REPORT ON
SUMMER SCHOOL 2022
(MAY 4TH 2022 - 17TH MAY 2022)

ON

APPLIED NLP AND UNSTRUCTURED DATA ANALYTICS

The Summer School 2022, on **"Applied NLP and Unstructured Data Analytics"**, organized by International Centre for Free and Open Source Software (ICFOSS) in collaboration with **Kerala State IT Mission (KSITM)**, was held from **May 4th to 17th 2022** at International Centre for Free and Open Source Software (ICFOSS). The school convened participants who were Post Graduate students, Research scholars, Faculty and Professionals from industry. The aim of the school was to improve their participation in FOSS communities as well as mould the individuals in the field of Natural Language Processing. The invited talks and hands-on sessions provided an opportunity for the participants to acquire individual attention of mentors and to develop network between themselves. The Summer school encouraged the participants to come up with the best and brightest ideas in various areas of natural language processing and machine learning. Summer School sessions are Project Oriented i.e. the content covers practically with concepts of theoretical knowledge . As this event is part of the Gender and Technology program, 50% of the participants were women. It is the second installment of this program. Talks and workshops went hand in hand to deliver an immersive learning experience and underwent a project at the end of the camp

Participants :

A total of 26 participants attended the school. This program is designed by ICFOSS to empower the technologists with Industrial Knowledge in the particular domain through hands on trainings and workshops.

Inaugural Function

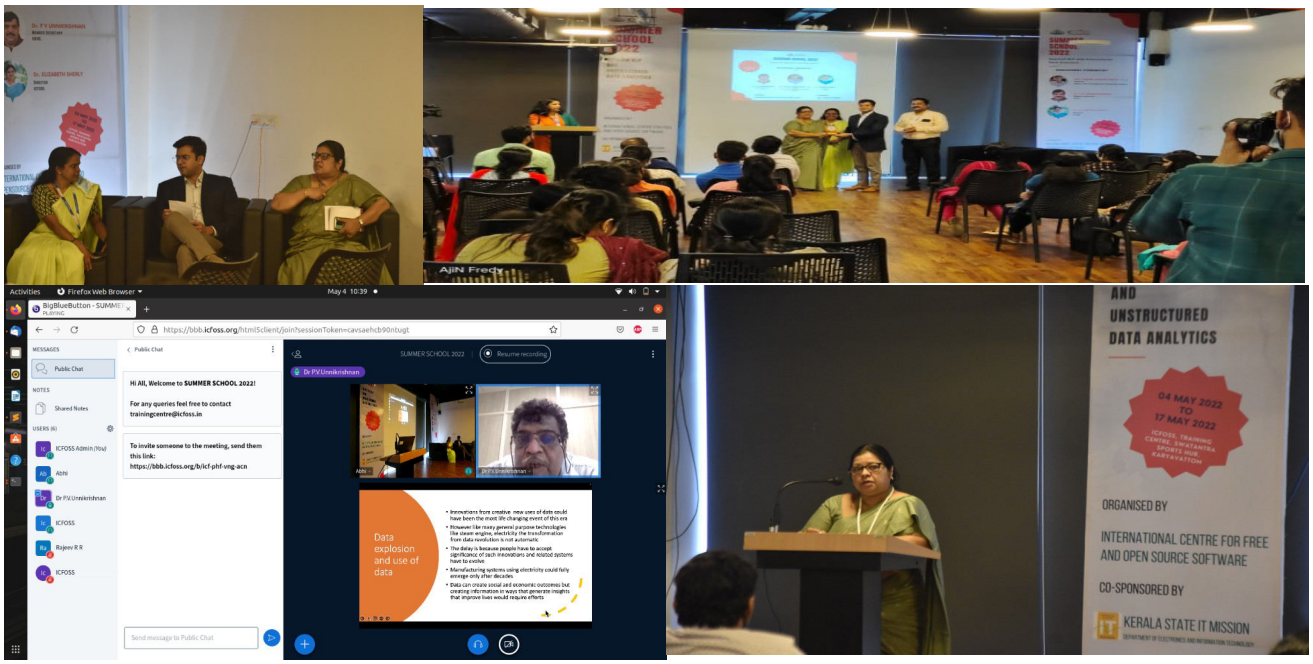
The Second edition of Summer School formally started with an inaugural function at 10.00 am on 04-05-2022. The welcome speech was delivered by Dr. Rajeev R R, Program Head (e-Governance and Development) of ICFOSS. He gave a brief introduction about ICFOSS, talked about the success of the previous edition of winter school which leads to the winter school for women in 2022. He then welcomed the honorable dignitaries present in the dais,Chief guest of the day-Mr Snehil Kumar Singh IAS-Director of Kerala State IT Mission, Keynote speaker-Dr. P V Unnikrishnan-Strategic Advisor of KDISC, Dr. Elizabeth Sherly-Director of ICFOSS , Mrs Chithra Secretary and Registrar of ICFOSS respectively.

Followed by the welcome speech, Dr. P V Unnikrishnan gave a keynote speech, he gave an insight about the concept of unstructured data analytics and its scope in current scenarios. In his speech he

pointed out that there is not only human to human communication but also a machine to human communication going on which is a deviation from the natural mode of communication. Then discussed the scope of NLP in industry and various types of data available for processing.

Dr. Elizabeth Sherly delivered a presidential speech, by pointing out the evolution of the brain computer interface and the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. She mentioned that BCI offers an extended degree of freedom either by strengthening or by substituting human peripheral working capacity and has potential applications in various fields such as rehabilitation, affective computing, robotics, gaming, and neuroscience.She also gave an overview of artificial intelligence having the ability to rationalize and take actions that have the best chance of achieving a specific goal.

Mr Snehil Kumar Singh IAS gave the inaugural speech. He praised the initiative taken by ICFOSS for conducting such a programme to bring young people to a thriving technology area. He spoke about the importance of data analytics in any sector;government organizations as well as the corporate world, and also job opportunities in various government organizations. He mentioned that data analytics offers both estimation and exploration capability for information. It allows one to understand the market or process's current state and offers a solid base for forecasting future results.



Following the inaugural speech, felicitation was given by Mrs Chitra. She talked about the innovations in unstructured data analytics that are happening now. Vote of thanks was given by Mrs.Hasiya Noohu, She thanked each and everyone in the dais and the entire ICFOSS team. The formal inaugural ceremony ended by 11.45 am.

DAY 1 - 04/05/2022
AFTERNOON SESSION:
(Time: 2pm-5pm)
TOPIC: NLP in Indian Languages
Resource Person: Dr. Rajeev RR
Profile:Programme Head(E-Governance & Development),ICFOSS

He started the session at 2.15 pm with the Introduction of participants.After that he moved into the topic with a really nice introduction about data.
He started off with the introduction about data. And detailed about the following topics below:

Data is the lifeblood of business, and it comes in a huge variety of formats — everything from strictly formed relational databases to your last post on Facebook. All of that data, in all different formats, can be sorted into one of two categories: structured and unstructured data.

Structured vs. unstructured data can be understood by considering the who, what, when, where, and the how of the data:

1. Who will be using the data?
2. What type of data are you collecting?
3. When does the data need to be prepared, before storage or when used
4. Where will the data be stored?
5. How will the data be stored?

These five questions highlight the fundamentals of both structured and unstructured data, and allow general users to understand how the two differ. They will also help users understand nuances like *semi*-structured data, and guide us as we navigate the future of_ data in the cloud.

What is structured data?

Structured data is data that has been predefined and formatted to a set structure before being placed in data storage, which is often referred to as schema-on-write. The best example of structured data is the relational database: the data has been formatted into precisely defined fields, such as credit card numbers or address, in order to be easily queried with SQL.

What is unstructured data?

Unstructured data is data stored in its native format and not processed until it is used, which is known as schema-on-read. It comes in a myriad of file formats, including email, social media posts, presentations, chats, IoT sensor data, and satellite imagery.

What is semi-structured data?

Semi-structured data refers to what would normally be considered unstructured data, but that also has meta data that identifies certain characteristics. The metadata contains enough information to enable the data to be more efficiently cataloged, searched, and analyzed than strictly unstructured data. Think of semi-structured data as the go-between of structured and unstructured data.

Language Technology

Language technology is a multidisciplinary field. It often comes with the label computational linguistics, natural language processing or natural language engineering. In language technology we study methods and develop tools for processing human language (speech and writing) to be recognised by computers. Well-known applications of such research include automatic spelling and grammar checking, machine translation as well as automatic speech recognition.

Natural Language Processing and its different terms

Natural Language Processing (NLP) is a subfield of artificial intelligence(AI) It helps machines process and understand the human language so that they can automatically perform repetitive tasks. Examples include machine translation, summarization, ticket classification, and spell check.

Take sentiment analysis, for example, which uses natural language processing to detect emotions in text. This classification task is one of the most popular tasks of NLP, often used by businesses to automatically detect brand sentiment on social media. Analyzing these interactions can help brands detect urgent customer issues that they need to respond to right away, or monitor overall customer satisfaction.

He then explained the pro and cons of natural language processing and its applications.Then he started explaining sentiment analysis in brief.

Sentiment Analysis

Sentiment Analysis is the process of analyzing emotions within a text and classifying them as positive, negative, or neutral. By running sentiment analysis on social media posts, product reviews, NPS surveys, and customer feedback, businesses can gain valuable insights about how customers perceive their brand.

He concluded the session with discussion on computational linguistics,morphological analyser,parsing and its analysis using various techniques and given a brief idea on malayalam language.

DAY 2 - 05/05/2022
MORNING SESSION:
TOPIC: DATA MINING/DATA ANALYTICS
Resource Person: Dr. Satheesh Kumar
Profile:Professor,University of Kerala

He started the session with a brief introduction to Data Science which is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.He mentioned that Data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.Predictive analytics is a branch under advanced analytics primarily used to make predictions about the uncertain future events. Predictive analytics makes use of statistics, modeling, data mining, artificial intelligence, and machine language to work on the current set of data provided as instructions and predict the future events.He explained the various steps involved in the Predictive Analytics Process Cycle:Regression,Classification,Association rule mining, Clustering and Text mining.Then explained the installation of Base-R.

He gave a detailed description about **built-in data sets**, which are generally used as demo data for playing with R functions.Then mentioned that R supports majorly four kinds of binary operators between a set of operands. In this article, we will see various types of **operators in R Programming language** and their usage.Matrices are the R objects in which the elements are arranged in a two-dimensional rectangular layout and its calculation.R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms,R has four in-built functions to generate normal distribution. They are described below. **dnorm(x, mean, sd) pnorm(x, mean, sd) qnorm(p, mean, sd) rnorm(n, mean, sd).** Many of the statistical approaches used to assess the role of chance in epidemiologic measurements are based on either the direct application of a probability distribution (e.g. exact methods) or on approximations to exact methods.R makes it easy to work with probability distributions.A representation of the distribution of a numeric variable that uses a kernel density estimate to show the probability density function of the variable. In R Language we use the density() function which helps to compute kernel density estimates.The sample correlation coefficient (r) is a measure of the closeness of association of the points in a scatter plot to a linear regression line based on the plotted points in the graph.Regression analysis is a widely used statistical tool to establish a relationship model between two variables.The types of regression-Linear Regression,Polynomial Regression,Ridge Regression.A linear regression is a statistical model that

analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).*Polynomial Regression* is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial.Ridge regression is a regularized regression algorithm that performs L2 regularization that adds an L2 penalty, which equals the square of the magnitude of coefficients.

AFTERNOON SESSION:
(Time:2:00pm-5:00pm)

This session started with an introduction to R.R is a language and environment for statistical computing and graphics.R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

Then he moved on to Neural Network in R. Neural Network is just like a human nervous system, which is made up of interconnected neurons, in other words, a neural network is made up of interconnected information processing units.The neural network draws from the parallel processing of information, which is the strength of this method.A neural network helps us to extract meaningful information and detect hidden patterns from complex data sets.A neural network is considered one of the most powerful techniques in the data science world.This method is developed to solve problems that are easy for humans and difficult for machines.These problems are often referred to as pattern recognition.Then,he concluded the session.

DAY 3 - 06/05/2022

TOPIC: Math & Machine Learning

Resource Person: Dr. Umesh P
Profile:Assistant Professor,College of Engineering Thalassery

The session began with an overview of applications of Linear Algebra in Machine learning.He discussed selecting the right algorithm,
choosing parameter settings & validation strategies and understanding Bias-Variance tradeoff.He then discussed how mathematics is involved in machine learning and language computing.

After that he explained with the help of a presentation an overview of Linear Algebra and Machine Learning.Linear Algebra is the branch of mathematics concerning linear equations such as their representation in vector spaces through matrices.It is the study of vectors and linear functions.He also discussed Tensor,Notations,matrix formation,determinants,Basis,Linear independence,Eigenvalues and eigenvectors,Orthonormal-Orthogonal Bases,Diagonalizing symmetric matrices.

Then he explained various vector norms-Euclidean,Manhattan,Minkowski, Chebychev,Cosine similarity,and Hamming.

Then he conducted a **Lab Session** on Python for solving various Linear algebra applications.It covered L1 Norm,Squared L2 Norm,Cosine similarity and some applications like Eigen decomposition and Singular Value Decomposition(SVD) -and Image compression via SVD.

Afternoon session started with the introduction to calculus.Then he started explaining probability and statistics.In this session he covered various aspects related to probablility like distribution functions etc..He concluded the session by explaining the concepts related to statitics like hypothesis,central tendency parameters etc..

DAY 4 – 07/05/2022

Morning Session

TOPIC: Tagging In Malayalam- Parts Of Speech

Resource Person: Dr. Rajeev RR
Profile: Programme Head(E-Governance and Development), ICFOSS

Started the Session with structure of a sentence and transfer to different languages. And detailed about the following topics below:

Eg:Good boy → നല്ല കുട്ടി

Very good boy → വളരെ നല്ല കുട്ടി

Address different issues while translating in different languages.Mention Rule-based systems are not enough to work with the exact POS .So, Machine translation is required instead.

- POS Tagging

  POS(grammatical tagging).

- Tag Set.

Consists of grammatical tags, including morphological,morpho-syntactic, semantic-level tags:

Names given to set of tags from which tags are to be given to the input words in a text.Two types,

1. Flat Tag set
2. Hierarchical tag set.

Then he explained the various part of speech and explained the POS tagging.

Parts of Speech Tagging, a grammatical tagging, is a process of marking the words in a text as corresponding to a particular part of speech, based on its definition and context. This is the first step towards understanding any languages. It finds its major application in the speech and NLP like Speech Recognition, Speech Synthesis, Information retrieval etc. A lot of work has been done relating to this in NLP field. Chunking is the task of identifying and then segmenting the text into a syntactically correlated word groups. Chunking can be viewed as shallow parsing. This text chunking can be considered as the first step towards full parsing. Mostly Chunking occur after POS tagging. This is very important for activities relating to Language processing

Afternoon Session

Topic: Machine **Learning in IR**

Resource Person: **Dr.Anu Thomas**

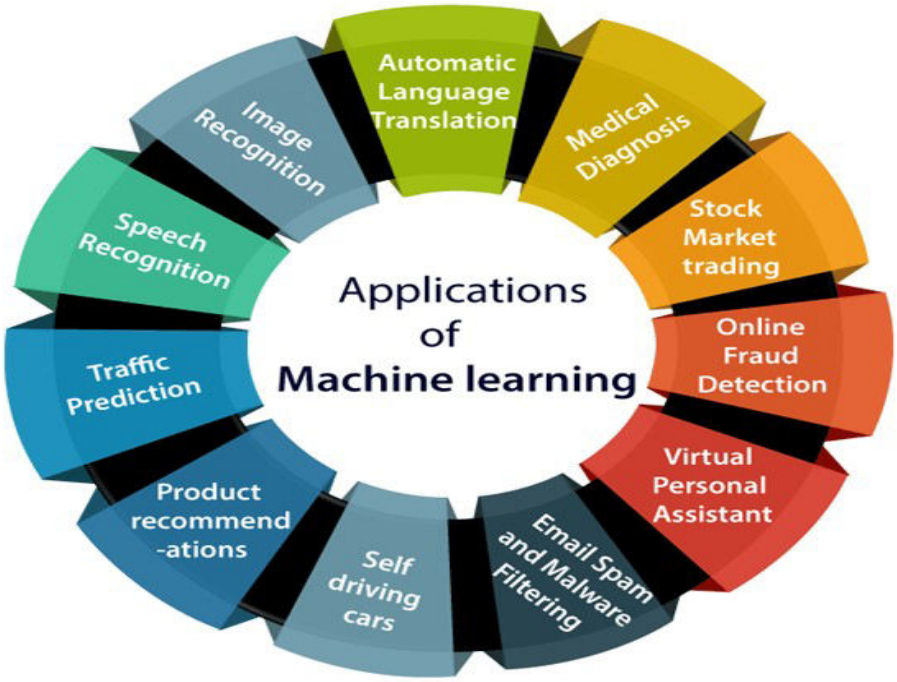
Profile: **Assistant Professor**

She started the class with basics of machine learning. she clearly explained the classification of supervised, unsupervised learning, semi-supervised learning and reinforcement learning. Supervised learning techniques like classification and regression is explained. The algorithms like logistic regression,support vector machine,K nearest neighbours,naive bayes etc.. are explained in detail.Given glimpses of deep learning also.

Deep Learning

- **Artificial Neural Networks or** ANN are the non-linear ML algorithms that work to process the brain and transfer information from one layer to another in a similar way.

Deep learning studies these neural networks, which implement newer and faster hardware for the training and development of larger networks with a huge dataset. All deep learning methods achieve great results for different challenging tasks such as machine translation, speech recognition, etc. The core of processing neural networks is based on linear algebra data structures, which are multiplied and added together. Deep learning algorithms also work with vectors, matrices, tensors (matrix with more than two dimensions) of inputs and coefficients for multiple dimensions.



Convolutional Neural Network based methods

-Is an effective feature extraction architecture,can identify the predictive n-gram vocabularies in a sentence automatically.

- Recurrent Neural Network based methods
- Able to capture long-dependence relationships.
- LSTM

- Methods of matching function learning
- Matching with word-level similarity matrix
- Attention models
- Methods of relevance learning-rank documents by relevance to a user's query.
- Relevance matching usually contains an asymmetric matching function and could contain all different forms, ranging from keywords to documents, to phrases and sentences and to documents.
- Similarity matching-paraphrase identification
- Relevance matching-ad hoc retrieval

Concluded by,

Simple Application **BERT to ad hoc IR tasks**

- Future ,more hybrid models can be built on top of BERT and other neural ranking models to produce better IR results.

Microsoft AI Challenge India 2018,Example task,describe about the data set and process.

DAY 5 -09/05/2022

Topic:Data Analysis Using Neural Networks

Resource Person:Dr.Gopakumar G

Profile:Assistant Professor,NITC

The session began with an overview of supervised learning. He mentioned that ANN comes under supervised learning. Then he briefly explained how the data is represented in the system. Two approaches for generating data were explained, namely Generative and Discriminant Approaches.

- Generative approaches estimate the discriminant function by first modeling the probability distribution of the data in each class.
- Discriminative approaches estimate the discriminant function explicitly, without modeling the data by a probability distribution.

In ANN, a discriminative approach is followed generally.
Then he proceeded to the explanation of discriminant functions.

The afternoon session started with Multilayer Neural networks. The goal is to classify objects by learning nonlinearity. There are many problems for which linear discriminants are insufficient for minimum error. There is no automatic method for determining the nonlinearities when no information is provided to the classifier. In using the multilayer Neural Networks, the form of the nonlinearity is learned from the training dataset.

Finally he gave a brief introduction to Deep Learning. In Deep Learning features are automatically learnt from the data. He introduced CNN, RNN, etc shortly. Then, he concluded the session.

DAY 6 – 10/05/2022

Topic: Visualization and streamlit deployment

Resource Person: Mr. Deepu C

Profile: Research Associate ,ICFOSS

It was a hands-on session on Streamlit. Streamlit is an open-source app framework for deploying Python projects. Streamlit is a tool to create a web-based dashboard with a focus for the ML scientist or engineer. It is an open-source framework. Specifically, Streamlit uses HTML, CSS and Javascript but does not need the developer to know all these. It has compatibility with all the major frameworks/ libraries in Python.
Then he introduced various frameworks for deployment.

- **Streamlit, Dash**, and **Panel** are full dashboarding solutions, focused on Python-based data analytics and running on the **Tornado** and **Flask** web frameworks.
- **Shiny** is a full dashboarding solution focused on data analytics with R.
- **Jupyter** is a notebook that data scientists use to analyze and manipulate data and can be used to visualize data.

- **Voila** is a library that turns individual Jupyter notebooks into interactive web pages.
- **Flask** is a Python web framework for building websites and apps – not necessarily with a data science focus.

Pros of Streamlit:

- Interactive
- Easy deployment
- Active Community

Cons of Streamlit:

- Security issues
- Customization is limited

He started with the installation of VSCode.
The following commands were used for the installation of VSCode on Ubuntu:

$ sudo apt update
$ sudo apt install software-properties-common apt-transport-https wget
$ wget -q https://packages.microsoft.com/keys/microsoft.asc -O- | sudo apt-key add -
$ sudo add-apt-repository "deb [arch=amd64] https://packages.microsoft.com/repos/vscode stable main"
$ sudo apt install code

Commands to install Streamlit:
$ pip install streamlit
$ streamlit hello
He shared with the participants the project file to be opened in VSCode and gave necessary instructions.

- Different aspects of creating a dashboard were discussed. Then he proceeded with implementation of Streamlit widgets like button, selectbox, multiselect, checkbox, text_area, slider, date_input, number_input, etc.
- Then visualization using streamlit was introduced. He shared a dataset as a csv file and demonstrated different methods for visualizing the dataset.
- Next topic was related to the deployment of machine learning models using streamlit. Streamlit allows us to deploy our machine learning project using simple python scripts. Pickle and joblib libraries are usually used for storing machine learning models. Joblib is preferably used for text data.
- He demonstrated deployment of a machine learning model in the cloud using Heroku.

Finally he explained how deep learning models are deployed using streamlit. Then he concluded the session.

AFTERNOON SESSION

Topic: Text Analysis
Resource Person: Mrs. Dhanya L K
Profile: Mar Baselios College of Engineering And Technology, Trivandrum
Time: 2.00pm - 5.00pm

She started with an introduction to Natural Language Processing. NLP refers to the branch of Computer Science concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. NLP is an intersection of Computer Science and linguistics.
How is the NLP area divided?
Speech recognition: Translation of spoken language into text

1. Natural Language Understanding: The computer's ability to understand what we say
2. Natural Language Generation: The generation of natural language by computer.

Then she mentioned some applications of NLP like information retrieval, sentiment analysis, question answering, machine translation, chatbot, spelling correction, etc.
Basics of NLP for a text:

- Sentence segmentation
- Words Tokenization
- Text Lemmatization
- Stop Words
- Dependency Parsing in NLP
- NER

  Text analytics is an AL technology that uses NLP to transform the unstructured text in documents into normalized, structured data suitable for analysis or to drive ML algorithms.

Then she proceeded with the hands-on session. She started with the installation of NLTK. NLTK was imported first. Different pre-processing steps for NLP were introduced.

Tokenization

Tokenization in NLP is the process by which a large quantity of text is divided into smaller parts called tokens. Then sentence tokenization and word tokenization were performed using NLTK..

Stop Words

Content words and stop words were explained and stop word removal using NLTK was demonstrated.
Stop words are a set of commonly used words in a language. Examples of stop words in English are “a”, “the”, “is”, “are” and etc.

Stemming

It is a process of linguistic normalization,which reduces words to their root word or chops off the derivational affixes.The most commonly used stemmer is Porter stemmer. It uses the suffix stripping algorithm. It may not always be accurate. So we need lemmatization.

Lemmatization

Lemmatization in NLTK is the algorithmic process of finding the lemma of a word depending on its meaning and context. It helps in returning the base or dictionary form of a word known as the lemma.
The NLTK Lemmatization method is based on WorldNet’s built-in morph function.
Then she explained some functions provided by WorldNet.

WordNet

WordNet is a lexical database of English. Using synsets, it helps find conceptual relationships between words such as hypernyms, hyponyms, synonyms, antonyms etc.

Part of Speech Tagging

Part-of-speech (POS) tagging is a popular Natural Language Processing process which refers to categorizing words in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context.
She mentioned different tag-sets used like Penn treebank tagset, BIS tagset, etc. Tagsets can be either hierarchical or flat.

Named Entity Recognition

Named entity recognition (NER)is probably the first step towards information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Cohen’s Kappa coefficient
Then she explained about Cohen’s Kappa coefficient. Cohen's kappa coefficient (κ) is a statistic to measure the reliability between annotators for qualitative (categorical) items. It is a metric often used to assess the agreement between two raters.
She briefly mentioned Glove, Fast text, word2vec, which are word-embedding models.
Then she introduced transfer learning, where we reuse a pre-trained model as the starting point for a model on a new task.
She mentioned some resources like nltk.corpus, TDIL, INLTK, etc.

SMOTE technique
SMOTE techniques used for unbalanced datasets were also mentioned. SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.
Logistic regression, KNN were mentioned briefly. Finally, she concluded the session.

DAY 7 - 11/05/2022
Morning Session
Topic:Social Media Analysis
Resource Person: Dr.Prem Sankar
Profile: Assistant Professor

He started the session with a brief description about data

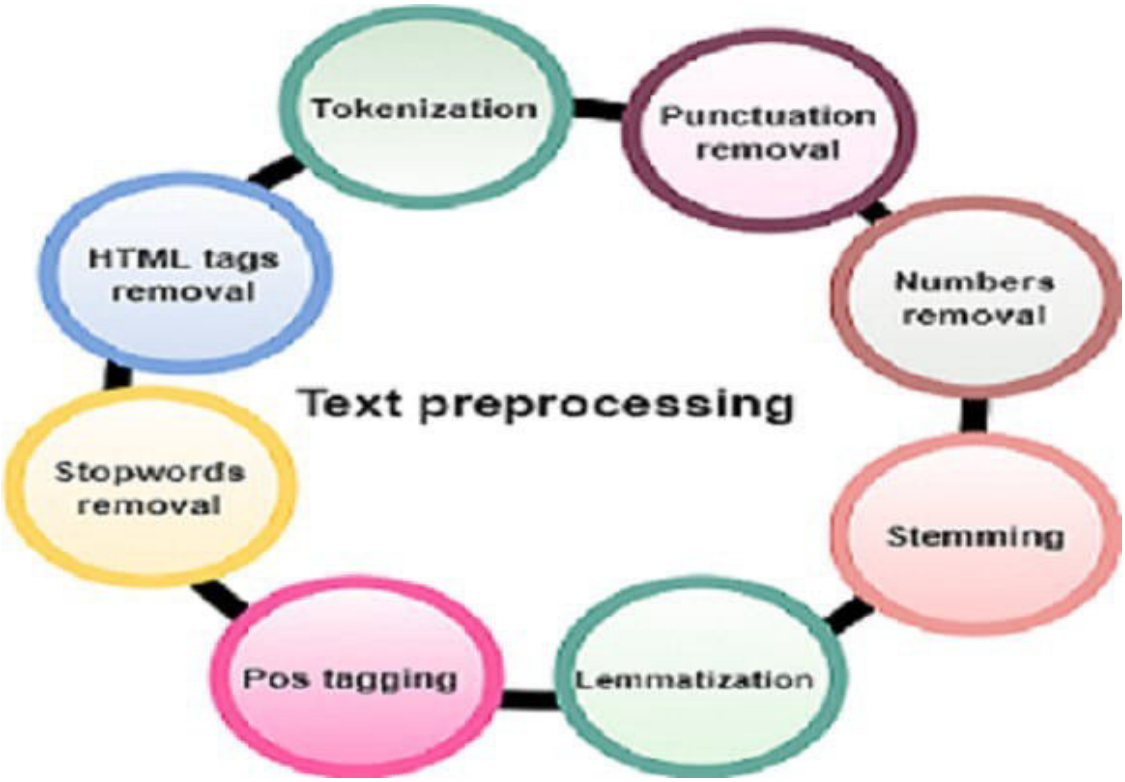
The data can be both quantitative and qualitative in nature. Quantitative data is in numeric form , which can be discrete that includes finite numerical values or continuous which also takes fractional values apart from finite values. For instance, the number of girls in a class can only take finite values, so it is a discrete variable, while the cost of a product is a continuous variable.

Qualitative data is not-numerical which can be based on methods such as interviews, grades given in an exam etc. It can be nominal and ordinal, where nominal data does not contains any order such as the

gender, marital status, while ordinal data has a particular order such as ratings of a movie, sizes of a shirt.

Explained the importance of social media analysis
What is Social Media?
To speak formally, as per the Cambridge Dictionary, social media is defined as “websites and computer programs that allow people to communicate and share information on the internet using a computer or mobile phone”. Now if you split the words social media, then, social which comes from the Latin word ‘socius’ means friend, and media refer to means of mass communication. Thus, social media in simple terms can be referred to as an informal platform for mass communication.

Basic text preprocessing timeline



A natural language processing system for textual data reads, processes, analyzes, and interprets text. As a first step, the system preprocesses the text into a more structured format using several different stages. The output from one stage becomes an input for the next—hence the name “preprocessing pipeline.”

An NLP pipeline for document classification might include steps such as sentence segmentation, word tokenization, lowercasing, stemming or lemmatization, stop word removal, and spelling correction. Some or all of these commonly used text preprocessing stages are used in typical NLP systems, although the order can vary depending on the application.

Segmentation

Segmentation involves breaking up text into corresponding sentences. While this may seem like a trivial task, it has a few challenges. For example, in the English language, a period normally indicates the end of a sentence, but many abbreviations, including “Inc.,” “Calif.,” “Mr.,” and “Ms.,” and all fractional numbers contain periods and introduce uncertainty unless the end-of-sentence rules accommodate those exceptions.

Stemming

The term word stem is borrowed from linguistics and used to refer to the base or root form of a word. For example, learn is a base word for its variants such as learn, learns, learning, and learned.

Stemming is the process of converting all words to their base form, or stem. Normally, a lookup table is used to find the word and its corresponding stem. Many search engines apply stemming for retrieving documents that match user queries. Stemming is also used at the preprocessing stage for applications such as emotion identification and text classification.

Lemmatization

Lemmatization is a more advanced form of stemming and involves converting all words to their corresponding root form, called “lemma.” While stemming reduces all words to their stem via a lookup table, it does not employ any knowledge of the parts of speech or the context of the word. This means stemming can’t distinguish which meaning of the word right is intended in the sentences “Please turn right at the next light” and “She is always right.”

The stemmer would stem right to right in both sentences; the lemmatizer would treat right differently based upon its usage in the two phrases.

A lemmatizer also converts different word forms or inflections to a standard form. For example, it would convert less to little, wrote to write, slept to sleep, etc.

He discussed about a journal paper :

Generating and visualizing topic hierarchies from microblogs:An iterative latent dirichlet allocation approach

He also familiarized about the open source data mining tool named Gephi.He demonstrated web scraping using Beautiful Soup.Also he gave us information about various API’s where we can extract data from social media.

He familiarized about some tools that are used for social network analysis

- PAJEK
- STOCNET
- NODE XL
- GEPHY
- SOCIO VIZ

AFTERNOON SESSION

Topic: Introduction to Neural Machine Translation

Resource Person: Dr.Asif Ekbal.

Profile: Associate Professor, IIT Patna

He started the session with history of Machine Translation. Machine translation is the automated translation of a source-language text into a target-language text. Human translators may be involved at pre-editing or post-editing **s**tages, i.e. at the beginning or the end, but they are not typically involved in the translation process.
Although concepts of machine translation can be traced back to the seventeenth century, it was in the 1950s when US-government-funded research stimulated international interest in the investigation and production of machine translation systems.
The original intention was to produce a fully automatic high quality machine translation system (FAHQMT) but by 1952 it was "already clear that objectives of fully automated systems were unrealistic and that human intervention would be essential" (Hutchins, 2006, p. 376). Many researchers were scientists rather than linguists and unconscious of the need for real world knowledge in the translation process. Many complex elements of language could not be easily programmed into a computer, e.g. understanding homonyms or metaphors.
Then he moved on the basics of machine translation.

Machine translation (MT) is the task to translate a text from a source language to its counterpart in a target language. There are many challenging aspects of MT: 1) the large variety of languages, alphabets and grammars; 2) the task to translate a sequence (a sentence for example) to a sequence is harder for a computer than working with numbers only; 3) there is no *one* correct answer (e.g.: translating from a language without gender-dependent pronouns, *he* and *she* can be the same).

Machine translation is a relatively old task. From the 1970s, there were projects to achieve automatic translation. Over the years, three major approaches emerged:
Rule-based Machine Translation (RBMT): 1970s-1990s

- Statistical Machine Translation (SMT): 1990s-2010s
- Neural Machine Translation (NMT): 2014-

He pointed out the Booth and weaver discussions about MT at Newyork. And also discussed about ALPAC report headed by John R Pierce of Bell Labs .Its conclusions and consequences.

He discussed MULTILINGUALITY in Indian situation.The challenges in Machine Translation has been clearly discussed.He clearly explained the relevance and importance of machine translation in

various domains such as Government,Software and Technology ,Health care,Social etc..The approches used in Machine Translation is clearly defined.
There are four different types of machine translation–Statistical Machine Translation (SMT), Rule-based Machine Translation (RBMT), Hybrid Machine Translation (HMT), and Neural Machine Translation (NMT). Here's an overview of each type:

Rule-Based Machine Translation (RBMT)

RBMT— the earliest form of MT— translates content based on grammatical rules. There have been significant advances in machine translation technology since RBMT was developed, so it has a few disadvantages. These drawbacks include the need for large amounts of human post-editing and adding languages manually. Despite this low translation quality, RBMT is useful in basic situations where a quick understanding of meaning is all that is required.

Statistical Machine Translation (SMT)

SMT works by building a statistical model of the relationships between text words, phrases, and sentences. It then applies this translation model to a second language and converts the same elements to the new language. SMT improves somewhat on RBMT but still shares many of the same problems.

Hybrid Machine Translation (HMT)

HMT is a blend of RBMT and SMT. HMT leverages a translation memory, making it far more effective in terms of quality. However, even HMT has its share of drawbacks, the greatest of which is the need for **human editing**.

Neural Machine Translation (NMT)

NMT employs artificial intelligence to learn languages and improve that knowledge constantly. In this way, it strives to mimic the neural networks in the human brain. NMT is more accurate than other types of AI translation. With NMT, it's easier to add languages and translate content. Because NMT provides better translations, it is rapidly becoming the standard in MT tool development.
NMT works by incorporating training data. Depending on the user's needs, the data can be generic or custom.

- Generic Data: This is the total of all the data learned from translations performed over time by the machine translation engine (MTE). This data produces a generalized translation tool for various applications, including text, voice, and documents.
- Custom or Specialized Data: This is training data fed to a machine translation engine to build specialization in a subject matter. Subjects include engineering, design, programming, or any discipline with its own specialized glossaries and dictionaries.

DAY 8 – 12/05/2022

Topic: Introduction to ML/Neural Networks in Keras

Resource Person: Dr.Sunil TT

Profile: Principal ,College of Engineering Attingal.

Time: **9:00am-12:00pm**

He started with the evolution of the human race and went through knowledge sharing and modern day technologies. Method of learning by children, how they grasp things around them. Various works on going on the areas of the human sensory system mainly in the areas of smell, touch etc,. Features, how does it help the system to make assumptions.

Feature space, to map features to n dimensional feature space.

Explain about data set and show some examples of dataset
Started with the evolution of the human race and went through knowledge sharing and modern day technologies. Method of learning by children, how they grasp things around them. Various works on going on the areas of the human sensory system mainly in the areas of smell, touch etc,. Features, how does it help the system to make assumptions.

Feature space, to map features to n dimensional feature space.

Explain about data set and show some examples of dataset
Several Sources are available for Machine Learning.

- Public dataset for ML.
- UC Irvine ML Repository
- Middlebury CV dataset
  It was then followed by a practical session.

Practical Session:

OVERVIEW

Imported packages are,

- Keras, sample dataset mnist,Dense from layers, Sequential from models
- (X_train, y_train)- used to train the model.
- (X_valid, y_valid)- for validation.Shape of validation data is 10000.
- mnist.load_data()-- load the dataset
- X_train.shape-(60000, 28, 28) as output.

Before train the model needs to do **preprocess the data.**

- Unpacked the matrix to vector(single dimensional)-**Reshape**
- **Normalize** the data value ranges in between 0 and 1.
- convert the labels to **one hot representation.**

- Building a model in **keras.**
- Define **neural network architecture**

- There are two ways to build **Keras models: sequential and functional.**

Sequential is a kind of layer- by layer architecture.

Functional API allows you to create models that have a lot more flexibility as you can easily define models where layers connect to more than just the previous and next layers.

The dense layer is a neural network layer that is connected deeply, which means each neuron in the dense layer receives input from all neurons of its previous layer.

There are 64 (hyper parameter)neurons in the hidden layer

Each neuron has **sigmoid activation function**. There are **784 inputs.**

activation='softmax'- categorical data, predicting

Loss-Function=The loss function is the function that computes the distance between the current output of the algorithm and the expected output . It's a method to evaluate how your algorithm models the data.

Sparse Matrix: Most of the entries are 0.

MSE(Mean-Squared Error)

Concluded with the **model evaluation** and **accuracy** of the mode and **visualization**

DAY 8 (12/05/22):

AFTERNOON SESSION

Topic: Optical Character Recognition(OCR)
Resource Person: **Aswathy. P**
Profile: **Assistant Professor (CUSAT)**
Time: **2:00pm-05:00pm**

The session began with **Smt.Aswathy** giving an explanation of the concept of Optical Character Recognition.

Then she discussed the need and importance of using Optical Character Recognition.

- A cost and time effective solution
- Saved time,decreased errors and minimized effort
- Searchability
- Editability
- Back-ups
- Accessibility
- Storability

- Transability

After that she discussed the different scenarios where the OCR is used.

1. passport recognition for airports

2. Traffic sign recognition for airports

3. Extracting contact information

4. Converting handwritten notes to machine readable text

5. Aids for the blind

6. Data entry for business docs

7.Autonomous driving-Driverless car where OCR can be used for reading sign boards

Then she discussed about Computer Vision Models

Then she discussed the

1.Methods before the Deep Learning Era:

- Connected Components Analysis (CCA)

2.Methods inspired by Object detection

- Based on CTC,like CRNN
- Based on Spatial transformer Networks
- Based on encoder-decoder network
- Based on attention network

With that she ended the session with a brief discussion at 5.00 pm.

DAY 9 (13/05/22):
MORNING SESSION
Topic: NEURAL MACHINE TRANSLATION
Resource Person: Mr.Santanu Pal
Profile: Lead Scientist,Wipro AI LAb
Time: 10:00am-01.00pm

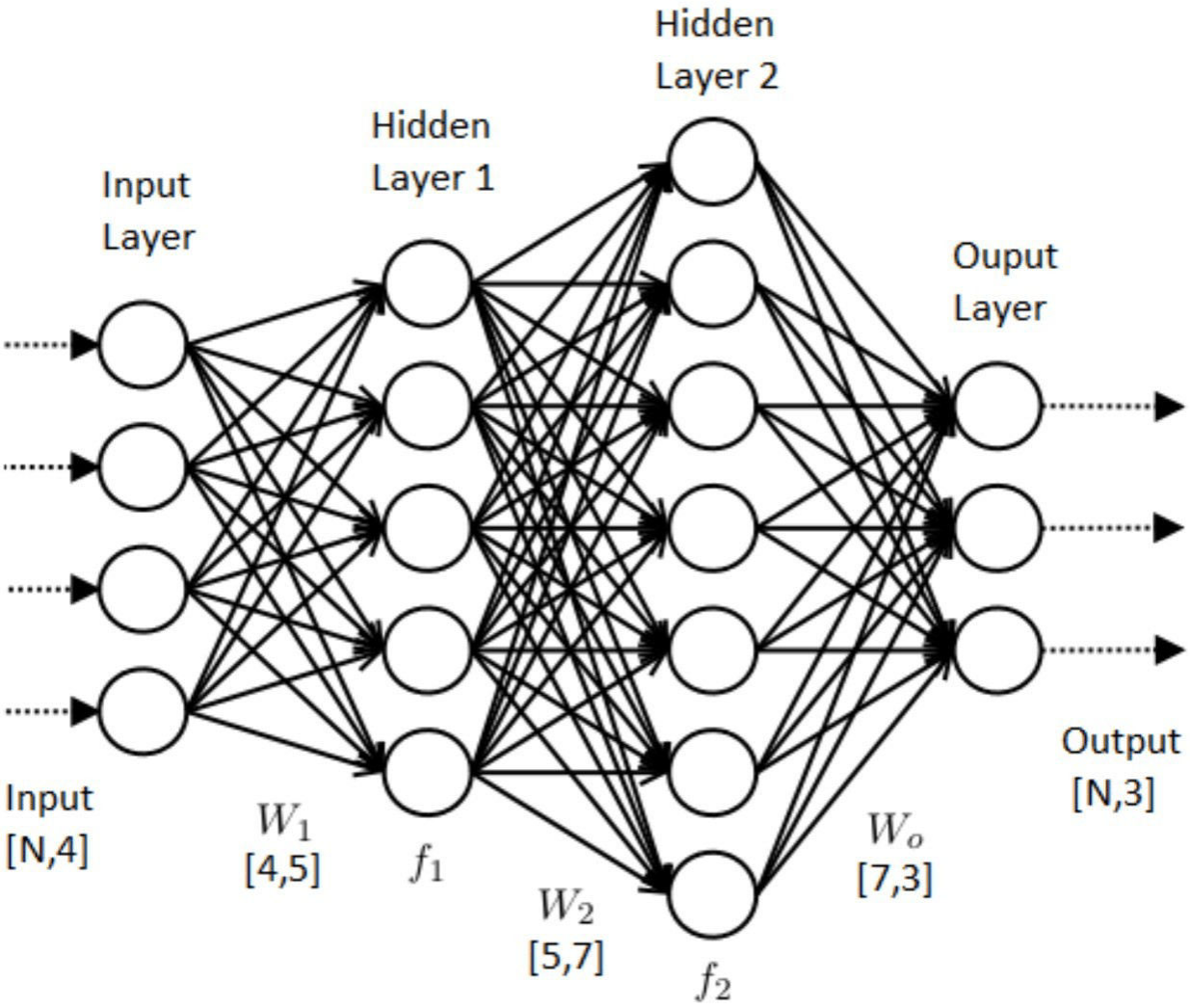
The Session started at 10 am.He started giving an introduction about Neural Machine Translation.
Introduction about embeddings
Introduction:Basic Neural Network

Key Terms

- **Neuron**: A building block of ANN. It is responsible for accepting input data, performing calculations, and producing output.
- **Input data**: Information or data provided to the neurons.
- **Artificial Neural Network(ANN)**: A computational system inspired by the way biological neural networks in the human brain process information.
- **Deep Neural Network**: An ANN with many layers placed between the input layer and the output layer.
- **Weights**: The strength of the connection between two neurons. Weights determine what impact the input will have on the output.
- **Bias**: An additional parameter used along with the sum of the product of weights and inputs to produce an output.
- **Activation Function**: Determines the output of a neural network.

Overview of Neural Networks

For example, the image dimensions might be 20 X 20 pixels that make 400 pixels. Those 400 pixels would make the first layer of our neural network.



A neural network learns from structured data and exhibits the output. Learning taking place within neural networks can be in three different categories:

- Supervised Learning - with the help of labeled data, inputs, and outputs are fed to the algorithms. They then predict the desired result after being trained on how to interpret data.
- Unsupervised Learning - ANN learns with no human intervention. There is no labeled data, and output is determined according to patterns identified within the output data.
- Reinforcement Learning - the network learns depending on the feedback you give it.

Then she discussed in detail about **How Neural Networks work**

Implementation

When talking about the implementation of the attention mechanism in the neural network, we can perform it in various ways. One of the ways can be found in the article. Where we can see how the attention mechanism can be applied into a Bi-directional LSTM neural network with a comparison between the accuracies of models where one model is simply bidirectional LSTM and other model is

bidirectional LSTM with attention mechanism and the mechanism is introduced to the network is defined by a function.

Calculating Attention, Hard Attention. Soft Attention, Disadvantages of Hard Attention, Global attention, Local Attention, Self Attention, Transfomer Model, Multi Head Attention, Positional encoding, Residual Encoding. Class ended with a question and answer session at 12.45 pm.

DAY 9 (13/05/22):

AFTERNOON SESSION
Topic: SPEECH TO SPEECH MACHINE TRANSLATION
Resource Person: Dr.Dipti Misra Sharma
Profile: Professor & Head,IIIT Hyderabad
Time: 2.00pm-5.00pm

She started the session explaining the languages speaking in India and its limitations. Then he discussed in detailed about Introduction about Speech to Speech translation, Direct speech to speech translation., Explained the basic process of translation both manually and with machine, Speech to Speech MT(Indian Context), Current Readiness, ASR, MT, TTS. Then briefed about Current Status of Technology in machine translation, Discussed about Punctuation Marks, Disfluencies, Post Processing ASR output for MT input, Domain Specific Term Translation, NLP, Prosody, Role of prosody in SSMT, Pauses, Stress, Equation(Spoken)

Translation involves

- Text translation
- Referensibility
- Cohesiveness
- Naturalness

Then detailed about Approaches to MT, Explained different types of MT, Statistical MAchine Translation, Advancement with NMT, NMT - Advantages and Disadvantages, SSMT, Hybrid Machine Translation, Sampark +NMT worked out pipeline, Machine Translation- Efforts on Sampark Systems, Corpus used for training, Morphe + BPE Segmentation, Machine Translation - Domain Adaptation, Models developed in IIIT hyderabad, Showed various translation systems live, Showed recent works developed - Swayam video lecture translation, Described the overall process, Explained the results:Tempoal Effects

: Technical Effort
: Cognitive Effort

Her Student Vandan explained about the recent researched going on at IIIT hyderabad.

DAY 10 (14-05-2022)-Morning Session
Topic: Convolutional Neural Networks/ Generative Adversarial Networks
Resource Person: Dr Sunil TT
Profile: Principal College of Engineering Attingal
Time: 9:00 am-12.00pm

Started with deep neural networks in keras library, and continued with hand on session, mainly with keras library, setting up the neural networks and training the dataset and gave a detailed lecture about accuracy of neural network models. And gave a detailed description of choosing epochs and batches for training.

He talked about normalized and unnormalized data, how they influence the training and ultimately the performance of the neural network. Then talked about neural network architecture, like how to choose input and output layers and what should be the criterion for choosing hidden layers. After that he delved into various activation functions how they influence

Given an overview of **Transfer_learning tutorial**
The reuse of a previously learned model on a new problem is known as transfer learning. It's particularly popular in deep learning right now since it can train deep neural networks with a small amount of data. This is particularly valuable in the field of data science, as most real-world situations do not require millions of labeled data points to train complicated models.

DAY 10 (14-05-2022)-Afternoon Session
Resource person's Profile: Sanjay K.Dwivedi
Prof & Head ,Dept.of Computer Science Dean ,
School of Information Science & Technology
BBA central university ,Lucknow.
TOPIC: Handling Ambiguity Issues In CLIR
Time : 2.00pm-4.30pm

The session started with what is ambiguity and handling the ambiguity issues in CLIR.

Information retrieval

An information retrieval (IR) system is a set of algorithms that facilitate the relevance of displayed documents to searched queries. In simple words, it works to sort and rank documents based on the queries of a user. There is uniformity with respect to the query and text in the document to enable document accessibility.

- **Ambiguity in IR**
  Word ambiguity is not something that we encounter in everyday life, except perhaps in the context of jokes.

Somehow, when an ambiguous word is spoken in a sentence, we are able to select the correct sense of that word without considering alternative senses. However, in any application where a computer has to process natural language, ambiguity is a problem.

Discussed about Data Warehouse

A data warehouse is **a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics**. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data.

Case study

1. Query expansion with or without ranking
2. Highest and lowest frequency terms
3. Adding terms at appropriate locations
4. (FIRE,Snippets and nearest neighbor).
5. The UAM corpus tool is used to find the frequency of terms.

Rank of retrieved documents by Okapi BM25

Documents obtained after Google searching against each query are ranked using Okapi BM25

Okapi BM25 (BM is an abbreviation of best matching) is **a ranking function used by search engines to estimate the relevance of documents to a given search query**.

Explain the below documents using okapi BM25

Query	Okapi BM25 Value & Rank									
	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
1	−2.71 Rank9	−2.708 Ran8	−3.276 Rank10	−2.564 Rank6	−2.712 Rank7	−2.169 Rank3	−1.534 Rank1	−2.268 Rank4	−1.552 Rank2	−2.39 Rank5
2	0.637 Rank1	0.582 Rank2	0.564 Rank3	0.499 Rank4	Query Terms Absent	Query Terms Absent	Query Terms Absent	Query Terms Absent	Query Terms Absent	Query Terms Absent
3	−1.966 Rank7	−1.718 Rank3	−1.006 Rank2	Query Terms Absent	−1.939 Rank5	−1.869 Rank4	−1.965 Rank6	−1.990 Rank9	−1.834 Rank1	−1.987 Rank8
4	−7.701 Rank9	−6.887 Rank6	−7.14 Rank8	−6.559 Rank5	−7.041 Rank7	−9.543 Rank10	−6.025 Rank1	−6.410 Rank2	−6.546 Rank4	−6.452 Rank3
5	−4.254 Rank6	−4.362 Rank9	−4.306 Rank8	−4.481 Rank10	−4.175 Rank4	−4.179 Rank5	−3.924 Rank2	−4.293 Rank7	−4.409 Rank3	−2.633 Rank1
6	−3.331 Rank8	−121.352 Rank10	−2.826 Rank6	−8.89 Rank9	−2.654 Rank4	−0.433 Rank2	0.55 Rank1	−2.617 Rank3	−2.849 Rank7	−2.666 Rank5
7	−2.93 Rank3	−3.844 Rank6	−4.053 Rank8	−4.141 Rank10	−3.055 Rank4	−4.075 Rank9	−2.665 Rank2	−2.3 Rank1	−4.05 Rank7	−3.797 Rank5
8	−1.016 Rank5	−1.566 Rank8	−1.625 Rank9	−1.006 Rank4	−1.002 Rank3	−1.542 Rank7	−0.72 Rank1	Query Term Absent	−1.402 Rank6	−0.728 Rank2
9	−7.016 Rank8	−6.941 Rank7	−7.174 Rank9	−6.047 Rank3	−6.736 Rank5	−6.426 Rank4	−6.792 Rank6	−5.777 Rank1	−5.842 Rank2	−7.196 Rank10

Conclusion and Trends

The main problem of CLIR is poor performance that occurs due to query term mismatching, un-translated query words, multiple representations of query terms, wrong translation and small size of query. TSV has shown potential in improving the relevancy. Among the strategies explored,the best results are obtained in case1 where query expansion performed by adding lowest frequency words
It can be extended to many Indian languages to address ambiguity issue in web queries.
Snippet as test collections are the best dataset for all most cases.

DAY 11 (16/05/22)

MORNING SESSION

Topic: Resource Creation for NLP
Resource Person: Dr. Girish Nath Jha
Profile: Professor, School of Sanskrit and Indic Studies, JNU
Time: 10:15am-1:00pm


The Professor started the session by discussing about **NLP system in Multilingual societies**. He mentioned about 2011 census which identified about 217 mother tongues present in India. There are even more varieties of languages present. To deal with such variety of languages we need strong NLP systems. He also talked about the study conducted at JNU, regarding how people uses Hindi.

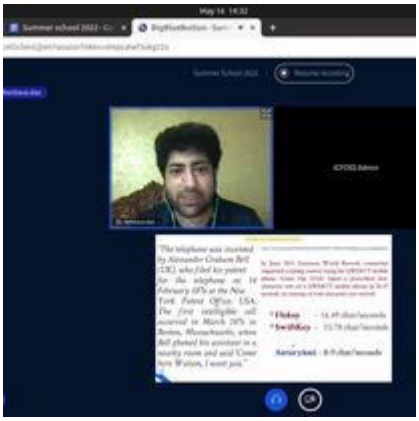
Requirements for resource creation for NLP

- Availability of data
- Standards(creation/collection, annotation, storage, versioning)
- Manpower
- Tools(collection/ creation, annotation, management, bootstrapping)

He also mentioned about topics like annotating data, languages used in judiciary, etc. Then he shown the demo of certain resources like the one by ILCI for data annotation. Finally, he concluded the session with a question answering session.

Afternoon Session

Topic: CodeMix Data Analysis
Resource Person: Dr. AmitavDas
Profile: Lead Scientist , Wipro AI Lab
Time: 2.00pm-5:00pm



Started with antaryami english hindi codemix tool. Then went on explaining how codemix happens globally by a detailed graph.

Conditional probability, Maximum likelihood estimate, Bigram probabilities, Normal and gaussian distribution of languages, smoothing, Language model evaluation, Language model perplexity
Lower perplexity = Better Model, Heaps law, Positional Encoding, various word embedding models
CM with neural networks, BERT Model

DAY 12 (17/05/22):
MORNING SESSION
Topic: Speech Processing
Resource Person: Ms. Lekshmi K R
Profile: Digital University
Time: 10:00 am-1:00 pm

The Speaker started the session by briefly discussing the current trends in Speech recognition by explaining various techniques used in each steps such as Preprocessing, Feature extraction, acoustic modeling, linear modeling ,acoustic analysis .

MORNING SESSION

Topic: Automatic Speech Recognition
Resource Person:Dr. Elizabeth Sherly

Profile: Director ,ICFOSS
Time: 11:45 am-1:00 pm

The Speaker started the session with the history of automatic speech recognition,

Speech visualization and its challenges.

Brain computing
How might computers use it?

1. Digitization
2. signal processing
3. Prosodic features
4. Formant features
5. Cepstrum features
6. Acoustic features

ASR Architecture

Showed a video on ASR

<u>Project Presentation</u>

Project presentation of the participants started at 2.10 pm wherein 9 groups presented their project. The best projects won prizes

Valedictory Function

The session concluded with a valedictory function,distribution of prizes for best projects,for the gaming sessions that we have conducted and distribution of certificates.

==========