Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

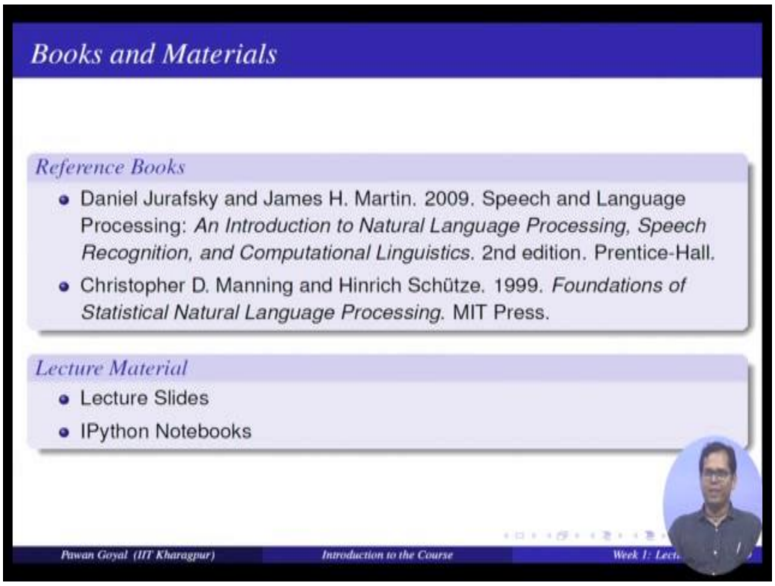
Lecture – 01
Introduction to the Course

So, hello everyone and welcome to this course on Natural Language Processing; so in this course we will be having this course for 12 weeks and in which week will have 5 modules. So, today we are starting this course and this is the first module, first lecture for this week and this is basically the course introduction: what are the different topics we will be covering this course, what are the different text books that you might be using for this and some other details that might be necessarily for you.

So firstly, my contact - if you have certain questions you may also write to me on this email ID. So, this is my official email id and you might also want to visit my web page that is provided in this link.

So, in this course we will have two teaching assistants. So, Amrith Krishna and Mayank Singh both are my PhD students and they are working on NLP. So, idea is that they can; they will be able to help you with any of the queries that you having during this course and they will also help you with all the assignments that we will provide this course. So, in addition to me they will also be your primary contacts during this course.

(Refer Slide Time: 01:37)



Books and Materials

Reference Books

- Daniel Jurafsky and James H. Martin. 2009. Speech and Language Processing: *An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.* 2nd edition. Prentice-Hall.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press.

Lecture Material

- Lecture Slides
- IPython Notebooks

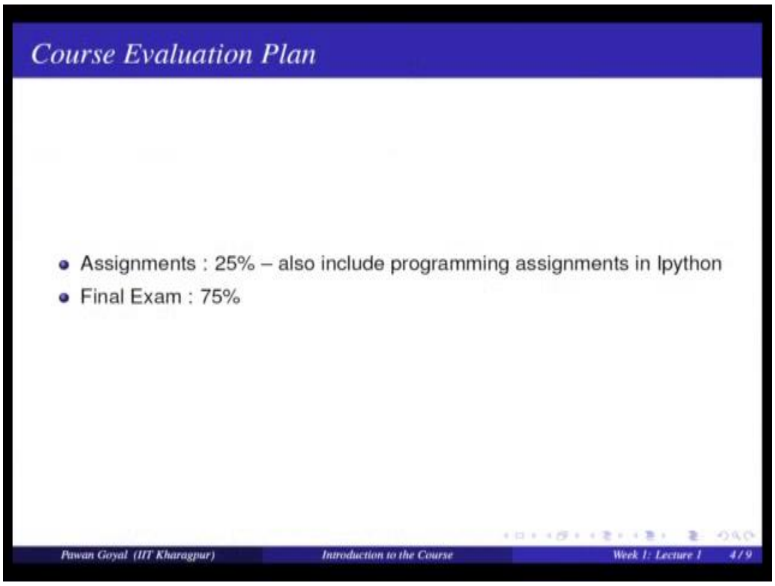
So, in this course we will be following two of the main very very popular books, on natural language processing. So, Jurafsky and Martin is a very popular book, title is speech natural language processing. So, this is your second addition, but there may be also be the third addition that is available. So, you might you second additional and any later additional of this book for this course.

Then there is a book by Manning and Schutze on foundations of a statistical natural language processing by MIT press, that is again a very very popular book and we will be using concepts from both the books along with some other sources in this course. So, you can try to avail any of these two books or both the books if possible and with the course some addition reading from this book will be helpful for addition the concepts and getting a good class on the course.

So, in addition to the books you will also get the lecture slides. So, from the lectures that I will be taking this course, all those slides will be made available to you and if necessary I will also point out to some additional a person and any other material that might be helpful for you, so that also will provide on the course website and we also thinking of giving you some IPython note books so that you can also do some sort of hands on.

So, these note books are so you will be using python for doing certain text (Refer Time: 03:05) tasks and so this we will also provide you some basic instructions on how to start teaching this IPython notebooks. So, they will also very very helpful for this course because this is NLP is mostly and hands on. So, it is not very very nice just to have an NLP course to have only the theoretical knowledge it is important also so that given a problem, you can a start doing some processing over that ok.

(Refer Slide Time: 03:36)



Course Evaluation Plan

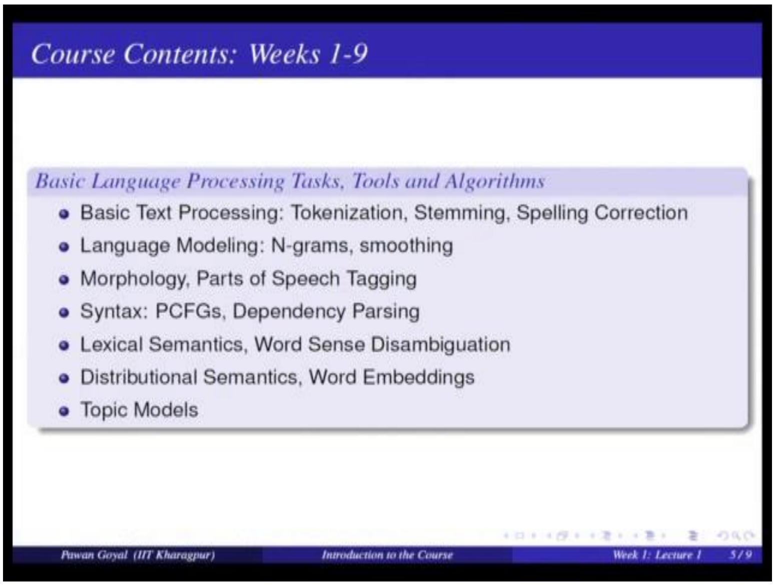
- Assignments : 25% – also include programming assignments in Ipython
- Final Exam : 75%

Pawan Goyal (IIT Kharagpur) Introduction to the Course Week 1: Lecture 1 4 / 9

So, in for the evaluation, we will have assignments that will be given after every week throughout the course. So, that will constitute 25 percent of the whole evaluation of for this course. So, they will assignments will be on the lectures that are covered in this course; also we might also give you some sort of programming assignments time to time in this course

So, they will be all been in python plus, there will with the final exam, that will be constituting 75 percent of the overall weightage, that mean the after the end of this course.

(Refer Slide Time: 04:13)



Course Contents: Weeks 1-9

Basic Language Processing Tasks, Tools and Algorithms

- Basic Text Processing: Tokenization, Stemming, Spelling Correction
- Language Modeling: N-grams, smoothing
- Morphology, Parts of Speech Tagging
- Syntax: PCFGs, Dependency Parsing
- Lexical Semantics, Word Sense Disambiguation
- Distributional Semantics, Word Embeddings
- Topic Models

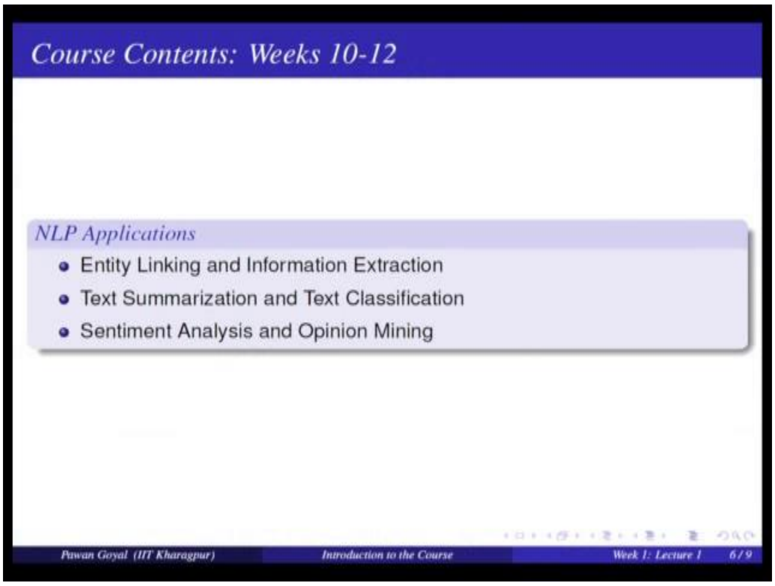
Pawan Goyal (IIT Kharagpur) Introduction to the Course Week 1: Lecture 1 5 / 9

So, what are the different topics that we will be covering this course? So, this course we have starting with some of the basic topics that are required for understanding all the concepts in NLP then we will move to some applications. So, what are the various basic topics that we will be covering this course? So, we will start with the text processing, given a text data how do you start doing some processing over there. So, that may include how do I tokenize it that is breaking in to various words, how do you I start doing lemmatization stemming find out the root words and so on.

Then we will be start with very foundational topic on language modelling that is how do I use the ordering information inside the language for (Refer Time: 05:01) certain applications. So, how can I use this statistics? Then we will go to the morphology, that is what to different categories of words and given a text data, how do I start finding out different categories of words. So, what are the different applications or algorithms that I can use? Then we will go to high level finding out, what are different groups in the sentence, how they are connected to each other in the topic of syntax? Then we will move to semantics, where we will have various models of semantics using lexicon and lexical semantics and using the distributions and distribution semantics and we will also touch upon the topic on word embeddings that is very very popular as of now.

So, finally, we will also cover topic models. So, how do you find out what are the various topics that I have covered in a given text data and how do you I make use of this in various applications.

(Refer Slide Time: 05:59)



Course Contents: Weeks 10-12

NLP Applications

- Entity Linking and Information Extraction
- Text Summarization and Text Classification
- Sentiment Analysis and Opinion Mining

Pawan Goyal (IIT Kharagpur) Introduction to the Course Week 1: Lecture 1 6 / 9

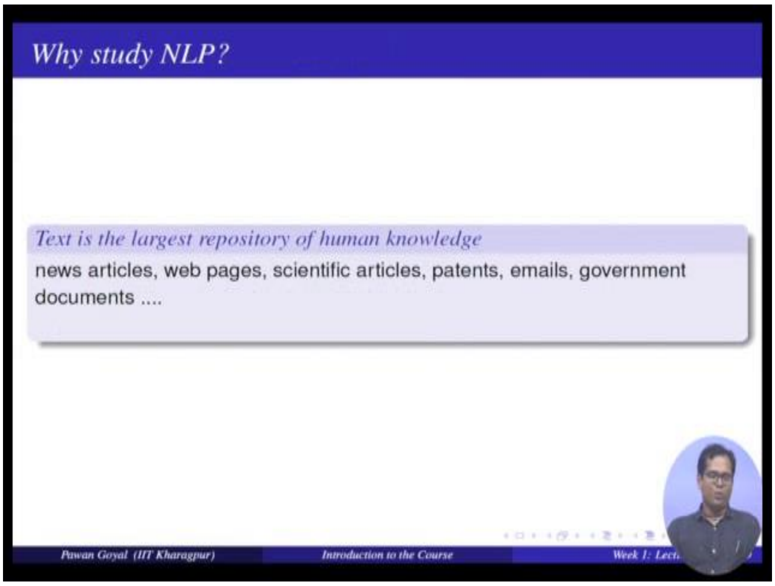
Then after cub once the basic topics are covered, we will also devote some time especially in the last three weeks on how do you a start applying these basic concepts for certain applications. So, NLP is very very broad topic and you will see enumerable applications where you can apply all these concepts, but to give you an idea, so we will take 3 very very interesting and important application in NLP.

So, will start with the topic of entity linking and information extraction, this will be the first application that will be cover then we will go to text summarization and classification. Finally we will end up with opinion analysis and opinion mining. So, this will be the final application that we will be cover and we will hope that whatever basics and some aspects of application that we cover in this course, will help you in taking any new problem and starting to think of your own approach for solving that.

So, idea of this course is that you are not only aware of what are the tools available you can use them on your own, but also you are you know what are the basic algorithms, what are the foundations and given a new problem you can think of a approach on your own and build your own tools. So, that is the goal of this course and as I said earlier this is million introductory courses; but all the basics that are required for dealing with any advanced topic will be covered.

So, if you take any new research topic that is not covered in this course, the idea would be whatever knowledge is covered will help you to at least start understanding the topic and then go deep in to that.

(Refer Slide Time: 07:41)



Why study NLP?

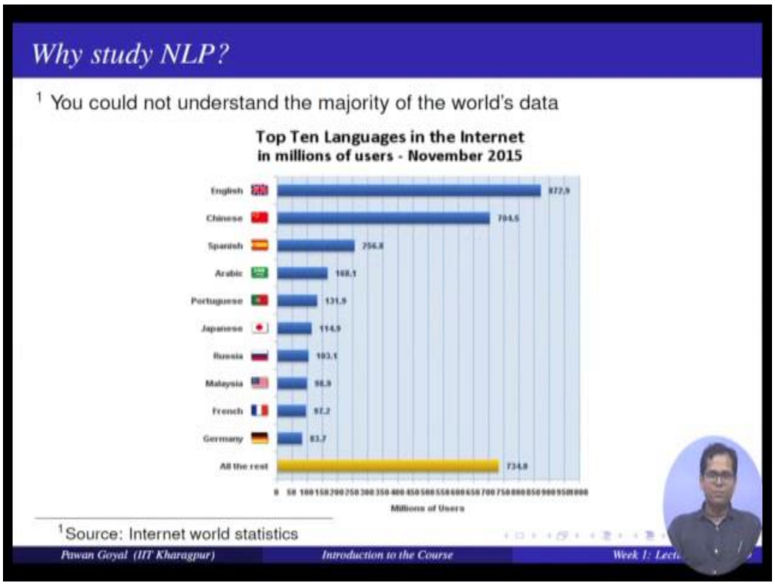
Text is the largest repository of human knowledge

news articles, web pages, scientific articles, patents, emails, government documents ....

So, now, if you work on why do we need to study NLP? So, what is NLP? So, NLP is all about processing text data. So, now, you see on everywhere, you can find out this abundance of text data, now you can also see the text is the largest human knowledge a pastry that you have. So, what are the various sources where you find all these knowledge in text data? So, you can think of all the Wikipedia articles for instance, all the news articles that come dealing, all the scientific articles they are available and text format, patents, all with the all the social media all they have tweets, face book posts and everything is also available in the text format.

So, now this abandons of text data and this is all quite unstructured. So, to be able to make use of this information and build some nice applications, you should know how to process this data and that is why NLP comes into picture. So, with recently with all the tweets, face book posts, comments over all the news articles everywhere and Quora. So, you have an abundance of text data and you should you should be aware of some tools by which you can do certain analysis of this data.

(Refer Slide Time: 09:03)



Why study NLP?

[1] You could not understand the majority of the world's data

[1] Source: Internet world statistics

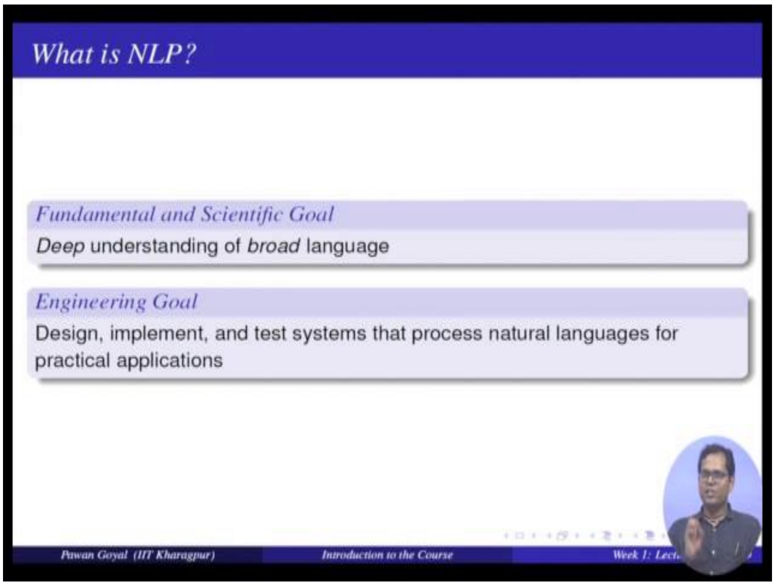
Pawan Goyal (IIT Kharagpur) Introduction to the Course Week 1: Lect...

So, this is one thing that we have a lot of text data available, this is another problem that so much data is available that is unstructured, but this is not in the single language for example, in its not in English. So, this gives you some, this chart gives you some a statistics, so what you will see here. So, this is from November 22 2015 how many millions of users are there on internet and different languages. You see yes English users are the highest, at the same time you have lots and lots of Chinese users and then Spanish and Arabic and again you see large number for all the rest language that are not covered in this chart.

So, you will not be able to understand all the languages, and you will need a tool that can take any given language and try to put it in the language that you can understand. So, some sort of tools are necessary so that they can automatically find out what is the language provided in this document, how do I translate it into some languages that I know and so on, doing some information summarization over there and translation; you need NLP in all of these aspects. So, now this lecture I will just end with dealing saying what is NLP. So, what is the main goal of NLP as such, what for this research field of NLP? So, if you think about it, there is a very very broad scientific goal of this field of NLP and that is can be understand language, can we have a very very deep understanding on how human process language, can we teach computers how to understand language.

So, that is I will say a deep scientific goal behind all the NLP. So, can we teach computers on understanding language and they can may be respond with humans the way humans do and so on. So, that is very very longest and in goal of NLP.

(Refer Slide Time: 11:05)



What is NLP?

Fundamental and Scientific Goal

Deep understanding of *broad* language

Engineering Goal

Design, implement, and test systems that process natural languages for practical applications

So, this we would say is mainly a fundamental scientific goal. I want to have a deep understanding of broad natural language, but with that we also be very very practical and engineering goal and what is that? So, the goal is that. So, yes there is lot of text it available, even if I do not know how to make computers understand that fully, can I still make use of this data to build some very very nice applications that will be helpful for people in they really life and that you can see from you own life, so how many different applications that you might be using. Starting from the search, that is very very common application that everyone uses and so you will see a lot of other applications starting from news recommendations and all that, that you are using a new daily life.

So, this is the engineering goal for NLP, that is can I design implement and test systems that will process natural language and that I design for very very practical applications that we can use in day to day life. So, this is the engineering goal of NLP and that is what we will be mainly focusing in this course and NLP. So, all the concepts and algorithms that we will discuss in this course are mainly focus towards, various engineering applications of NLP.

So with that I again welcome you to this course. So in the next module or the next lecture, we will discuss what do we actually do in NLP by taking some simple examples.

Thank you.