Master of Data Science and Innovation

36118 Applied Natural Language Processing (ANLP)

Session 1

!!! While we wait for the session to start, please make sure that you have access to canvas modules

Dr. Antonette Shibani

Senior Lecturer, MDSI

Acknowledgement of country

I would like to acknowledge the Gadigal people of the Eora Nation upon whose ancestral lands our City campus now stands. I would also like to pay respect to the Elders both past and present, acknowledging them as the traditional custodians of knowledge for this land.

https://www.smh.com.au/culture/books/welcome-back-the-recovery-of-australia-s-indigenous-languages-20201120-p56gfp.html

Teaching staff

Dr. Antonette **Shibani**

Teaching team members

Sarah Fawcett

Core teaching principles

We are here

- to build your foundational skills in NLP and its modern applications
- not just to teach core technical skills, but to develop your potential as an all-round data professional
- to help you think through ethics, value, and applications in the real-world
- to build your higher-order learning skills and interpersonal skills
- to enable a respectful learning community that can build off each other’s knowledge through transdisciplinarity

Communication and Support structure

Canvas subject site: The main platform containing all subject related information and announcements (Make sure you have access and receive announcement emails!). It also lets you request for extensions.

In-class sessions: F2F attendance is mandatory for all sessions (recordings not provided). Email the subject coordinator for exceptional circumstances.

Discussion board: Post all general queries on the canvas discussion board. Peers can respond in addition to the teaching team! (Emails containing general queries will be copied to the channel so information is shared with everyone)

Email the subject coordinator: Individual/ personal matters

Teaching team members: Available for guidance in the F2F sessions on campus, and online drop-ins, discussion forums

Learning strategies

- In-class lectures, tutorials
- Self-directed study **You** are in charge of your learning!
- Coding practice (you will only get better with hands-on experience and learning by doing, there is no shortcut!)
- Peer learning [Pro tip: Form your study group(s)]

Tools for engagement

- Menti meter for short surveys
- Undoing activities, in-class quizzes

- Miro board for collaborative activities – check out the cohort resource board (there's homework for you!)
- Two-minute writings for teaching feedback (After session polls on Canvas)

ANLP Spring 2025 Key dates

Subject schedule (Spring 2025)

Below table shows indicative topics we'll cover in the teaching weeks. All lectures run on-campus and are mandatory to attend (recordings are not provided).

Online tutorials in self-directed study weeks will be facilitated by the tutor and will focus answering questions and providing advice on subject content/ code - these are highly recommended but are optional to attend. In self-directed study weeks, you are encouraged to:

- Revisit study material and do additional reading
- Practise coding for NLP
- Try new data sets in exercises, build your learning portfolio
- Work on assessments

Teaching and Learning Schedule

Teaching Week	Week Commencing (2025)	Content * On-campus sessions are mandatory to attend	Assessment/ Key dates
Orientation	21 July	**Getting started** • Familiarise with the subject canvas site • Review subject information and assessments • Mark your calendar for in-class sessions (attendance is mandatory!) • Set up IDE/ Jupyter notebook environment/ Colab for Python • programming	

Canvas Subject Schedule

Week 1	28 July	**In-class session 1:** • Subject Introduction • Foundations of Natural language • Natural Language Processing (NLP) basics • Text analysis introduction using python • Assignment 1 brief release ***Lecture and tutorial session (On-campus): Monday 28 July, 5:30-8:30pm**	
Week 2	4 Aug	**In-class session 2:** • Regular expressions • Visualisation and sense making of text analysis • Topic Modelling • Clustering ***Lecture and tutorial session (On-campus): Monday 4 Aug, 5:30-8:30pm**	
Week 3	11 Aug	**Self-directed study week (Tutorial 1)** **Online Tutorial: Monday 11 Aug, 6-7pm**	Last day to enrol in the subject
Week 4	18 Aug	**In-class session 3 :** • Machine learning and text classification • Sentiment Analysis • Summarization • Assignment 2 brief release ***Lecture and tutorial session (On-campus): Monday 18 Aug, 5:30-8:30pm**	**AT1 Due Monday, 18 Aug**
Week 5	25 Aug	**In-class session 4:** • Vectorisation and embeddings • Deep Learning Basics • Perceptron • Convolutional Neural Network (CNN) • Recurrent Neural Network (RNN) • LSTM (Extended topic)	Census date Thursday 28 Aug 2025 (Last day to withdraw from the subject)

ANLP Spring 2025 Key dates (Ctd..)

[Canvas Subject Schedule]

Week 6	1 Sep	**Self-directed study week (Tutorial 2):** • Revisit study material and do additional reading • Practise coding for NLP • Try new data sets in exercises, build your learning portfolio • Work on assessments **Online Tutorial: Monday 1 Sep, 6-7pm**	
Week 7	8 Sep	**In-class session 5:** • Transformers • BERT • Language models • Large Language Models (LLMs) Intro • Generative AI tools ***Lecture and tutorial session: Monday 8 Sep, 5:30- 8:30pm**	
Week 8	15 Sep	**In-class session 6:** • LLM Deep Dive • Prompting Techniques • LLM Challenges and Risks • LLM Evaluation (Extended topic) • **Assignment 3 brief release** • ***Lecture and tutorial session: Monday 31 Mar, 5:30- 8:30pm**	
Stu Vac	22 Sep	**Self-directed study week (Tutorial 3):** • Revisit study material and do additional reading • Practise coding for NLP • Try new data sets in exercises, build your learning portfolio • Work on assessments **Online Tutorial: Monday 22 Sep, 6-7pm** ****Assessment 2 (Part A) due Wednesday 24 Sep, 11:59pm:** **End to end NLP project: Project Poster**	**AT2a Due Wednesday, 24 Sep**

Week 9	29 Sep	**In-class session 7:** • Building applications with LLMs • Retrieval Augmented Generation (RAG) intro • LLM deployment • Agentic AI ***Lecture and tutorial session: Monday 29 Sep, 5:30- 8:30pm**	
Week 10	6 Oct	**Self-directed study week (Monday 6 Oct public holiday)**	
Week 11	13 Oct	**In-class session 8:** • Ethics in NLP • NLP applications for social good • AI-augmented thinking ***Lecture and tutorial session: Monday 13 Oct, 6-9pm** ***Note the change in start time due to classroom availability** ****Assessment 2 (Part B) due Monday 13 Oct, 11:59pm:** **End to end NLP project - Project report and peer review (Group + Individual Rating on SPARK)**	**AT2b Due Monday, 13 Oct**
Week 12	20 Oct	**In-class session 9:** Project showcase and networking *Lecture and tutorial session: **Monday 20 Oct, 5:30- 8:30pm**	
Stuvac	27 Oct	**Self-directed study** Complete assessments*** ****Assessment 3 due Wednesday 29 Oct, 11:59pm:** **Critical and Ethical Reflection**	**AT3 Due Wednesday, 29 Oct**

Let’s do a quick check-in!

Instructions to join Menti are provided in class

ANLP - Subject Intro

Applied Natural Language Processing

- Human language in textual form has enormous potential, but is often more complex to collect, process, and interpret than numerical data
- Clear need for Natural Language Processing (NLP) as unstructured data grows rapidly in organisations
- The subject aims to build a working knowledge of NLP techniques with practical applications; a strong foundation for your text analysis skills
- Programming language used: Python

Subject Learning Objectives (SLOs)

Upon successful completion of this subject students should be able to:

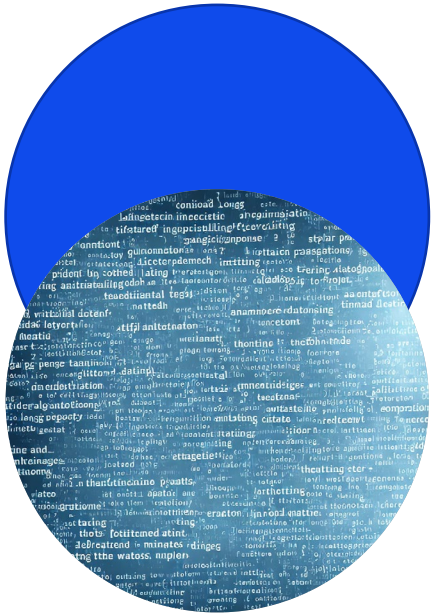
1. Understand core concepts of Natural Language Processing (NLP) and computational linguistics including its limitations (CILO 2.2, 2.3)

2. Evaluate complex challenges for problem solving and build practical NLP applications (CILO 2.3, 4.2)

3. Apply text mining techniques on unstructured data sets using advanced NLP programming packages (CILOs 1.2, 2.2)

4. Interpret, extract value and effectively communicate insights from text analysis and create real-world applications suitable to a range of audiences (CILOs 2.4, 3.2, 4.2)

5. Articulate the strengths, weaknesses and underlying assumptions of NLP and text analysis to apply ethical practices (CILO 5.1, 5.2)



What the subject is not intended to do

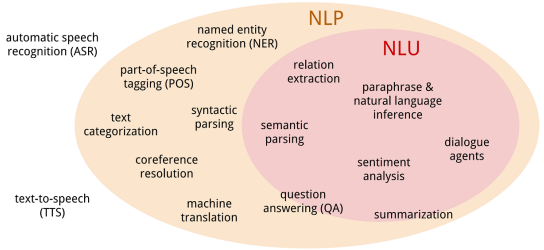


Image Source: Bill MacCartney

❌ Teach programming

(Will be application oriented with coding exercises for key concepts - you should practise a lot at your own time)

❌ Cover all techniques and its mathematical foundations

(NLP is huuugggeee, we will cover the most common applications)

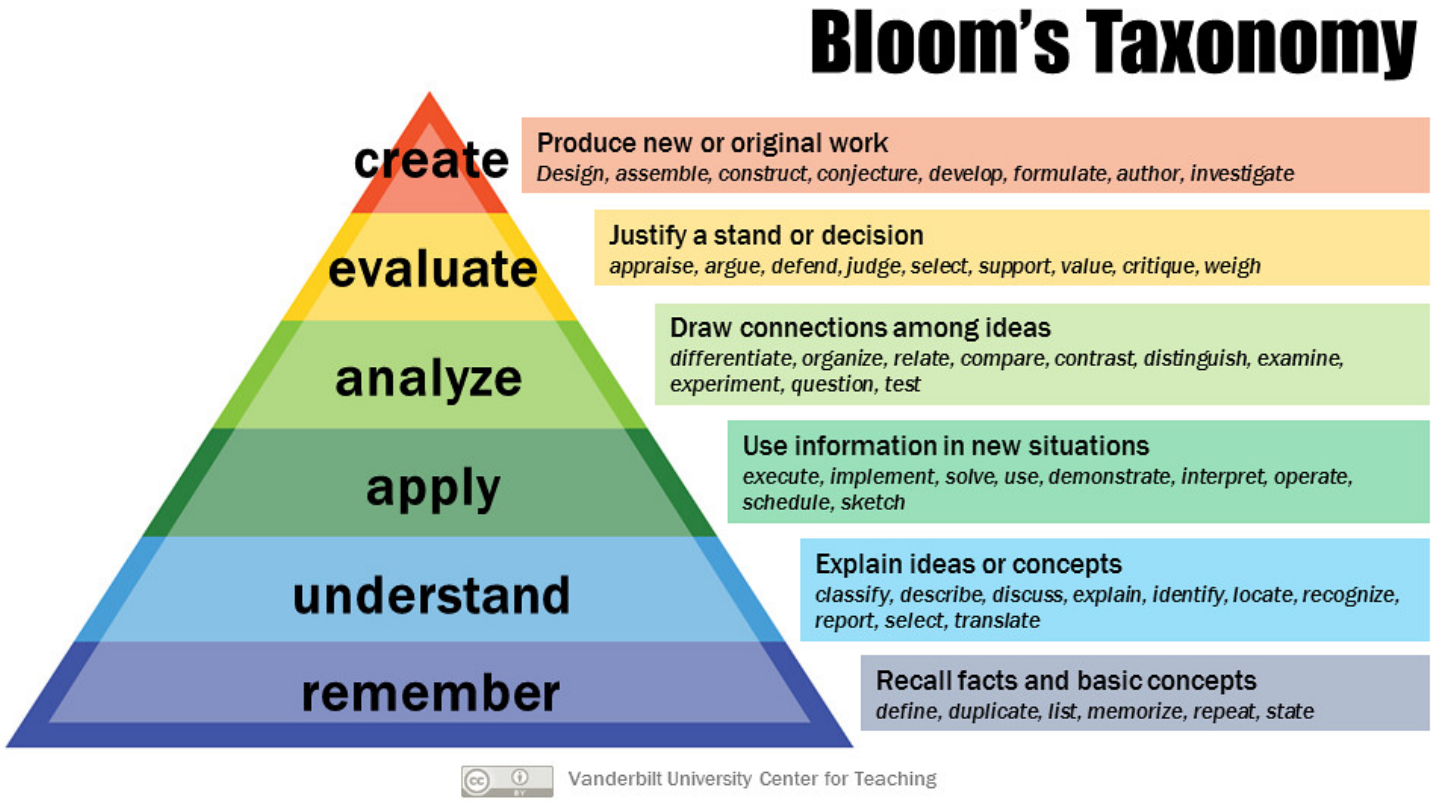
❌ Make you a NLP expert

(It is an ongoing process, lots of self-learning required)

Education is not the learning of facts but **the training of the mind to think**.

-Albert Einstein

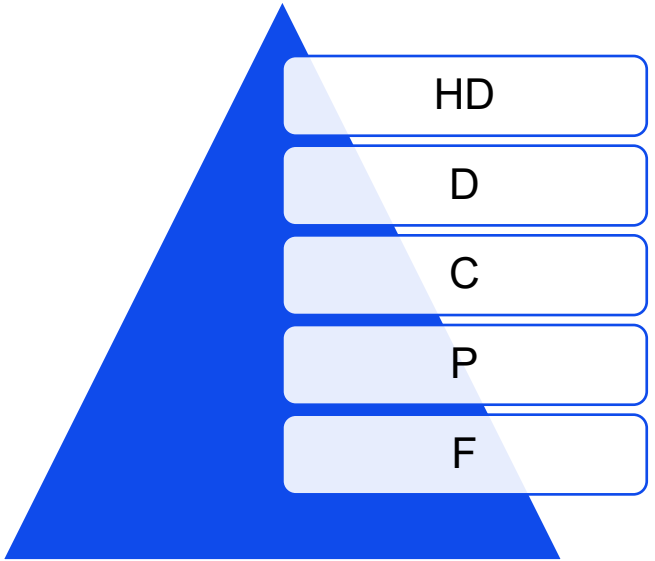
Develop higher-order skills



Assessments

Challenging, but not designed to make your life hard

- Practice-oriented
- Open-ended to showcase creativity and expertise
- Individual (AT1, AT3), and team tasks (AT2a and b, Team size: 4-5)

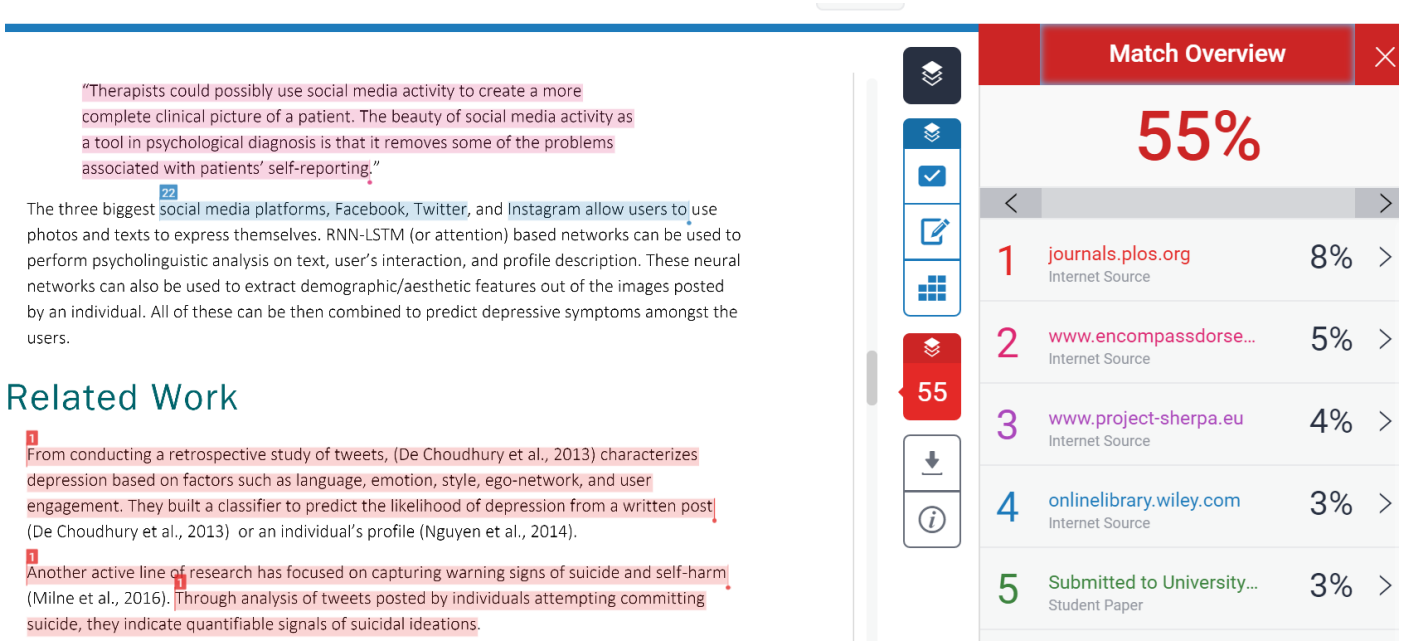


Assessment guidelines

- Clear expectations
- Transparency
- Learning outcomes
- Academic integrity

A detailed assessment brief will be provided on canvas for each assignment after release!

Plagiarism and academic misconduct



Can get you expelled from the degree!

Generative AI (GenAI) and academic integrity

You are allowed to use GenAI in this subject as a tool/ study aid to help with your learning. Indicative examples of allowed use are below:

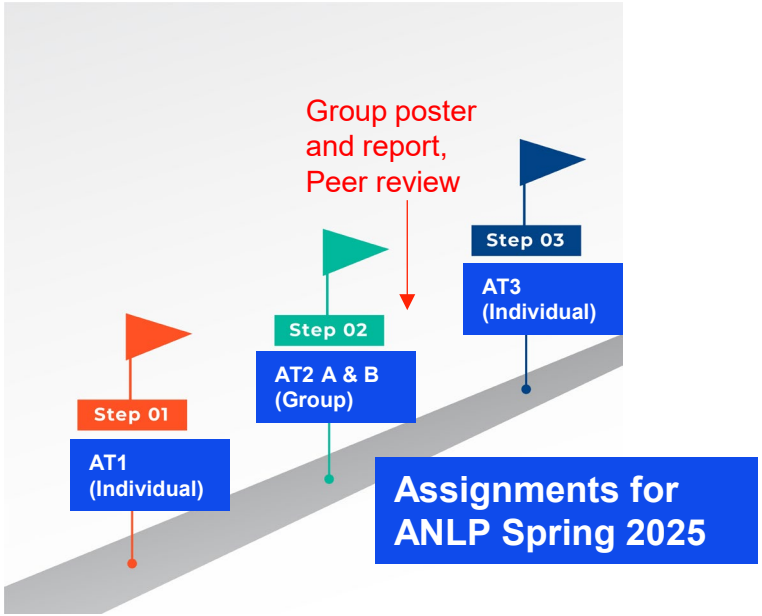
- Getting explanation and examples for a complex topic
- Getting starter ideas/ structuring for your writing
- Correcting your spelling and grammar
- Getting advice for debugging code (use only when stuck, or you might become over reliant!)
- Researching information (should be validated with reliable sources)

In all cases above, and others, you **must acknowledge** use of any copied or reworked materials (e.g. text, images, code) from a GenAI tool by declaring it's use. Find out how to reference your use of GenAI. UTS recommended tool is CoPilot - Logged in via your student email, it provides secure access.

Remember, your assignment submissions must always be original and written by you!

Assessment Structure

! Download the assessment brief from Canvas which contains more information and deadlines



Updated on canvas!

Be passionate about learning

Be kind and empathetic

Build your community/ network

but above all,

TAKE CARE OF YOURSELF!

Census date: Thursday 28 Aug 2025

Before you withdraw, find out about all of your options!

Consult your:
- tutor
- lecturer
- academic advisor
- academic liaison officer
- course coordinator.

UTS Counselling is here to support you!

Contact 9514 1177 to book a face-to-face, Zoom or phone consultation now.

Email queries:
student.services@uts.edu.au

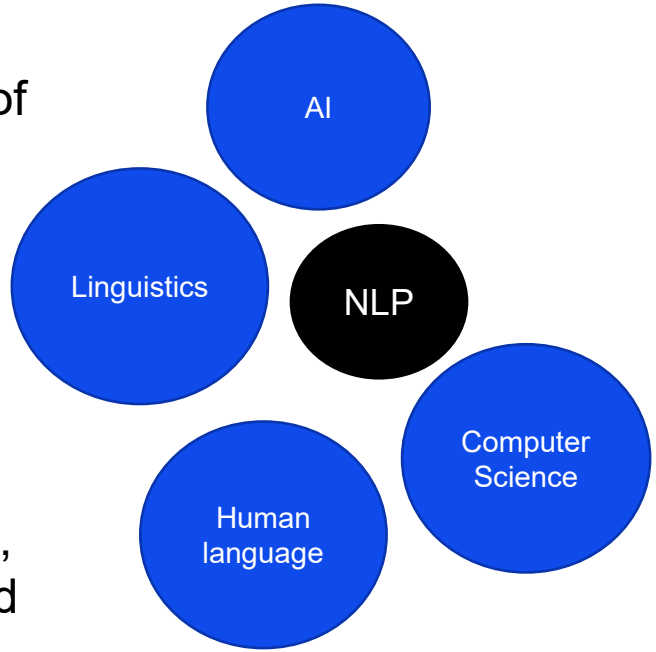
E-therapy options:
- *Mindspot*: mindspot.org.au
- *Mycompass*: mycompass.org.au

Student Service Unit supports at UTS:
Accessibility
Financial assistance
Counselling
Health Service
HELPS
Careers

Foundations of NLP

What is Natural Language Processing (NLP)?

Natural Language Processing (NLP) is a sub-field of artificial intelligence that helps computers understand, interpret and manipulate human language.



“Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.” (Wikipedia)

Towards natural human-to-computer communication

Computers understand
0s and 1s

We write programming
languages to teach it
tasks to do

Ultimate goal:
Understanding human
language in its form

But, it’s not that easy…

Language complexities

Subtlety:

“If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.” (Perfume review in *Perfumes: the Guide*)

Thwarted expectations and ordering effects:

“This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it **can’t hold up**.”

Sarcasm:

“Great idea, now try again with a real product development team” (e-reader)

Implicit knowledge is difficult for AI to learn

“Please take a seat”

INVESTMENTS = CONTAINERS FOR MONEY

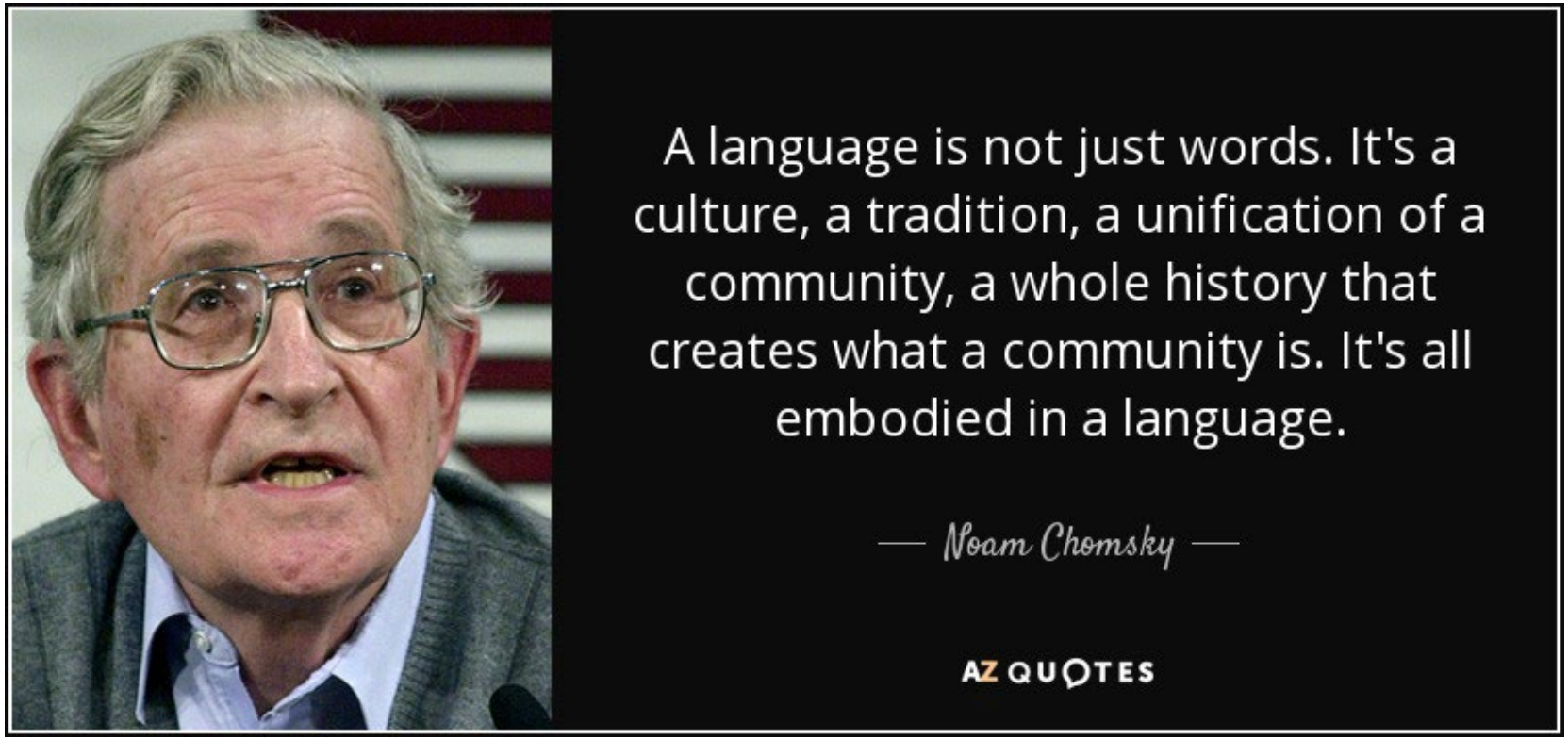
“Put your money in bonds.”

“The bottom of the economy dropped out.”

“I’m down to my bottom dollar.”

“This is an airtight investment.”

“We pooled our funds for the venture.”


A language is not just words. It's a culture, a tradition, a unification of a community, a whole history that creates what a community is. It's all embodied in a language.
— Noam Chomsky —
AZ QUOTES

What you will learn in this subject

Foundations of NLP (core concepts)

- What is NLP, and why is it distinctive?
- Basics of dissecting language using Python

NLP for text analysis and reporting

- Named entity recognition, Keywords and phrases, Lengths and frequencies
- Pre-processing
- Text visualisations
- Storytelling for text data

Topic modelling

Clustering

What you will learn in this subject (2)

Machine learning

- Classification algorithms
- Sentiment Analysis

Neural networks and deep learning

NLU and NLG

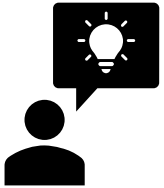
Large language models (LLMs)

LLM deployment

NLP applications for social good and NLP ethics

A brief history of NLP

1940
(after world war II)



I wish a machine could translate…

But the machine can't tell correct grammar

"Colorless green ideas sleep furiously" =/= Furiously sleep ideas green colorless

1958 – Noam Chomsky

1957 – 1970s

NLP
1) Symbolic
2) Stochastic

A brief history of NLP (contd..)

After 1970

- **Logic-based paradigms**

-Prolog

- **SHRDLU -> NLU**
- **Discourse modelling**

```
Person: PUT THE LITTLEST PYRAMID ON TOP OF
IT.
Computer: OK. (does it)
Person: DOES THE SHORTEST THING THE
TALLEST PYRAMID'S SUPPORT SUPPORTS SUPPORT
ANYTHING GREEN?
Computer: YES, THE GREEN PYRAMID.
Person: WHAT COLOR IS IT?
Computer: BY "IT", I ASSUME YOU MEAN THE
SHORTEST THING THE TALLEST PYRAMID'S
SUPPORT SUPPORTS.
Computer: RED
```

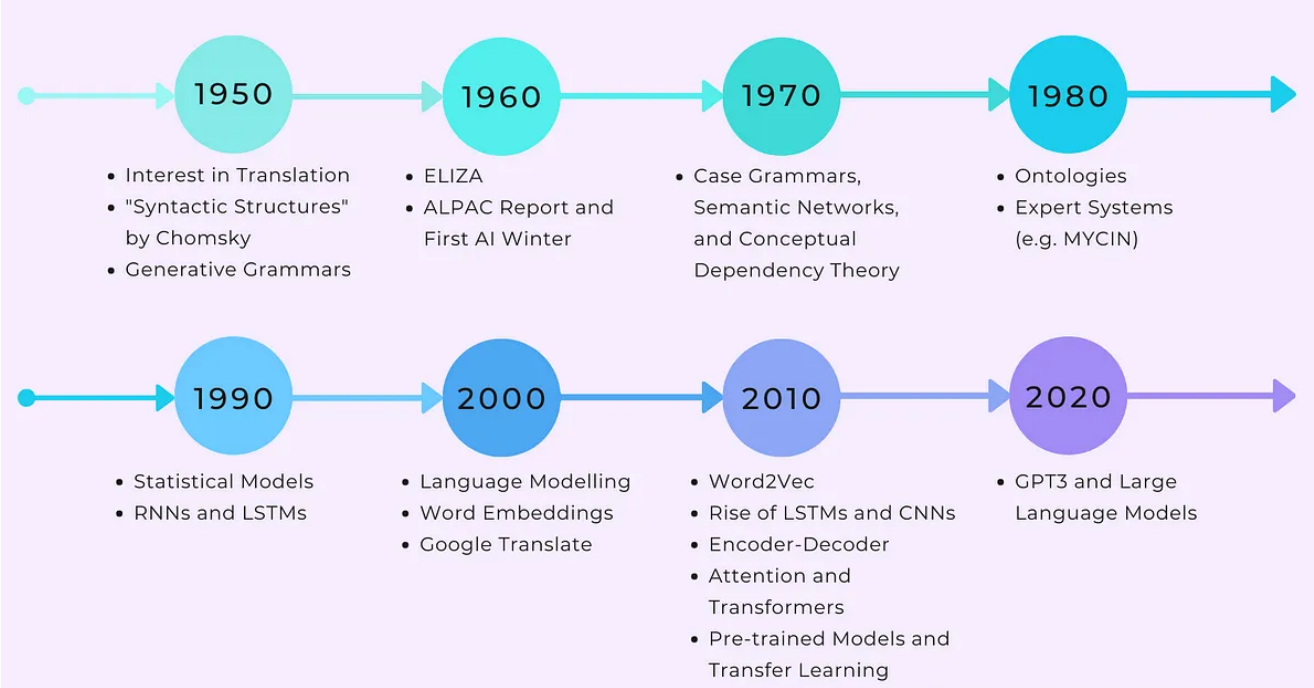
1983 - 1993

- **Empiricism**
- **Probabilistic and statistical methods**

2000s onwards

- **Annotated corpora, resources**
- **Information extraction, generation**

Key advances



https://medium.com/nlplanet/a-brief-timeline-of-nlp-bc45b640f07d

Levels of processing and understanding

To study NLP, it is important to understand core concepts from linguistics about how language is structured.

NLP applications need knowledge of the building blocks, starting from the basic sounds of language (phonemes) to texts with some meaningful expressions (context) for pragmatic usage.

Phonetical and Phonological	understanding the patterns present in the sound and speeches
Morphological	understanding the structure of the words and the systematic relations
Lexical	understanding the part of speech
Syntactic	understanding the structure of the sentence
Semantic	understanding the literal meaning of the words, phrases, and sentences
Discourse	understanding units larger than a single sentence
Pragmatic	real-world knowledge to understand the bigger context of the sentence

https://saxon.ai/blogs/natural-language-processing-in-ai/

Panel on the **Place of Linguistics and Symbolic Structures** NAACL 2022: https://underline.io/events/325/sessions/11355/lecture/56309-the-place-of-linguistics-and-symbolic-structures

An industry employee who starts at around 45:50 on the recording says:

“Very often an ensemble of simple classical methods like SVMs and HMMs and the like -- for many problems, that can be almost as good as a large pre-trained LM. Or even smaller pre-trained LMs -- even an ensemble of those can be ***almost as good as a large one and much easier to deploy*** and much less costly to maintain.

So I want to encourage you, ***please continue to teach your students classical methods***. They are important. Very important if they're going to go out into industry, especially if they're not working at a place like Google or Meta or Amazon (where you have a ton of resources) and the cost of your solution matters.

I interview people and sometimes I come across candidates who don't know anything but neural networks, and they use them to solve every problem whether it's appropriate or not. And ***so really the classical methods matter.***

And also teach ***them basic linguistic knowledge because it makes it much easier for them to communicate about the problem***. It doesn't necessarily have to be to the depth we used to go to, but at least some basic knowledge is useful. And also useful for engineering features.”

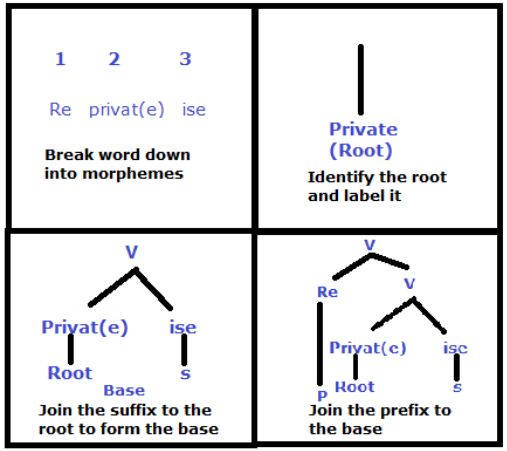
Terminology

Morphology: The study of words, how they are formed, and their relationship to other words

Lexicon: Vocabulary of a language/ dictionary

The bird is a *crane*

They used a *crane* to lift the block



https://all-about-linguistics.group.shef.ac.uk/branches-of-linguistics/morphology/what-is-morphology/

Tokenization

- Fundamental step in NLP that breaks down text into smaller units called tokens.
- The tokens can be words, characters, or sub-words, depending on the specific application and language.
- Typically split using whitespace characters discarding any spaces and punctuations

E.g. “The cat says meow”

(Word) Tokens: The cat says meow

Other kinds of tokenization

Character tokens: Text is split into individual characters.

E.g. "Hello" → ["H", "e", "l", "l", "o"]

Sub-word Tokens: Splits words into meaningful subunits.

E.g. "unhappiness" → ["un", "happiness"] or ["un", "happy", "ness"]

Sentence tokenization (also known as sentence segmentation) is also applied for NLP tasks at a sentence level, such as machine translation, text summarization, or sentiment analysis.

Not as simple as splitting text by spaces….

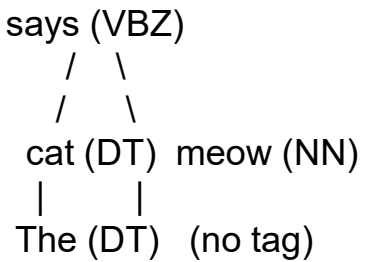
Tokenization challenges

1. Handling Punctuation: Should punctuation be separate tokens or part of words?
2. Contractions: How to handle words like "don't" or "I'm"?
3. Compound Words: Should "ice cream" be one token or two?
4. Special Characters: How to deal with emojis, hashtags, or URLs?
5. Language-Specific Issues: Different languages may require different approaches (e.g., Chinese doesn't use spaces between words)

Syntax: Rules for structuring the language

E.g. How are the words related

“He ate the fish” *not equal to* “The fish ate him”



Syntatic parsing (POS) and Dependency trees:

E.g. “The cat says meow”

"DT" stands for "determiner", that comes before a noun to indicate which one is being referred to (such as "the" or "a"). "VBZ" stands for "verb, 3rd person singular present", and "NN" stands for "noun, singular or mass". The tree shows that "says" is the main verb in the sentence, "cat" is the subject, and "meow" is the object. The determiner "The" is attached to "cat" as a dependent.

Universal POS tagset

Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

Parts of Speech (Pos) tagging

- The process of classifying words into their parts of speech and labeling them accordingly is known as part-of-speech tagging, POS tagging, or simply tagging.
- It helps identify the syntactic category or part of speech (e.g., noun, verb, adjective) of each word
- Parts of speech are also known as word classes or lexical categories.

E.g. Jim bought 300 shares of Acme Corp in 2006

Which POS can you find in this sentence?

POS tagging

Jim bought 300 shares of Acme Corp in 2006

Output from nltk: [('Jim', 'NNP'), ('bought', 'VBD'), ('300', 'CD'), ('shares', 'NNS'), ('of', 'IN'), ('Acme', 'NNP'), ('Corp', 'NNP'), ('in', 'IN'), ('2006', 'CD')]

Let's break down each word with its POS tag!

Jim (NNP):
NNP stands for Proper Noun, Singular
This correctly identifies "Jim" as a name (proper noun)

bought (VBD):
VBD stands for Verb, Past Tense
This correctly identifies "bought" as the past tense form of the verb "buy"

Input text: Jim bought 300 shares of Acme Corp in 2006

Output from nltk: [('Jim', 'NNP'), ('bought', 'VBD'), ('300', 'CD'), ('shares', 'NNS'), ('of', 'IN'), ('Acme', 'NNP'), ('Corp', 'NNP'), ('in', 'IN'), ('2006', 'CD')]

- 300 (CD): CD stands for Cardinal Number. This correctly identifies "300" as a number
- shares (NNS): NNS stands for Noun, Plural. This correctly identifies "shares" as a plural noun
- of (IN): IN stands for Preposition. This correctly identifies "of" as a preposition
- Acme (NNP): NNP stands for Proper Noun, Singular. This correctly identifies "Acme" as part of a company name (proper noun)
- Corp (NNP): NNP stands for Proper Noun, Singular. This correctly identifies "Corp" as part of a company name (proper noun)
- in (IN): IN stands for Preposition. This correctly identifies "in" as a preposition
- 2006 (CD): CD stands for Cardinal Number. This correctly identifies "2006" as a number (year)

POS for non-English languages

Bangla: ক ুঁড়েঘরগ ুলি/'NN' আক ার/'NN' ব াংল ার/'NNP' ব া/'CC' ভ ারতব র্ষে/'NNP' ?/None
ন য/'JJ' ?/None এঞ চলরে/'NN' প রচল তি/'JJ' ক ুঁড়ে/'NN' ঘর/'NN' নয়/'VM' ক্ি/'SYM'
Hindi: प ाकिस्त ान/'NNP' की/'PREP' प ूर्व/'JJ' प्रध ानम ंत्री/'NN' बेनजीर/'NNPC' भ ुट्टो/'NNP'
पर/'PREP' लगे/'VFM' भ्रष्ट ाच ार/'NN' के/'PREP' आर ोपों/'NN' के/'PREP' खिल ाफ/'PREP' भ ुट्टो/'NNP'
द्व ार ा/'PREP' द ायर/'NVB' की/'VFM' गई/'VAUX' य ाचिक ा/'NN' की/'PREP' स ुनव ाई/'NN'
म ंगलव ार/'NN' को/'PREP' वकीलों/'NN' की/'PREP' हड़त ाल/'NN' के/'PREP' क ारण/'PREP'
स्थ गित/'JVB' कर/'VFM' दी/'VAUX' गई/'VAUX' ।/'PUNC'
Marathi: ग्र ामीण/'JJ' जि ल्ह ाध्यक्ष/'NN' ब ाळ स ाहेब/'NNPC' भ ोसले/'NNP' य ांच्य ा/'PRP' ?/None
ध्यक्षतेख ाली/'NN' पक्ष ाची/'NN' आज/'NN' ब ै?/None क/'NN' झ ाली/'VM' ./'SYM'
Telugu: ఖ్చరుల/'NN' ను ంచి/'PREP' వచ్చిన/'VJJ' ప త్రాల/'NN' ను/'PREP' సాక్ష్య ా/'NN'

***Figure 2.1**: POS-Tagged Data from Four Indian Languages: Bangla, Hindi, Marathi, and Telugu*

Source: https://www.nltk.org/book/ch05.html

POS tags are useful for many downstream tasks…

- Lemmatization (select correct lemma given a word and its POS tag)
- Word Disambiguation ("I saw a bear." vs "Bear with me!")
- Named Entity Recognition (typically comprised of nouns and proper nouns)
- Information Extractions (e.g., verbs indicate relations between entities)
- Parsing (information of word classes useful before creating parse trees)
- Speech synthesis/recognition (e.g., noun "DIScount" vs. verb "disCOUNT")
- Authorship Attribution (e.g., relative frequencies of nouns, verbs, adjectives, etc.)
- Machine Translation (e.g., reordering of adjectives and nouns)

Let’s see a practical usage of pos tagging!

```
# Sample restaurant reviews
reviews = [
    "The food was absolutely delicious and the service was impeccable.",
    "Terrible experience. Rude staff and mediocre food.",
    "Average restaurant with okay food. Nothing special.",
    "The ambiance was cozy but the dishes were bland and overpriced.",
    "Fantastic flavors and friendly staff. Highly recommended!"
]
```

Can you spot the pos that seems to strongly contribute to the sentiment?

```
                                              Review                                Adjectives
Overall Sentiment  Adjective Sentiment
0  The food was absolutely delicious and the serv...           absolutely, delicious, impeccable
0.8176              0.9136
1  Terrible experience. Rude staff and mediocre food.                 terrible, rude, mediocre
-0.7351             -0.7579
2  Average restaurant with okay food. Nothing spe...                      average, okay,
special           0.0772              0.4404
3  The ambiance was cozy but the dishes were blan...                   cozy, bland, overpriced
-0.1779             -0.1027
4  Fantastic flavors and friendly staff. Highly r...                    fantastic, friendly
0.8513              0.7579

Correlation between Overall and Adjective-based Sentiment: 0.97
```

Named entity recognition (NER)

- The task of identifying and categorizing key information (entities) in text. Also known as entity chunking/ identification
- Pre-defined categories include person names, organizations, locations, quantities, etc.

$[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[\text{2006}]_{\text{Time}}$

Other NLP terminology

Semantics:

- How we infer meaning
- Can be applied to entire texts or to single words

The dog would stand and *bark*

The tree *bark* was dry

(You can see in the above example how POS can play a role in determining meaning)

Other NLP terminology

Discourse: Multiple sentences connected in a coherent and meaningful way

Person A: "Hey, did you hear about that new Italian restaurant that just opened up downtown?"

Person B: "No, I haven't. Have you tried it yet?"

Regular exercise has numerous benefits for physical and mental health. *It* can help prevent chronic diseases such as heart disease, diabetes, and cancer, as well as improve mood, sleep, and cognitive function.

Other NLP terminology

Corpus/ corpora

- Large collection of texts used for linguistic analysis or language modelling
- Can include a variety of text types, such as books, articles, web pages, social media posts, and spoken recordings
- Can be in various formats, such as plain text, HTML, XML, or JSON
- We can create our own corpus by compiling a collection of texts that are relevant to our research or analysis.

E.g. British National Corpus (BNC) with over 100 million words:
http://www.natcorp.ox.ac.uk/

Why POS, syntactic parsing (dependency trees)?

These form the foundation for many applications of NLP to perform semantic tasks (deriving meaning) and pragmatic tasks!

Word Sense Disambiguation (WSD)

Words in natural language usually have a fair number of different possible meanings.

- *Ellen has a strong interest in computational linguistics*
- *Ellen pays a large amount of interest on her credit card.*

For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Other NLP tasks

Anaphora Resolution/Co-Reference

• Determine which phrases in a document refer to the same underlying entity.

John put the carrot on the plate and ate *it*.

Bush started the war in Iraq. But *the president* needed the consent of Congress.

Some cases require difficult reasoning. E.g.:

Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Regular expressions

- A **regular expression** (or RE) is used to match strings of text such as particular characters, words, or patterns of characters. These come in quite handy for many NLP operations.

E.g. Extracting name from an email ID, Title from a name, or components of an address

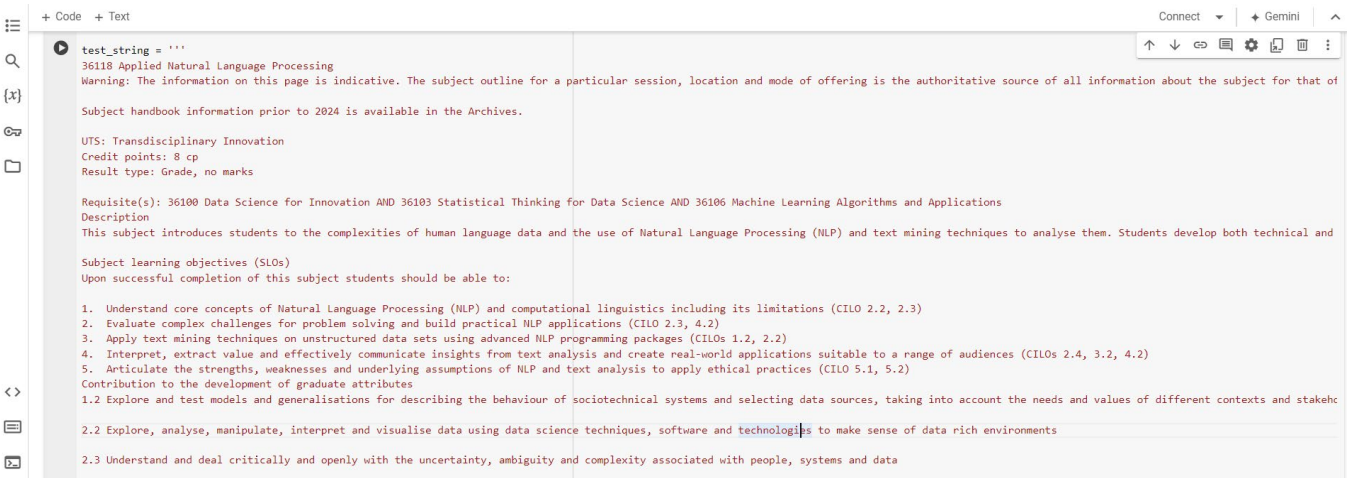
- Most languages support Regex, and have slightly different syntaxes (We'll use Python for regex)
- Useful links to play with RegEx: https://regexr.com/, https://regexone.com/, https://regex101.com/, https://www.nltk.org/book/ch03.html

RegEx examples

```
## Search for pattern 'iii' in string 'piiig'.
## All of the pattern must match, but it may appear anywhere.
## On success, match.group() is matched text.
match = re.search(r'iii', 'piiig') # found, match.group() == "iii"
match = re.search(r'igs', 'piiig') # not found, match == None

## . = any char but \n
match = re.search(r'..g', 'piiig') # found, match.group() == "iig"

## \d = digit char, \w = word char
match = re.search(r'\d\d\d', 'p123g') # found, match.group() == "123"
match = re.search(r'\w\w\w', '@@abcd!!') # found, match.group() == "abc"
```

https://developers.google.com/edu/python/regular-expressions

test_string




```
#Extracting string with integers with at least 4 digits and at most 7 digits
pattern = re.compile(r'\d{4,7}(?!\d)')
find_with_regex(pattern, test_string)
```

```
All matching texts:
['36118', '2024', '36100', '36103', '36106']
```

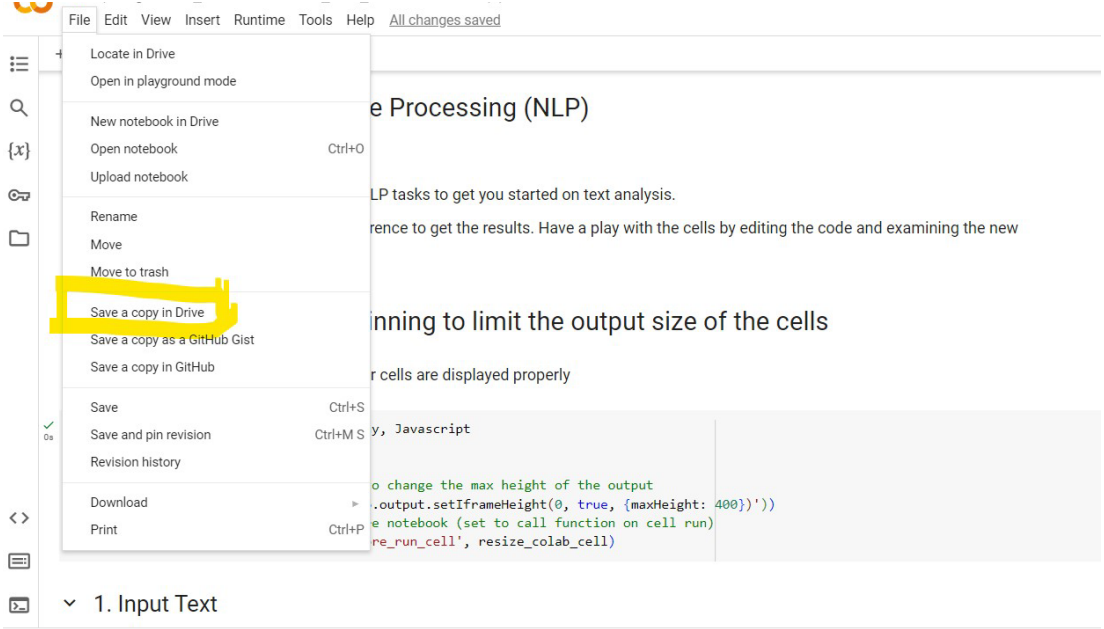
Learn more about Linguistics

https://www.coursera.org/learn/human-language

Tutorial: NLP Basics (1)

Link to notebook: tinyurl.com/ANLPcolab1

- Defining text as inputs
- Displaying readable text
- Tokenization
- POS tagging
- Named entity recognition
- Dependency trees



These have an impact on downstream NLP tasks!

Break

Text mining (aka Text analysis/ Text analytics)

Text mining is the use of automated NLP techniques to **extract valuable insights from unstructured text data**

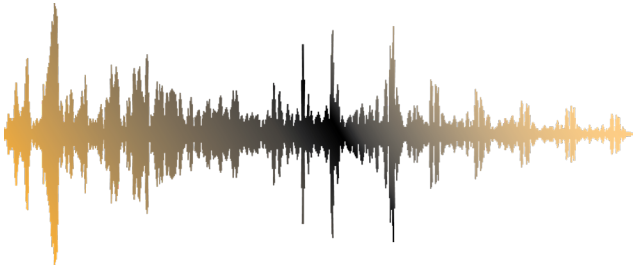
Structured data

Name	FName	City	Age	Salary
Smith	John	3	35	$280
Doe	Jane	1	28	$325
Brown	Scott	3	41	$265
Howard	Shemp	4	48	$359
Taylor	Tom	2	22	$250

Unstructured text

“Joe Bloggs readily admits when he doesn’t know the answer to a particular query. He outlines the steps that he will take to resolve a problem. However, he has difficulty saying no or tactfully telling customers that they must wait their turn. He also tends to refers too many queries to management for final resolution.”

Unstructured data



“Joe Bloggs readily admits when he doesn’t know the answer to a particular query. He outlines the steps that he will take to resolve a problem. However, he has difficulty saying no or tactfully telling customers that they must wait their turn. He also tends to refers too many queries to management for final resolution.”

Basic text analyses using NLP

- Word counts
- Word frequency (lists of words and their frequencies)
- N-grams (common two-, three-, etc.- word phrases)

Word counts

- Word frequency (lists of words and their frequencies) are simple but effective tools to derive insights from the lengths of texts

(Read: Word counts are amazing by Ted Underwood)

- Usefulness often depends on the questions you ask on the data. For example, questions on the length of newspaper articles can look like:

  Have article lengths decreased to suit the reducing attention span of readers?

That's a great data story!
(more in Week 2)

http://tedunderwood.com/2013/02/20/wordcounts-are-amazing/

Calculate word count

```
[ ]  #To make it easier to reuse in the future, we can create a function that returns word count
     def word_count(text):
         wc = len(text.split())
         return wc
```

Now now we can apply the word_count function to our text variable to create a new variable with the number of words in the news article text.

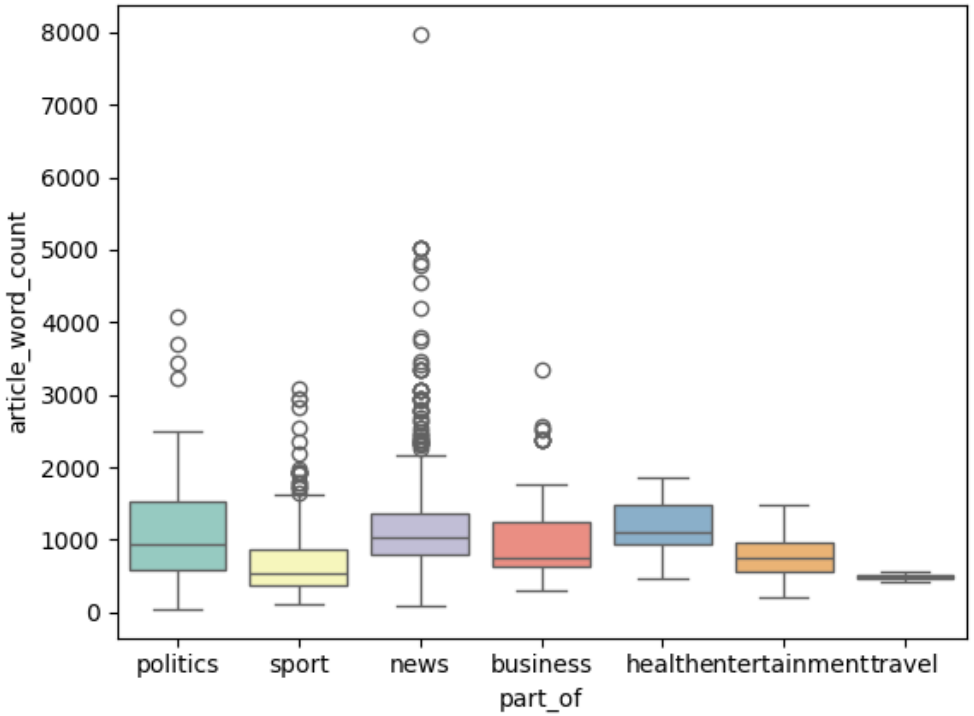
```
[ ]  newsdf['article_word_count'] = newsdf['text'].apply(word_count)
```

We can use describe, hist, and scatter functions to provide some information on the length of articles in our dataset

```
[ ]  newsdf['article_word_count'].describe()
```

	article_word_count
count	1708.000000
mean	1026.977752
std	657.996654
min	48.000000
25%	589.750000
50%	909.000000

- Which category of articles is usually the longest?

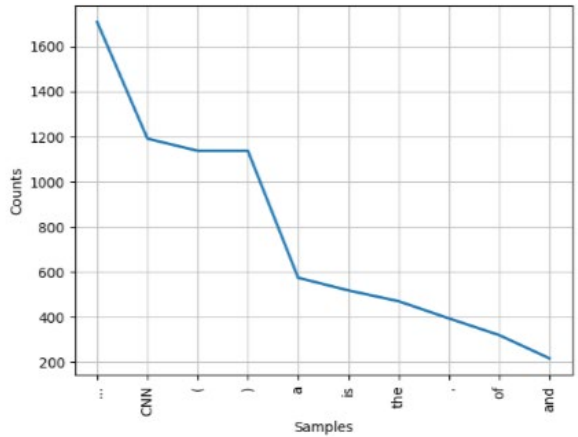


What about the most frequently occurring words?

Calculate frequencies to determine the most common words in the corpus

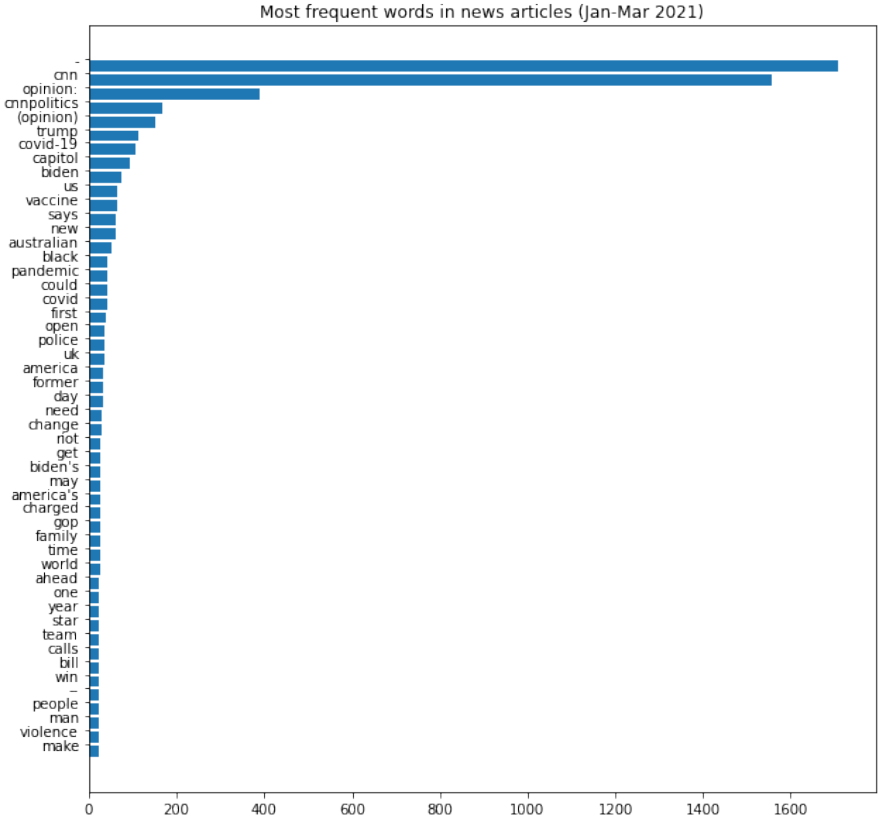
```
[21] # converting series to string
     article_text = newsdf['text'].to_string()

     #create word tokens
     tokenized_words=word_tokenize(article_text)
```

```
[22] all_words=nltk.FreqDist(tokenized_words)
     all_words.plot(10);
     print(all_words.most_common(20))
```




```
[('...', 1707), ('CNN', 1190), ('(', 1136), (')', 1136), ('a', 573), ('is', 517), ('the', 469), (',', 393), ('of', 319), ('and', 216),
```

Do you notice anything odd here?


Most frequent words in news articles (Jan-Mar 2021)
-
cnn
opinion:
cnnpolitics
(opinion)
trump
covid-19
capitol
biden
us
vaccine
says
new
australian
black
pandemic
could
covid
first
open
police
uk
america
former
day
need
change
riot
get
biden's
may
america's
charged
gop
family
time
world
ahead
one
year
star
team
calls
bill
win
–
people
man
violence
make
0
200
400
600
800
1000
1200
1400
1600

This is why we clean up the text first

Let’s do some basic cleaning – removing punctuations and stop words before we try again!

Stopwords removal

- Stopwords do not usually have meaningful information but appear frequently in text.
- Examples in English include:
  - a, an, the
  - is, are, was, were
  - in, on, at, to
  - and, or, but
  - I, you, he, she, it
- Libraries have in-built stop word lists, but we can create our custom list to add stopwords specific to our application!

Let’s see an example

Original sentence: "The quick brown fox jumps over the lazy dog."

After stopword removal: "quick brown fox jumps lazy dog"

Is the core meaning still preserved without stop words?

Mostly, Yes!

Removing punctuations

```
sentence1 = "I love this movie! It's amazing."
```

```
sentence2 = "The plot was confusing... or was it? I'm not sure."
```

```
Tokens with punctuation:
['i', 'love', 'this', 'movie', '!', 'it', "'s", 'amazing', '.']

Tokens without punctuation:
['i', 'love', 'this', 'movie', 'its', 'amazing']
```

```
Tokens with punctuation:
['the', 'plot', 'was', 'confusing', '...', 'or', 'was', 'it', '?', 'i', "'m", 'not', 'sure', '.']

Tokens without punctuation:
['the', 'plot', 'was', 'confusing', 'or', 'was', 'it', 'im', 'not', 'sure']
```

What differences do you see in tokenization results?

Why remove punctuation

- **Consistency:** "It's" and "I'm" are transformed into single tokens ("its" and "im"), which can help in treating contractions consistently.

- **Reduced noise:** Punctuation marks like "!", "?", and "..." are removed, which might help in focusing on the actual words for tasks like sentiment analysis or topic modeling.

- **Simplified text:** The removal of punctuation can make the text easier to process for certain NLP tasks, potentially improving performance.

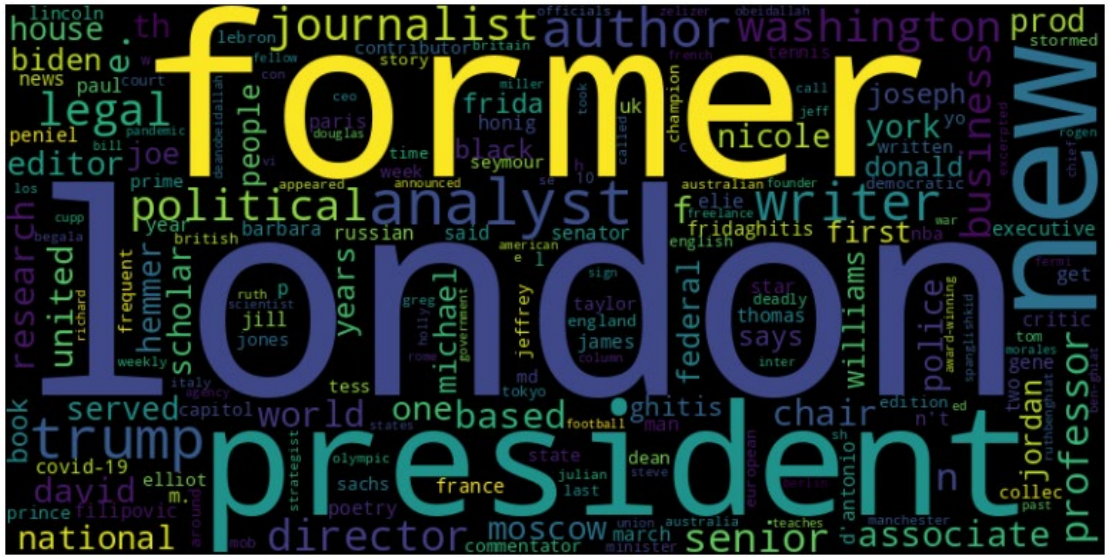
Cleaning text

```
[24] # converting article text to lowercase as Python is case-sensitive
     article_text_lower = article_text.lower()

     #create word tokens
     tokenized_words=word_tokenize(article_text_lower)

     #Set up stop words for removal
     nltk.download('stopwords')
     from nltk.corpus import stopwords
     #stopwords
     stop_words=stopwords.words("english")
     print(stop_words)
     #Add custom stopwords to the list
     stop_words.extend(["cnn", "'s", "a", "the"])
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', '
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
[25] #Create a new variable to store filtered tokens
     filtered_tokens=[]
     for w in tokenized_words:
         if w not in stop_words:
              #add all filtered tokens excluding stopwords in this list below
              filtered_tokens.append(w)

     import string
     # punctuations
     punctuations=list(string.punctuation)
     #Add custom punctuations to the list
     punctuations.append("...")  ← Add what is relevant to your data iteratively

     #Create another variable to store all clean tokens
     filtered_tokens_clean=[]
     for i in filtered_tokens:
         if i not in punctuations:
             filtered_tokens_clean.append(i)
```

Updated list of frequent words



(Also, the most popular text visualisation – word clouds!)

Time to dig into the notebook!

tinyurl.com/ANLPcolab1part2

"File" -> "Save a Copy in Drive..."

Take home exercise:
https://colab.research.google.com/drive/1wzZ_IM858GJsnaHDE_PwU0GeC78lr2oF?usp=sharing

More advanced pre-processing steps and analysis will be covered in Week 2!

Take home activities – Week 1

- Complete homework exercise in the notebook provided
- Start experimenting and building your portfolio (start exploring your AT1 data!)
- Complete week 1 activities on the cohort resource board, add interesting resources/ reading:

  tinyurl.com/**anlpspr25miro**
  Password: anlpspring2025

- Perform additional reading

Assessment details

Download the detailed assessment brief that contains all assignment related info (from Canvas)

You should start working on AT1!

General Q & A