8am – 5pm |

| | | |
|---|---|---|
| | ➢ Sequence to Sequence<br>➢ Encoder Decoder<br>➢ TextRank<br>➢ Summarizing Text Using Gensim<br>➢ Summarizing Text Using Word Frequency<br>➢ Generating Text with Markov Chains<br>➢ Markov Chains | |
| Eight | **MODULE 7: VECTOR REPRESENTATION**<br>➢ Vector Definition<br>➢ Why Vector Representations?<br>➢ Encoding<br>➢ Character-Level Encoding<br>➢ Positional Character-Level Encoding<br>➢ One-Hot Encoding<br>➢ Key Steps in One-Hot Encoding<br>➢ Word-Level One-Hot Encoding<br>➢ Word Embeddings<br>➢ Word2Vec<br>➢ Using Pre-Trained Word Vectors<br>➢ Document Vectors<br>➢ Uses of Document Vectors | 8am – 5pm |
| Nine and Ten | **MODULE 8: SENTIMENT ANALYSIS**<br>➢ Why is Sentiment Analysis Required?<br>➢ Growth of Sentiment Analysis 0<br>➢ Monetization of Emotion<br>➢ Types of Sentiments<br>➢ Key Ideas and Terms<br>➢ Applications of Sentiment Analysis<br>➢ Tools Used for Sentiment Analysis<br>➢ NLP Services from Major Cloud Providers<br>➢ Online Marketplaces<br>➢ Python NLP Libraries<br>➢ Deep Learning Libraries<br>➢ TextBlob<br>➢ Understanding Data for Sentiment Analysis<br>➢ Training Sentiment Models | 8am – 5pm |

## TRAINER'S BIOGRAPHIES

# RAYNER ALFRED

## ASSOCIATE PROFESSOR OF COMPUTER SCIENCE

Certified IBM DB2 Academic Associate, Certified Tester Foundation Level (CTFL)

**AREAS OF SPECILIZATION:** Advanced Machine Intelligence, Data Analytics, Data Mining, Information Retrieval, Artificial Intelligence, Machine Learning, Knowledge Discovery

**ADDRESS:** Knowledge Technology Research Group, Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah.

**CONTACT:** Mobile: 6013-881-9966, eMail: ralfred@ums.edu.my

Rayner Alfred is an Associate Professor of Computer Science at the Faculty of Computing and Informatics, Universiti Malaysia Sabah in Malaysia that focuses on Data Science and Software Engineering programmes. He leads and defines projects around knowledge discovery, information retrieval and machine learning that focuses on building smarter mechanism that enables knowledge discovery in structured and unstructured data. His work addresses the challenges related to big data problem: How can we create and apply smarter collaborative knowledge discovery and machine learning technologies that bridge the structured and unstructured data mining and cope with the big data problem.

Rayner completed his PhD in 2008 looking at intelligent techniques using machine learning to model and optimize the dynamic and distributed processes of knowledge discovery for structured and unstructured data. He holds a PhD degree in Computer Science from York University (United Kingdom), a master's degree in computer science from Western Michigan University, Kalamazoo (USA) and a Computer Science degree from Polytechnic University of Brooklyn, New York (USA) where he was the recipient of the *Myron M. Rosenthal Academic Achievement Award* for the outstanding academic achievement in Computer Science in 1994. He has authored and co-authored more than 100 journals/book chapters and conference papers, editorials, and served on the program and organizing committees of numerous national and international conferences and workshops.

Rayner is currently a member of IEEE, a Certified Software Tester (CTFL) from the International Software Testing Qualifications Board (*ISTQB*), and a certified IBM DB2 Academic Associate (IBM DB2 AA). He leads the Advanced Machine Intelligence (AMI) research group in UMS and he has led several projects related to knowledge discovery and machine learning on Big Data. Rayner is also the recipient of the Research Fellow at Japan Advanced Institute of Science and Technology (JAIST), Japan. He is also the recipient of multiple GOLD awards at national and international research exhibitions in Data Mining and Machine Learning based solutions (Face Recognition and Knowledge Discovery), that include International Trade Fair Ideas in Nuremberg, Germany (iNEA2018) International Invention Innovation Competition in Toronto, Canada (iCAN 2018), Seoul International Invention Exhibition in Seoul, Korea (SIIF 2010). He has secured RM6,931.433.00 worth of project grants. Some of his project researches include biometric authentication using face recognition, building security based on plate number recognition using deep learning, sentiment analysis for Malay and English in measuring public opinion, news-news correlation trending, machine learning algorithm-based solution for predicting diseases in health care, smart monitoring using an ensemble based face recognition system and smart information management and retrieval to name a few. Some of the completed projects include Semantic Multi-Agent For Knowledge Sharing, developing an Evolutionary-Based Ensemble Classifier Framework for Learning Big Relational Data, developing a genetic-based hierarchical agglomerative clustering technique for parallel clustering of bilingual corpora based on reduced terms, enhancing document Clustering By Integrating Semantic Background Knowledge and Syntactic

Features Into the BOW Representation and the fundamental Study on an Evolutionary Based Features Construction Methods for Data Summarization Approach to Predict Survival Factors of Coral Reefs in Malaysia, to name a few and also infrared face recognition based on ensemble approach.