Natural Language Processing (CS-472) Spring-2023

Muhammad Naseer Bajwa

Assistant Professor,
Department of Computing, SEECS
Co-Principal Investigator,
Deep Learning Lab, NCAI
NUST, Islamabad
naseer.bajwa@seecs.edu.pk

NUST
Defining futures
School of Electrical Engineering
& Computer Science

What do you expect from this course?

Overview of this week’s lecture

Introduction to NLP

- Rationale of NLP

- NLP Pipeline

- Course Plan

Objectives of this course are three fold

- To establish the foundation of effective modern methods of deep learning applied to NLP.

- To provide a broader understanding of natural languages and challenges in understanding and producing them.

- To afford sound command of and ability to build systems for some of the major NLP problems.

    - Word meaning

    - Machine translation

    - Question answering

    - and more …

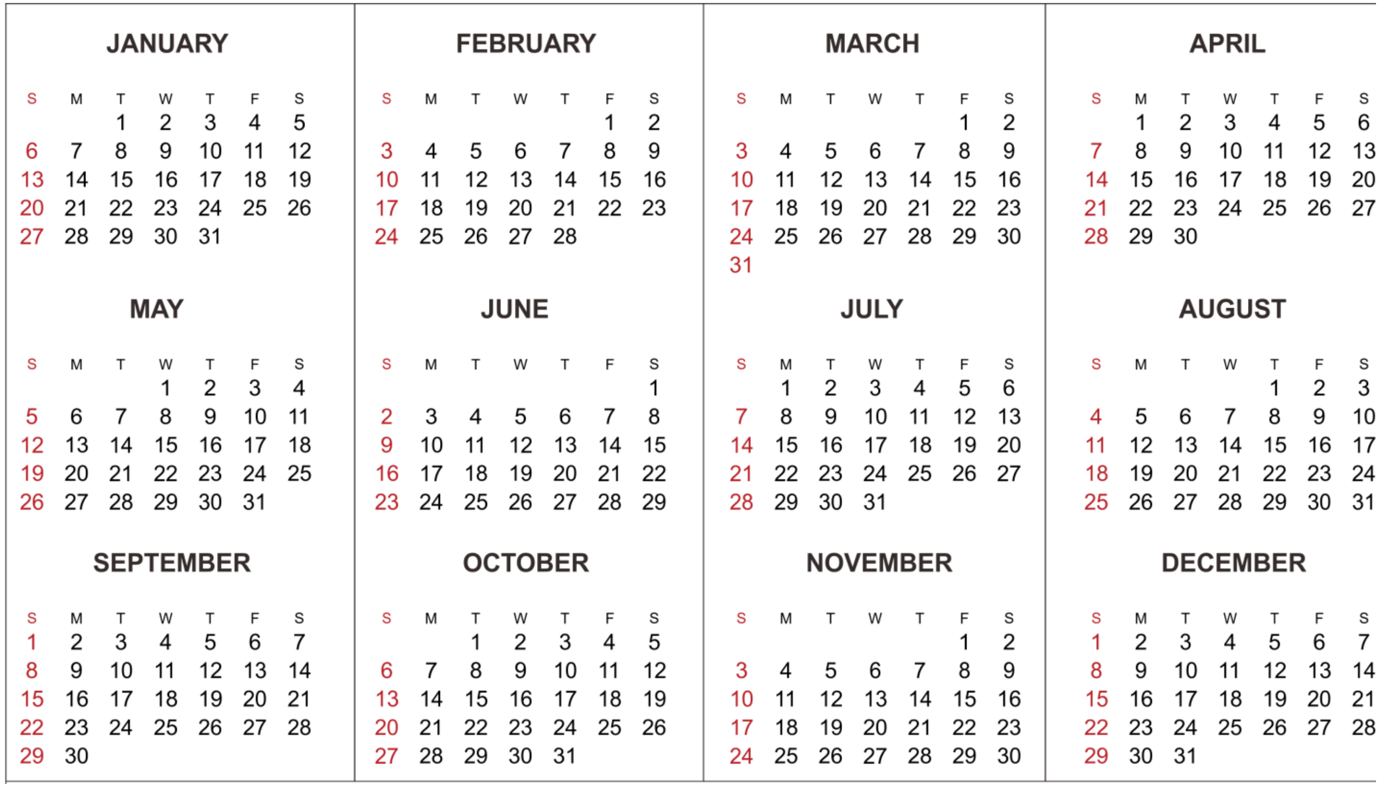
Natural language is a relatively new evolutionary feat

- Language developed around 100,000 to 1 million years ago. (age of the universe is around 13.7 billion years)

Natural language is a relatively new evolutionary feat

- Language developed around 100,000 to 1 million years ago. (age of the universe is around 13.7 billion years)

- If we project 13.7 billion years on 365 days;



JANUARY

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

FEBRUARY

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

MARCH

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

APRIL

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

MAY

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

JUNE

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

JULY

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

AUGUST

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

SEPTEMBER

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

OCTOBER

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

NOVEMBER

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

DECEMBER

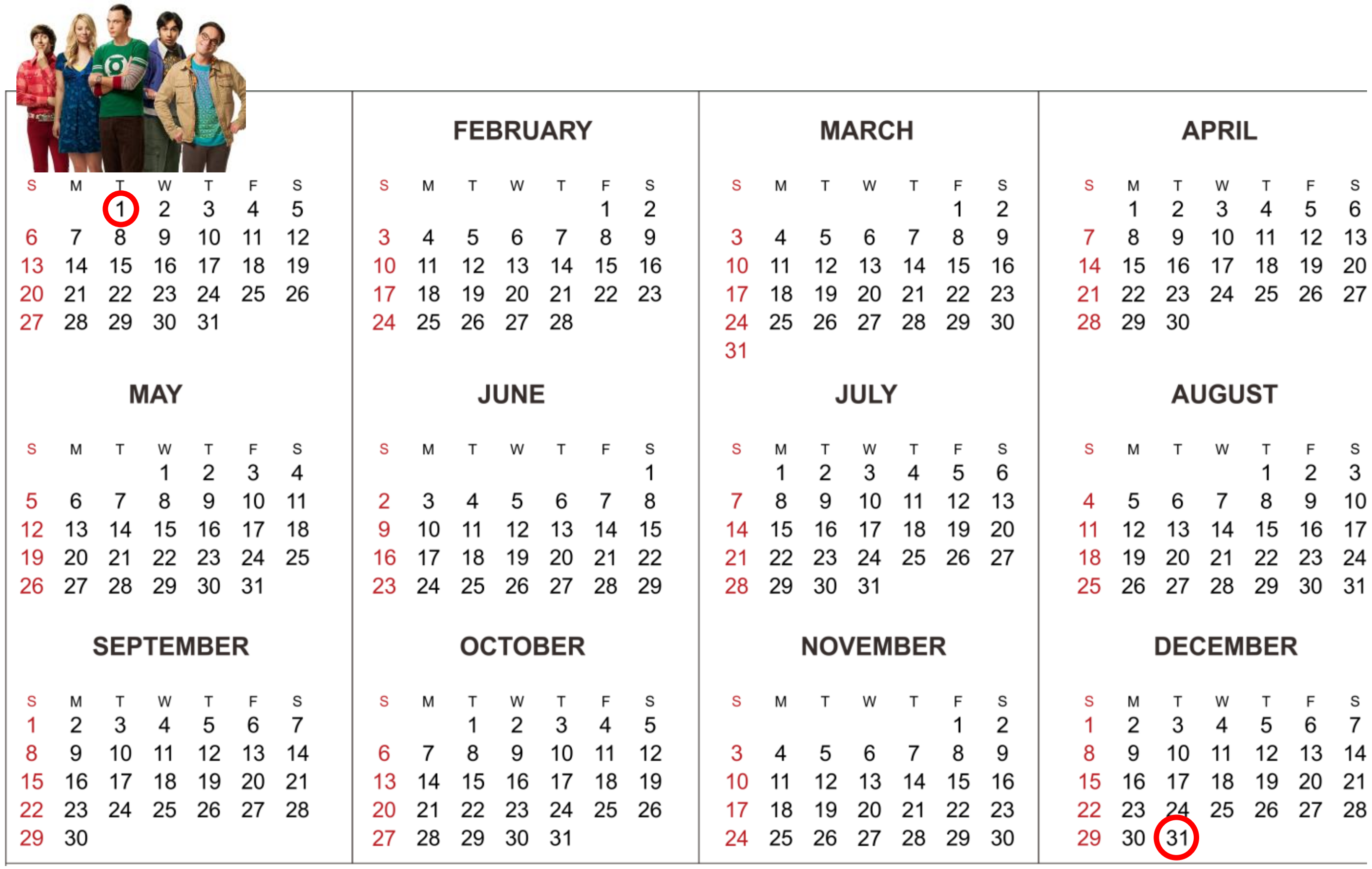
S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Natural language is a relatively new evolutionary feat

- Language developed around 100,000 to 1 million years ago. (age of the universe is around 13.7 billion years)

- If we project 13.7 billion years on 365 days;

    - Each day = 37.5 million years
    - Each hour = 1.56 million years
    - Each minutes = 26000 years

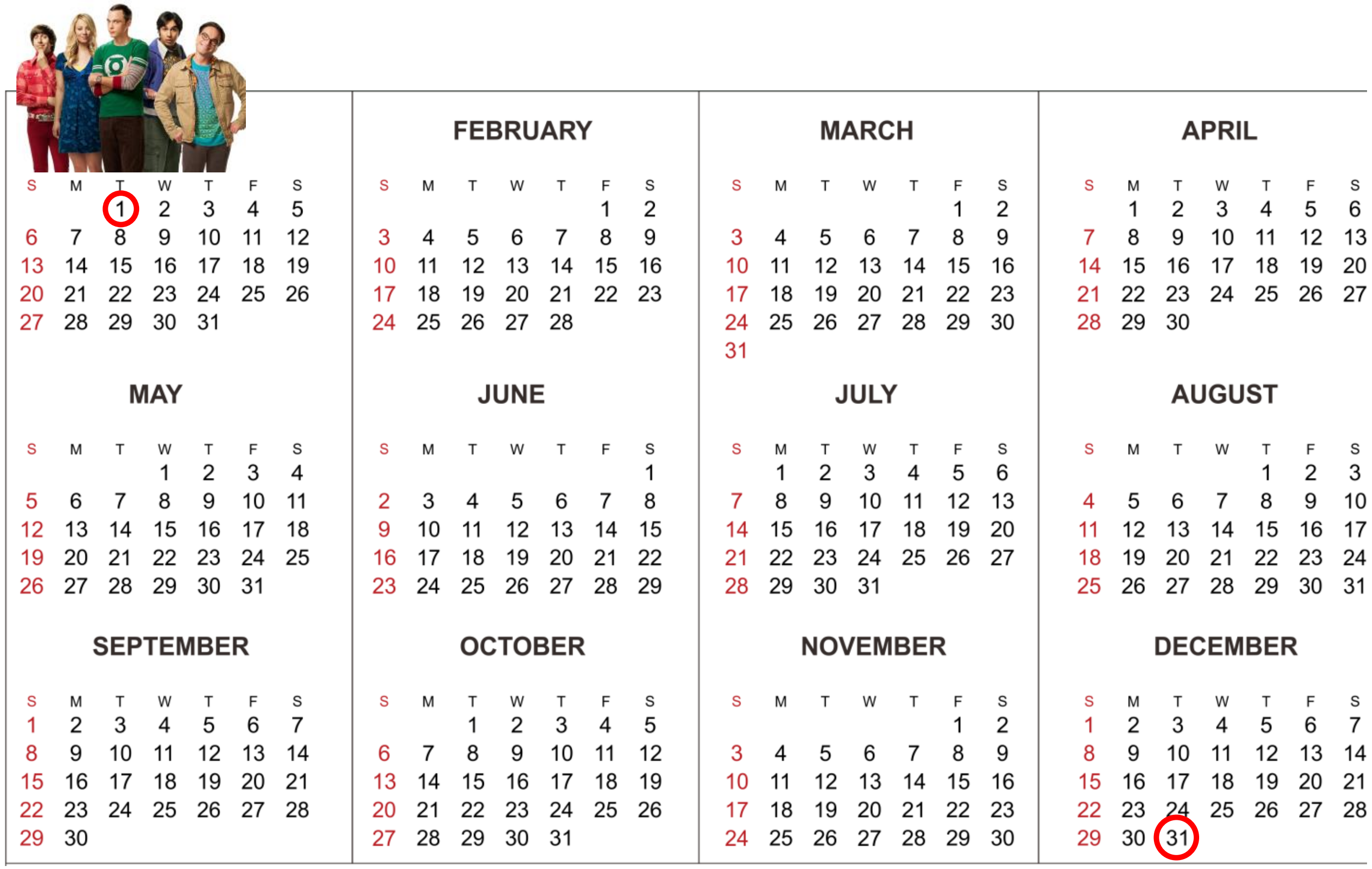


Natural language is a relatively new evolutionary feat

- Language developed around 100,000 to 1 million years ago. (age of the universe is around 13.7 billion years)

- If we project 13.7 billion years on 365 days;
  - Each day = 37.5 million years
  - Each hour = 1.56 million years
  - Each minutes = 26000 years

- Language emerged 4 – 40 minutes ago.



S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

FEBRUARY

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28		

MARCH

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

APRIL

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

MAY

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

JUNE

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

JULY

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

AUGUST

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

SEPTEMBER

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

OCTOBER

S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

NOVEMBER

S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

DECEMBER

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Natural language is a relatively new evolutionary feat

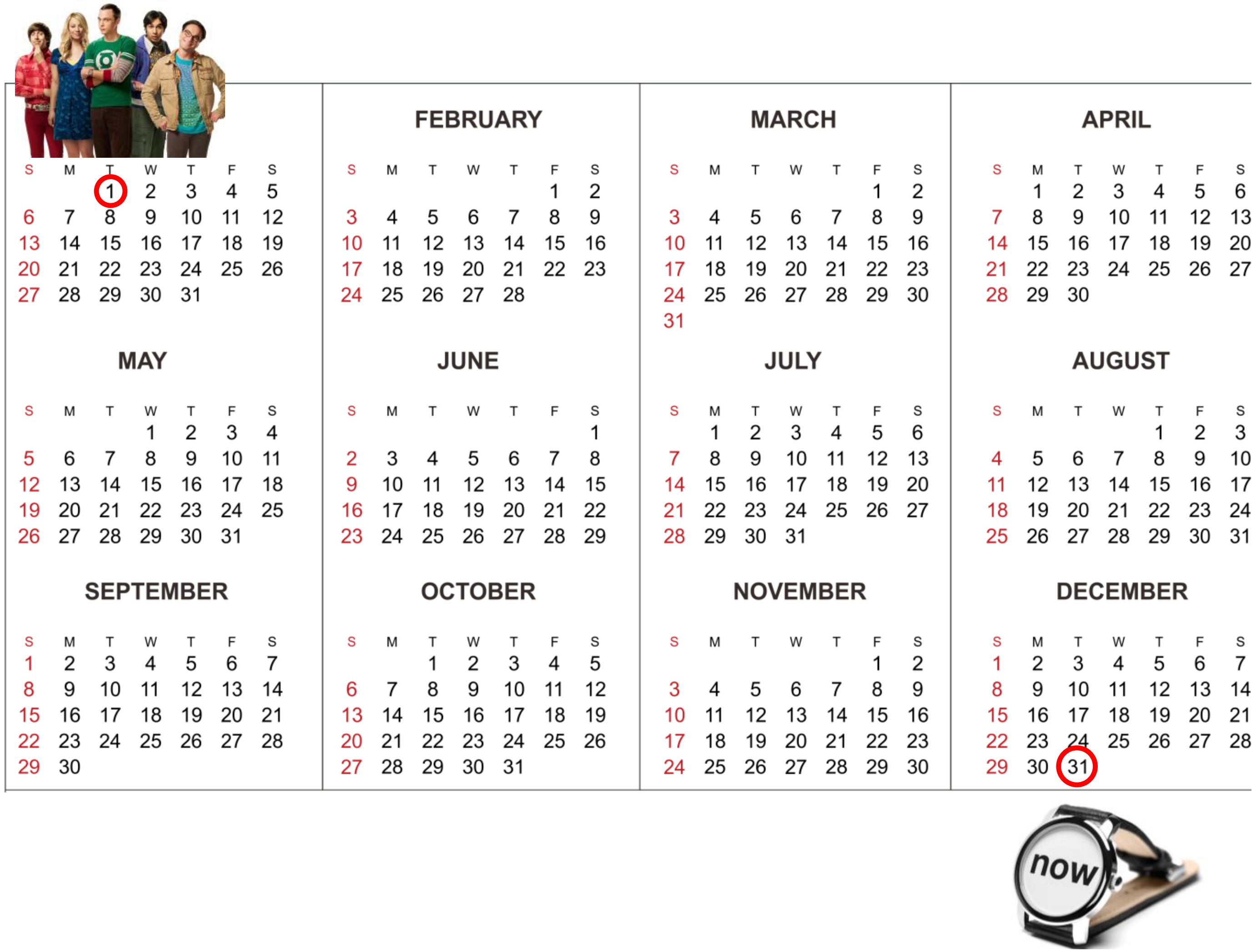
- Language developed around 100,000 to 1 million years ago. (age of the universe is around 13.7 billion years)

- If we project 13.7 billion years on 365 days;

  - Each day = 37.5 million years
  - Each hour = 1.56 million years
  - Each minutes = 26000 years

- Language emerged 4 – 40 minutes ago.

- Writing emerged around 5000 years ago. (11 – 12 seconds ago.)



What is NLP?

- NLP is a branch of AI that,
  - Receives human language (spoken or written) as input.
  - Processes it to understand the meaning.
  - Act upon it or respond.

What is NLP?

- NLP is a branch of AI that,
  - Receives human language (spoken or written) as input.
  - Processes it to understand the meaning.
  - Act upon it or respond.
- NLP is a multidisciplinary field
  - It borrows concepts from AI, Computational Linguistics, and Cognitive Science among others.

Why is NLP?

- Language helps preserve and propagate knowledge.
  - Across time, across space

Why is NLP?

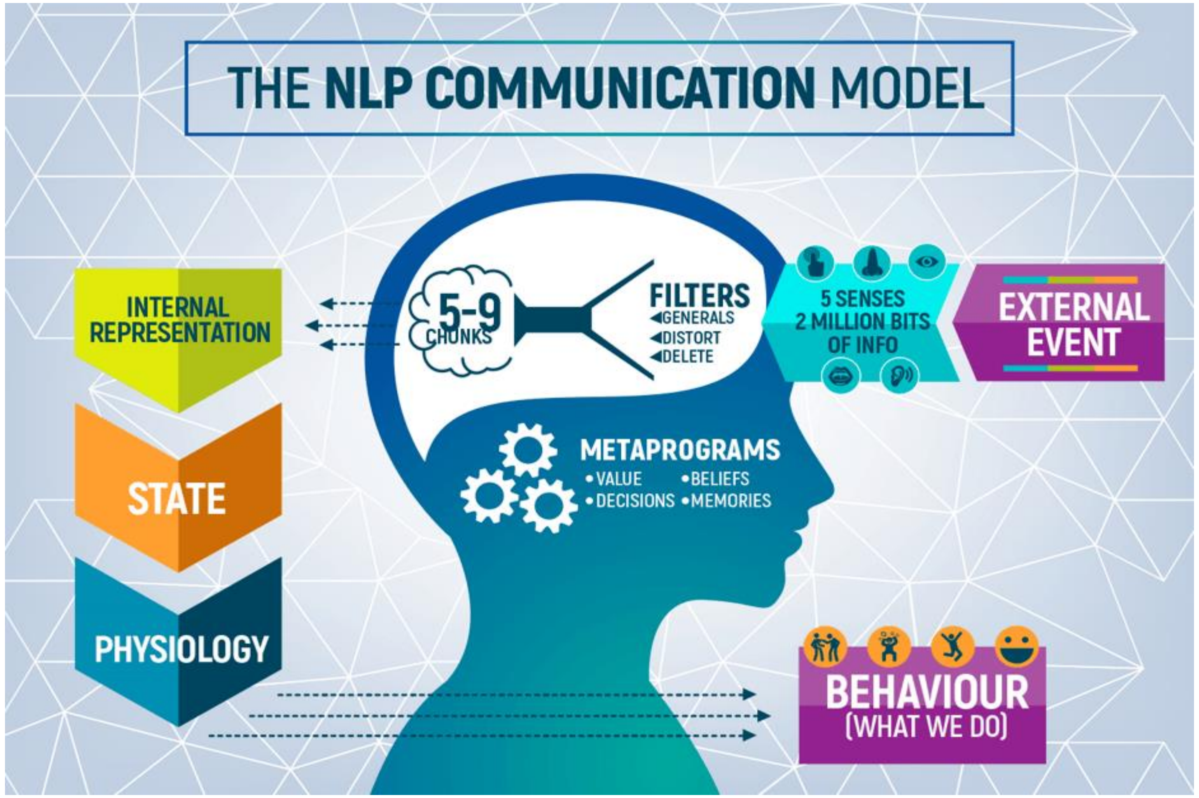
- Language helps preserve and propagate knowledge.
    - Across time, across space
- Bandwidth of humans to generate, process and propagate knowledge is limited.

HOW HAVE YOU FARED, SINCE LAST WE SAW EACH OTHER?
SINCE LAST YOU SAW ME, OR LAST I SAW YOU?
I SAW YOUR ATTEMPT TO SEE ME. TWAS' SUCH A SIGHT TO SEE.
AH, BUT I DO BELIEVE I HAVE SEEN YOU EVEN MORE RECENTLY THAN WHEN YOU SAW ME, SEEING YOU.
YOU... I... UH...
ANYBODY HERE?
NOPE.
CAN WE JUST SPEAK NORMALLY THEN?
OH MY GOD YES! THIS CONVERSATION HAS BEEN SIMPLY EXHAUSTING.
WANNA GRAB A CUP OF COFFEE AND TALK ABOUT THE REALM?

"VERBOSITY" - A COMIC BY 1D4ROUNDS

Why is NLP?

- Language helps preserve and propagate knowledge.
    - Across time, across space
- Bandwidth of humans to generate, process and propagate knowledge is limited.
- If computers can understand natural language it can greatly help advance human civilisation.



How can we teach computers human languages?

- To make computers understand human language, first we need to understand it ourselves.
  - She poured the water from jug into the cup until it was full.

How can we teach computers human languages?

- To make computers understand human language, first we need to understand it ourselves.
  - She poured the water from jug into the cup until it was full.
  - She poured the water from jug into the cup until it was empty.

How can we teach computers human languages?

- To make computers understand human language, first we need to understand it ourselves.
  - She poured the water from jug into the cup until it was full.
  - She poured the water from jug into the cup until it was empty.
- Human languages are complex.
  - Words only don’t convey the meaning.

How can we teach computers human languages?

- To make computers understand human language, first we need to understand it ourselves.
  - She poured the water from jug into the cup until it was full.
  - She poured the water from jug into the cup until it was empty.
- Human languages are complex.
  - Words only don’t convey the meaning.
- They have a very small bandwidth.
  - Cannot convey large amount of information in short amount of time.

How can we teach computers human languages?

- To make computers understand human language, first we need to understand it ourselves.
  - She poured the water from jug into the cup until it was full.
  - She poured the water from jug into the cup until it was empty.
- Human languages are complex.
  - Words only don’t convey the meaning.
- They have a very small bandwidth.
  - Cannot convey large amount of information in short amount of time.
- Natural languages make use of compression.
  - Requires knowledge of listener and context to fill the gaps.

Pipeline of NLP consists of three main components

- Text Processing
    - Take raw input text, clean it, normalize it, and convert it into a form that is suitable for feature extraction.

Pipeline of NLP consists of three main components

- Text Processing
  - Take raw input text, clean it, normalize it, and convert it into a form that is suitable for feature extraction.
- Feature Extraction (word/document representation)
  - Extract and/or produce feature representations that are appropriate for the type of NLP task at hand and the type of model to be used.

Pipeline of NLP consists of three main components

- Text Processing
    - Take raw input text, clean it, normalize it, and convert it into a form that is suitable for feature extraction.
- Feature Extraction (word/document representation)
    - Extract and/or produce feature representations that are appropriate for the type of NLP task at hand and the type of model to be used.
- Modelling
    - Design a model, fit its parameters to training data, use an optimization procedure, and then evaluate it to make predictions about unseen data.

Pipeline of NLP consists of three main components

- Text Processing
  - Take raw input text, clean it, normalize it, and convert it into a form that is suitable for feature extraction.
- Feature Extraction (word/document representation)
  - Extract and/or produce feature representations that are appropriate for the type of NLP task at hand and the type of model to be used.
- Modelling
  - Design a model, fit its parameters to training data, use an optimization procedure, and then evaluate it to make predictions about unseen data.

There could be additional steps depending upon application

Pre-processing operations depend on the use case

- **Cleaning:** Remove irrelevant items like HTML tags, symbols and non-alphabetic characters.

- **Normalisation:** Convert all words to lowercase and removing punctuation and extra spaces.

- **Tokenisation:** Split the text into words, also known as tokens.

- **Stop Words Removal:** Remove the most common words (a, an, the, etc.).

- **Parts of Speech Tagging:** Identify the parts of speech for the remaining words.

- **Named Entity Recognition:** Recognize the named entities in the data

- **Stemming and Lemmatisation:** Convert words into their canonical / dictionary forms, using stemming and/or lemmatization.

Computers cannot understand words, they understand numbers

- Text is represented by ASCII or Unicode which maps each character to a number.
  - Individual characters don’t carry much information/meaning.
  - Representing words as a sequence of ASCII/Unicode numbers does not capture the meaning of a word and its relationship with other words.

Computers cannot understand words, they understand numbers

- Text is represented by ASCII or Unicode which maps each character to a number.
  - Individual characters don’t carry much information/meaning.
  - Representing words as a sequence of ASCII/Unicode numbers does not capture the meaning of a word and its relationship with other words.
- Suitable representation of text data depends on the task.
  - For document-level tasks like sentiment analysis, BOW or doc2vec may be used.
  - For word-level tasks like language generation or machine translation word2vec or GloVe may be used.

Finally an NLP model is trained on the data and evaluated

- Common examples of models are $seq2seq$ models and Transformers.

Finally an NLP model is trained on the data and evaluated

- Common examples of models are $seq2seq$ models and Transformers.

- Once again, picking the right model depends upon the task.

What will we study in this course?

- Feature extraction: How to convert words into vectors.

What will we study in this course?

- Feature extraction: How to convert words into vectors.

- Network Architectures:

  - Brief revision of deep learning, 1D CNNs and RNNs/LSTMs
  - $seq2seq$ models and attention
  - Transformers, self-attention and pretraining transformers (basis for GPT)

What will we study in this course?

- Feature extraction: How to convert words into vectors.

- Network Architectures:

  - Brief revision of deep learning, 1D CNNs and RNNs/LSTMs
  - $seq2seq$ models and attention
  - Transformers, self-attention and pretraining transformers (basis for GPT)

- Applications:

  - Machine Translation
  - Question Answering
  - Natural Language Generation, and more

What will we study in this course?

- Feature extraction: How to convert words into vectors.

- Network Architectures:
    - Brief revision of deep learning, 1D CNNs and RNNs/LSTMs
    - $seq2seq$ models and attention
    - Transformers, self-attention and pretraining transformers (basis for GPT)

- Applications:
    - Machine Translation
    - Question Answering
    - Natural Language Generation, and more

- Advanced Topics:
    - Biases in AI
    - Model Analysis and Explanations

The course logistics will be as follows

- A Term Project (up to 10 marks): Make a group of 2-3.

  - Four deliverables:
    - Proposal (Introduction)
    - Mid-Semester Report (Progress Report)
    - Final Report (IEEE conference format)
    - Presentation (5 minutes each group)

The course logistics will be as follows

- A Term Project (up to 10 marks): Make a group of 2-3.
  - Four deliverables:
    - Proposal (Introduction)
    - Mid-Semester Report (Progress Report)
    - Final Report (IEEE conference format)
    - Presentation (5 minutes each group)
- Quizzes (up to 10 marks):
  - Truly unannounced
  - Whenever, Whatever
  - All quizzes will be graded. No make-ups

The course logistics will be as follows

- A Term Project (up to 10 marks): Make a group of 2-3.
    - Four deliverables:
        - Proposal (Introduction)
        - Mid-Semester Report (Progress Report)
        - Final Report (IEEE conference format)
        - Presentation (5 minutes each group)
- Quizzes (up to 10 marks):
    - Truly unannounced
    - Whenever, Whatever
    - All quizzes will be graded. No make-ups
- Assignments (up to 10 marks):
    - Individual/Group assignments
    - Discussion/exchanging notes is allowed. Copying is prohibited.

Do you have any problem?

Some material (images, tables, text etc.) in this presentation has been borrowed from different books, lecture notes, and the web. The original contents solely belong to their owners, and are used in this presentation only for clarifying various educational concepts. Any copyright infringement is ***not at all*** intended.