Introduction to Natural Language Processing

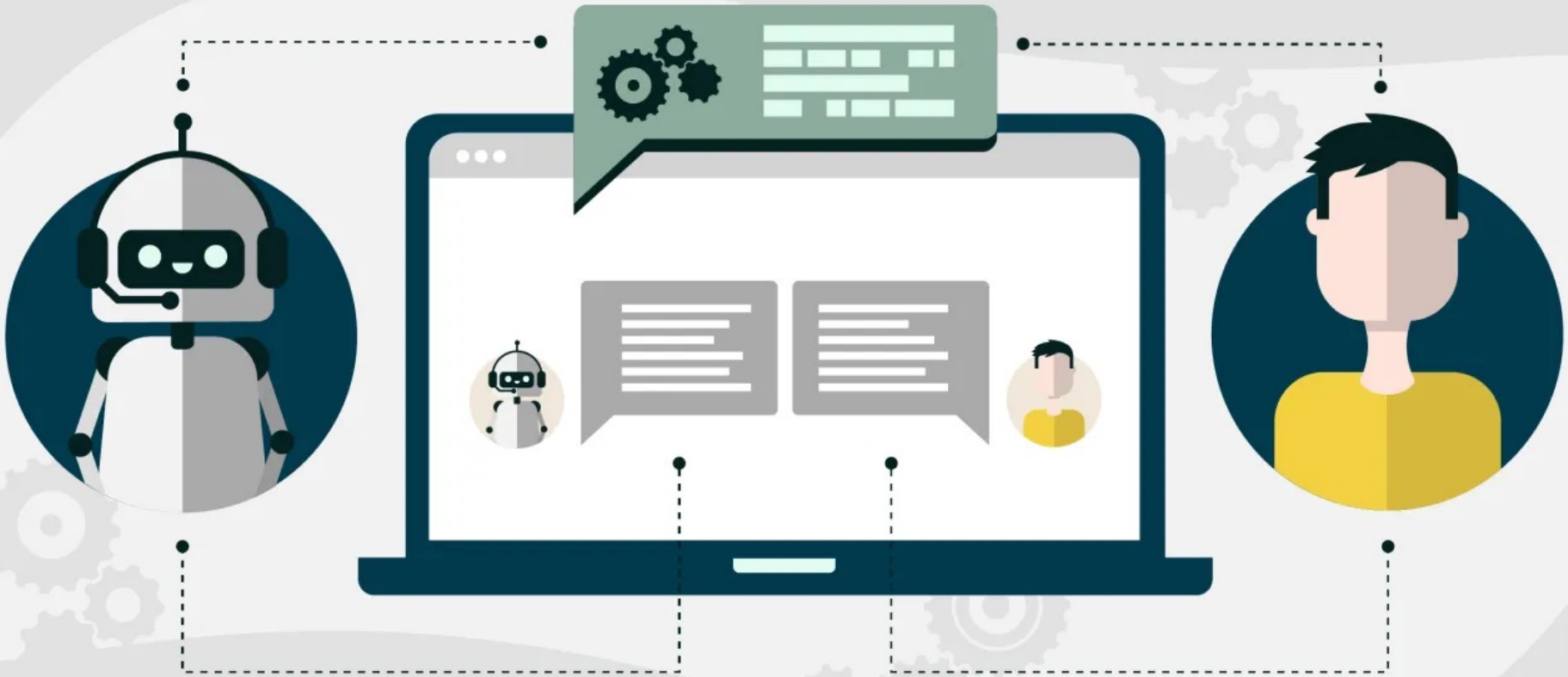
Maria Balaet

Department of Brain Sciences

Imperial College London

@emmsskyyy

NLP in the field of neuroscience



Then ~~someone~~ made Galactica...

Meta AI

a large language model that can store, combine and reason about scientific knowledge
(outperforms the latest GPT-3 by 68.2%)

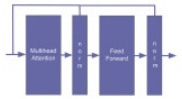
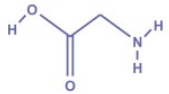
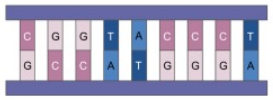
Modality	Entity	Sequence	
Text	Abell 370	`Abell 370 is a cluster...`	
LaTeX	Schwarzschild radius	`r_{s} = \frac{2GM}{c^2}`	$r_s = \frac{2GM}{c^2}$
Code	Transformer	`class Transformer(nn.Module)`	
SMILES	Glycine	`C(C(=O)O)N`	
AA Sequence	Collagen α-1(II) chain	`MIRLGAPQTL..`	
DNA Sequence	Human genome	`CGGTACCCTC..`	

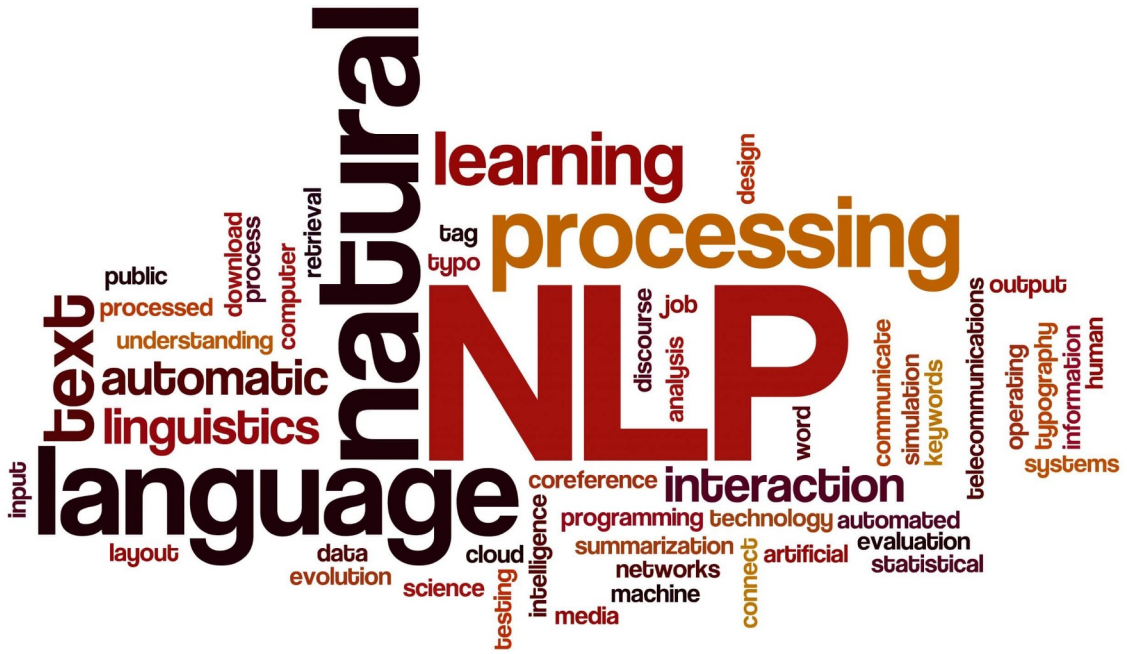
Table 1: Tokenizing Nature. Galactica trains on text sequences that represent scientific phenomena.

Total dataset size = 106 billion tokens			
Data source	Documents	Tokens	Token %
Papers	48 million	88 billion	83.0%
Code	2 million	7 billion	6.9%
Reference Material	8 million	7 billion	6.5%
Knowledge Bases	2 million	2 billion	2.0%
Filtered CommonCrawl	0.9 million	1 billion	1.0%
Prompts	1.3 million	0.4 billion	0.3%
Other	0.02 million	0.2 billion	0.2%

this is an insane amount of data

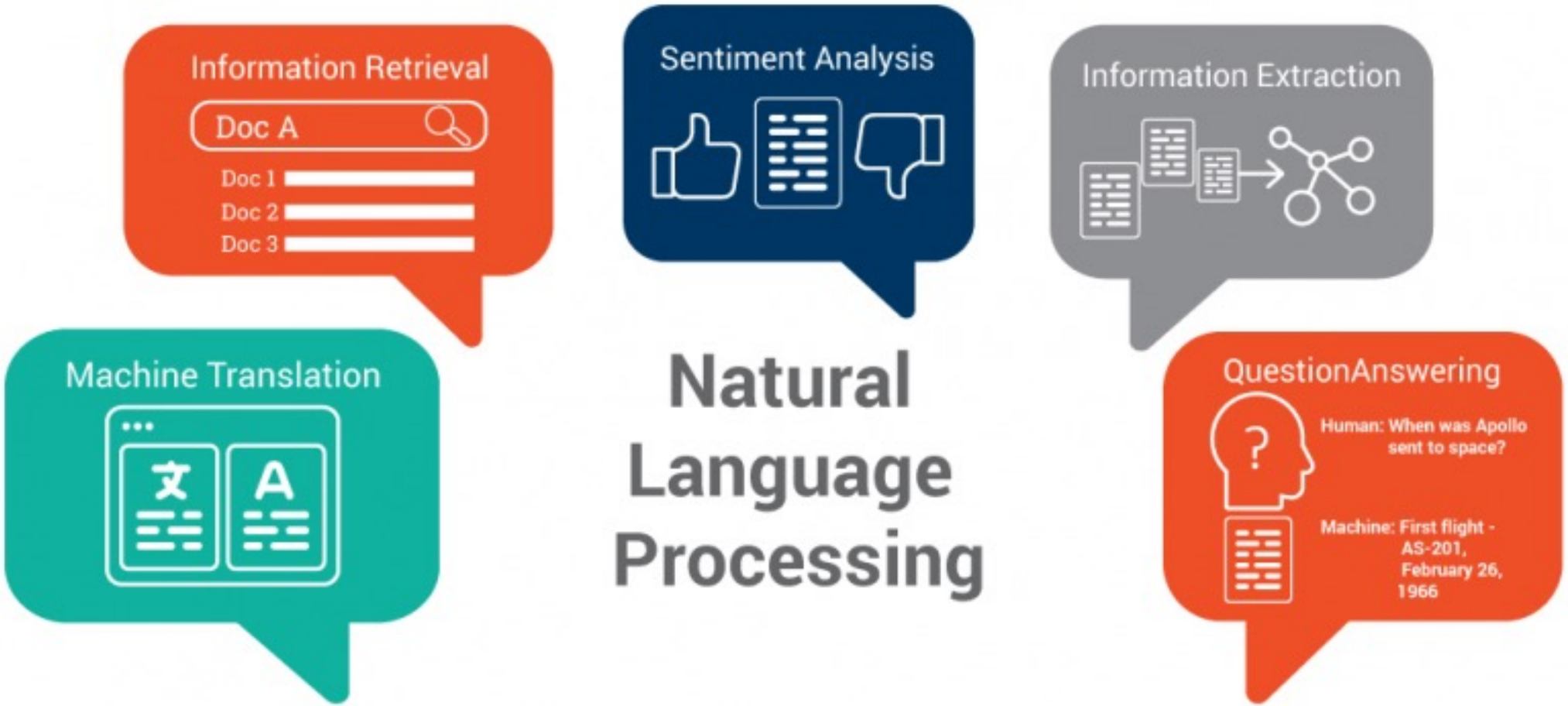
https://arxiv.org/pdf/2211.09085.pdf

What is NLP

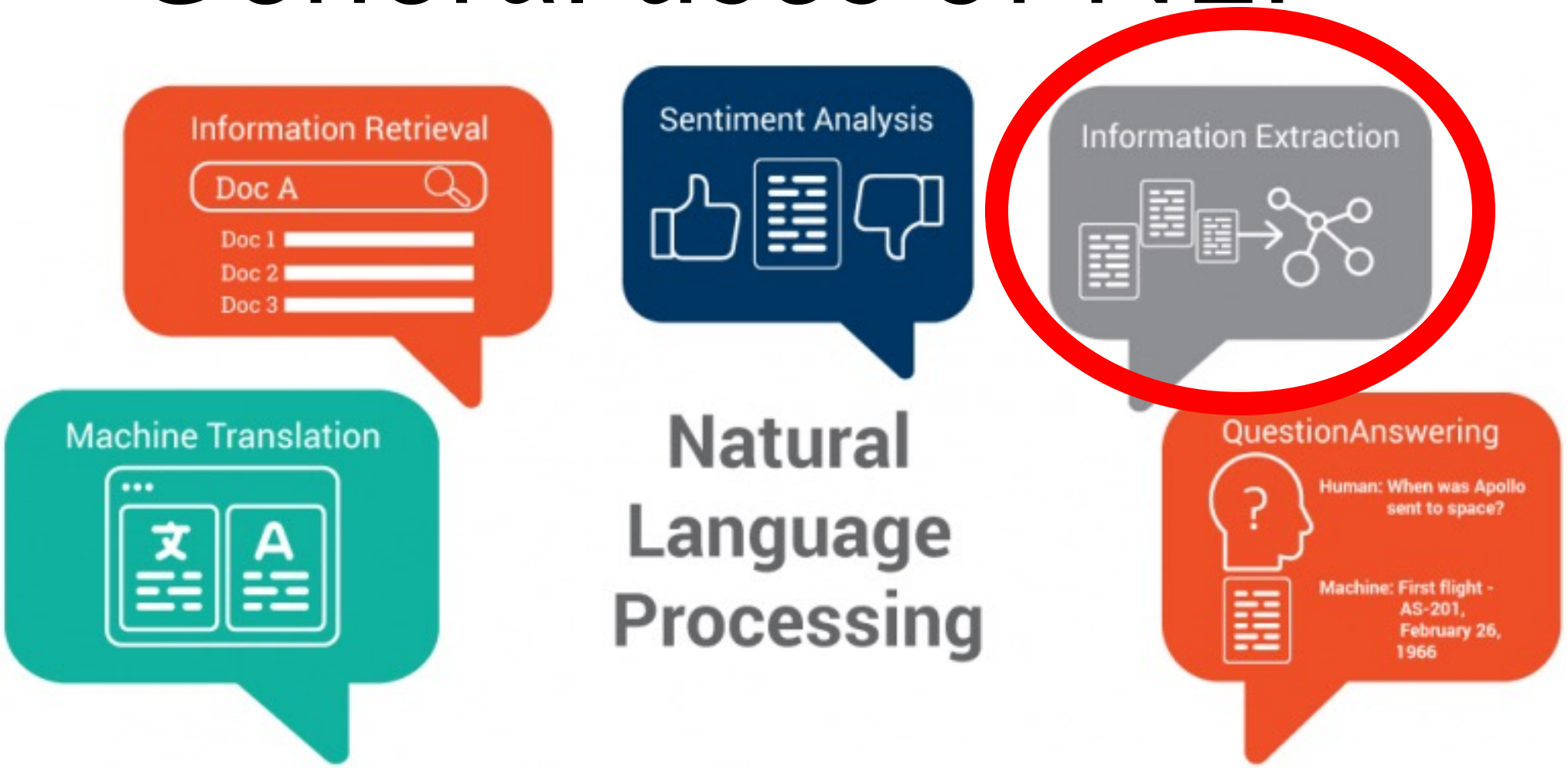


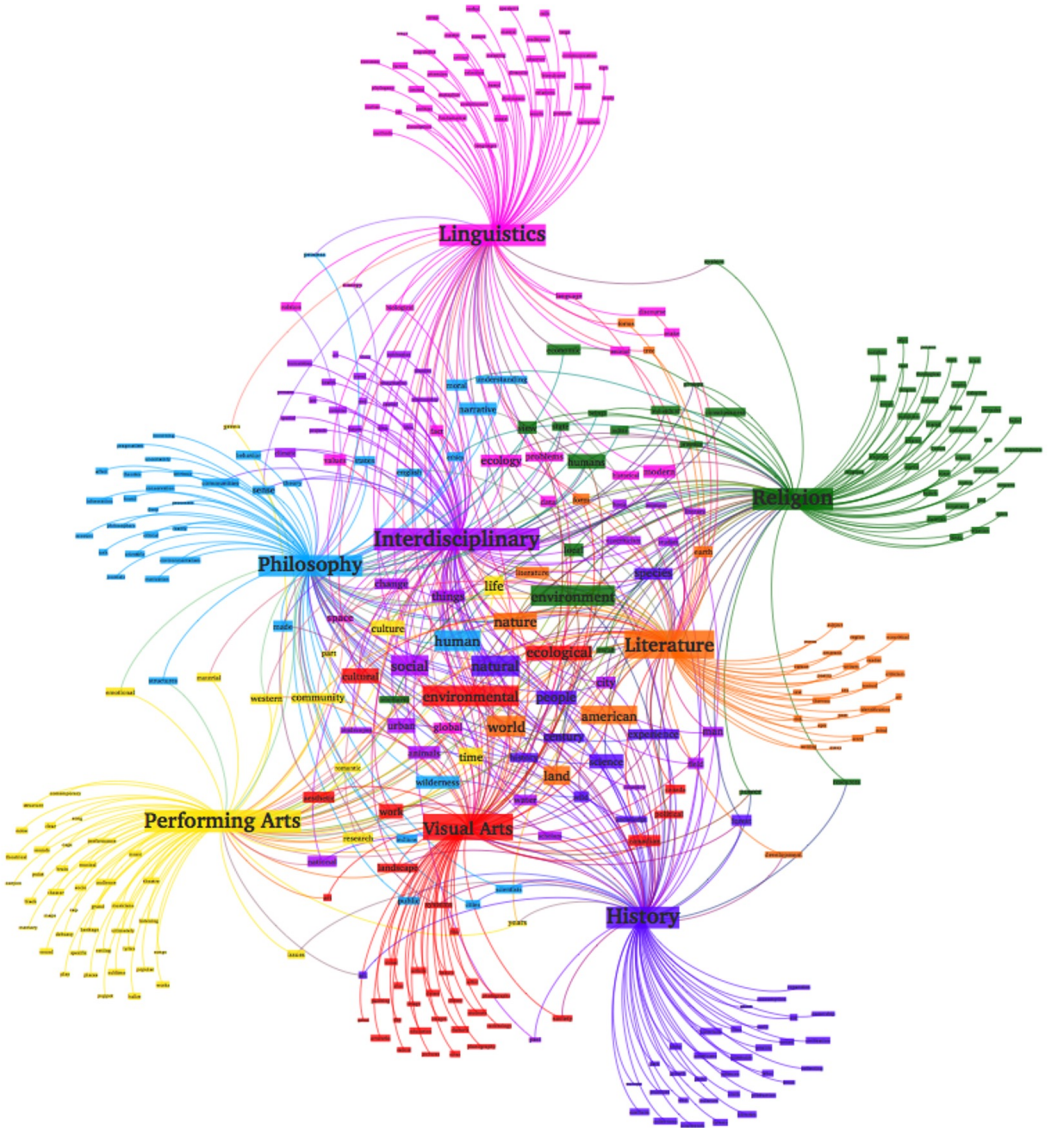
Official definition: *"Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyse large amounts of natural language data."*

General uses of NLP



General uses of NLP





Topic modelling

“a type of statistical model for discovering the abstract "topics" that occur in a collection of documents”

How do we achieve this?

Data preprocessing

text	created_at	preprocessed_text	tokenized_text
Thank you for joining us at the Lincoln Memorial tonight- a very special evening! Together we are going to MAKE AM… https://t.co/OSxa3BamHs	2017-01-20 00:40:51	thank joining us lincoln memorial tonight- special evening! together going make am…	[thank, joining, us, lincoln, memorial, tonight, special, evening, together, going, make, am]
Thank you for a wonderful evening in Washington D.C. #Inauguration https://t.co/a6xpFQTHj5	2017-01-20 04:24:33	thank wonderful evening washington d.c. #inauguration	[thank, wonderful, evening, washington, inauguration]
It all begins today! I will see you at 11:00 A.M. for the swearing-in. THE MOVEMENT CONTINUES - THE WORK BEGINS!	2017-01-20 12:31:53	begins today! see 11:00 a.m. swearing-in. movement continues - work begins!	[begins, today, see, swearing, in, movement, continues, work, begins]
Today we are not merely transferring power from one Administration to another or from one party to another – but we are transferring...	2017-01-20 17:51:25	today merely transferring power one administration another one party another – transferring...	[today, merely, transferring, power, one, administration, another, one, party, another, transferring]
power from Washington D.C. and giving it back to you the American People. #InaugurationDay	2017-01-20 17:51:58	power washington d.c. giving back american people. #inaugurationday	[power, washington, giving, back, american, people, inaugurationday]

Latent Dirichlet Allocation

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \alpha, \beta) = \prod_{i=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\phi; \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi z_{j,t}),$$

LDA is an unsupervised, machine learning, clustering technique that we commonly use for text analysis. It's a type of topic modelling in which words are represented as topics, and documents are represented as a collection of these word topics.

PS: there are other algorithms too, but LDA is the most established one, acting as the field's gold standard

Latent Dirichlet Allocation

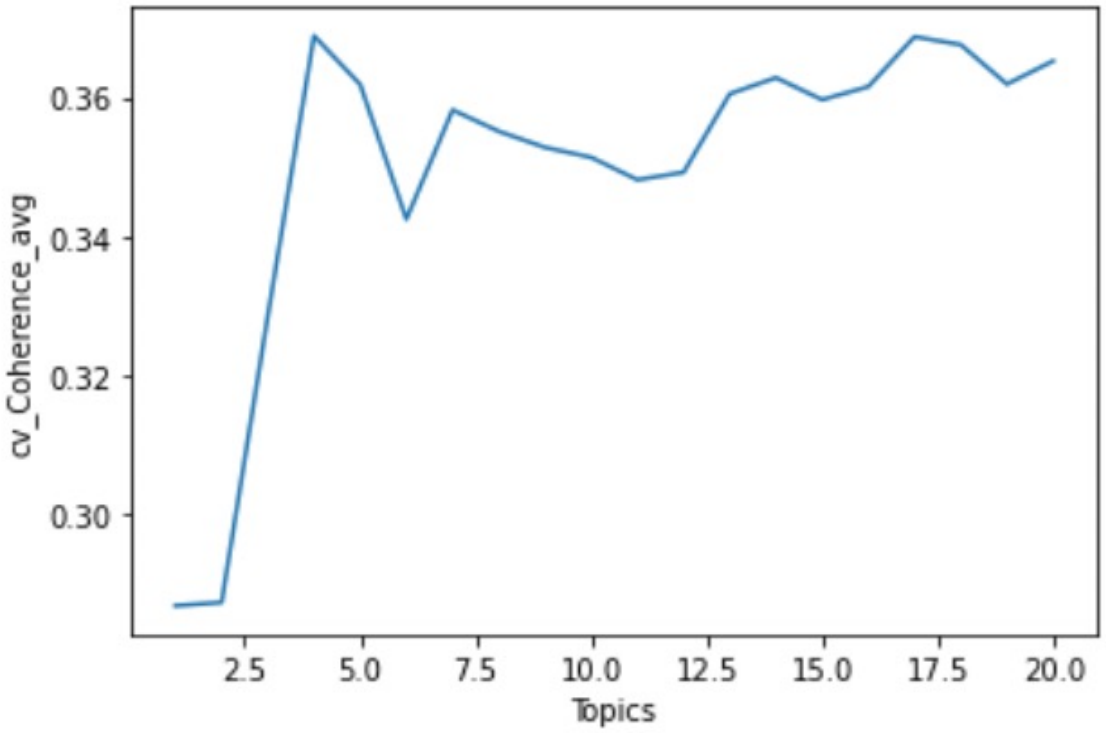
- **Latent**: indicates that the model discovers the ‘yet-to-be-found’ or hidden topics that are common across the documents that people have written.
- **Dirichlet**: indicates the two assumptions of LDA - that both the distribution of topics within a document and the distribution of words within each topic are Dirichlet distributions (which is a type of probability distribution).
- **Allocation** indicates the distribution of topics in the document.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Topic coherence

- measure how interpretable the topics are to humans
- topics are represented as the top N words with the highest probability of belonging to that particular topic
- the coherence score measures how similar these words belonging to a topic are to each other
- A measure of words co-occurence



	Original text	Topic1	Topic2	Topic3	Topic4
30	quite some time ago, i decided to take magic m...	0.000673	0.000676	0.000667	0.997985
31	so this was my first trip on shrooms, yesterda...	0.000729	0.000749	0.000718	0.997803
32	other posts on this sub have been so helpful t...	0.000757	0.000750	0.000754	0.997739
33	so i know this is an lsd page but i wanted to ...	0.000756	0.000778	0.000764	0.997702

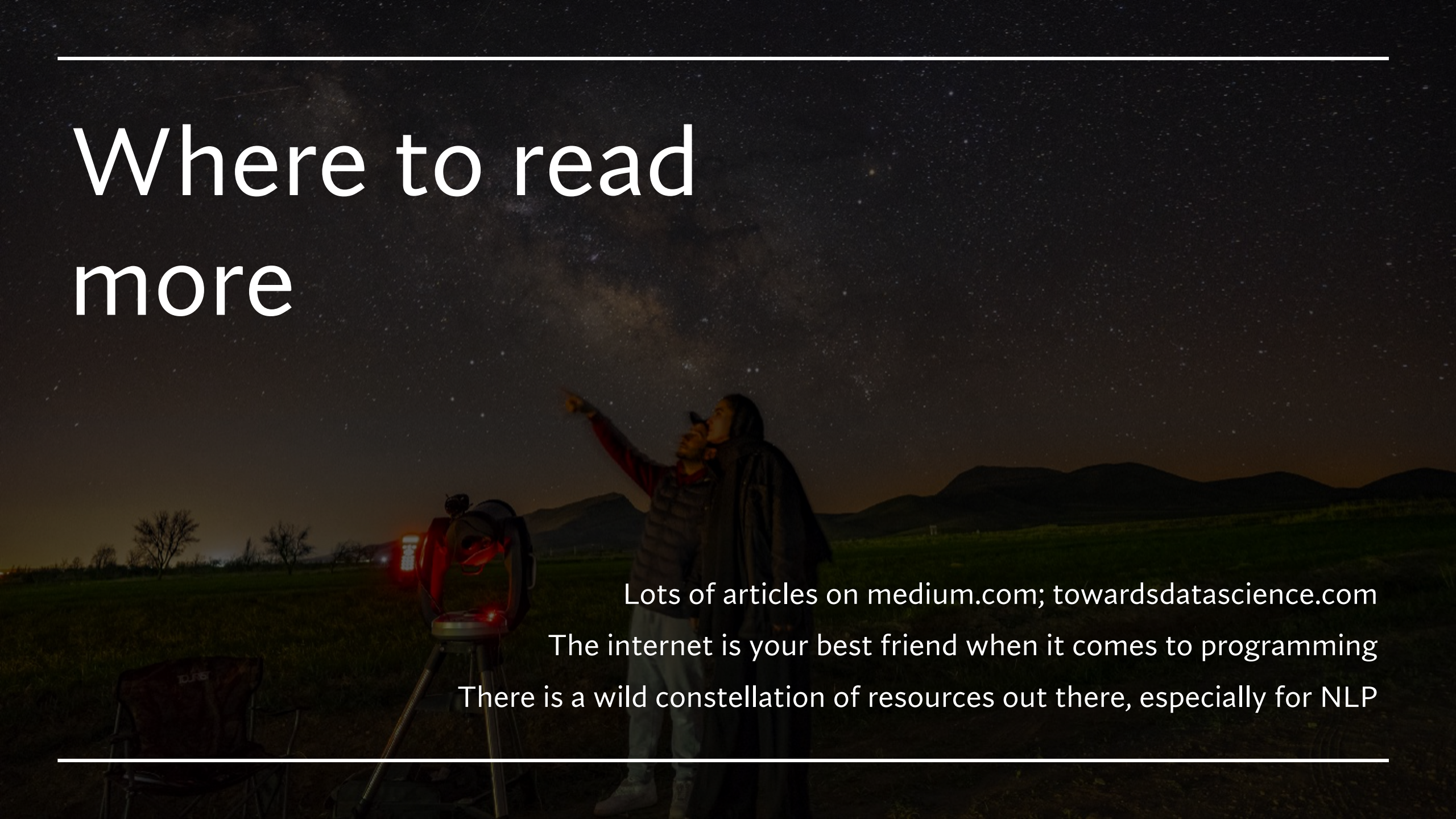
Interpreting topics

TOP WORDS

TOP DOCUMENTS BELONGING TO A CERTAIN TOPIC


Understanding the world in
a data-driven way

Where to read more



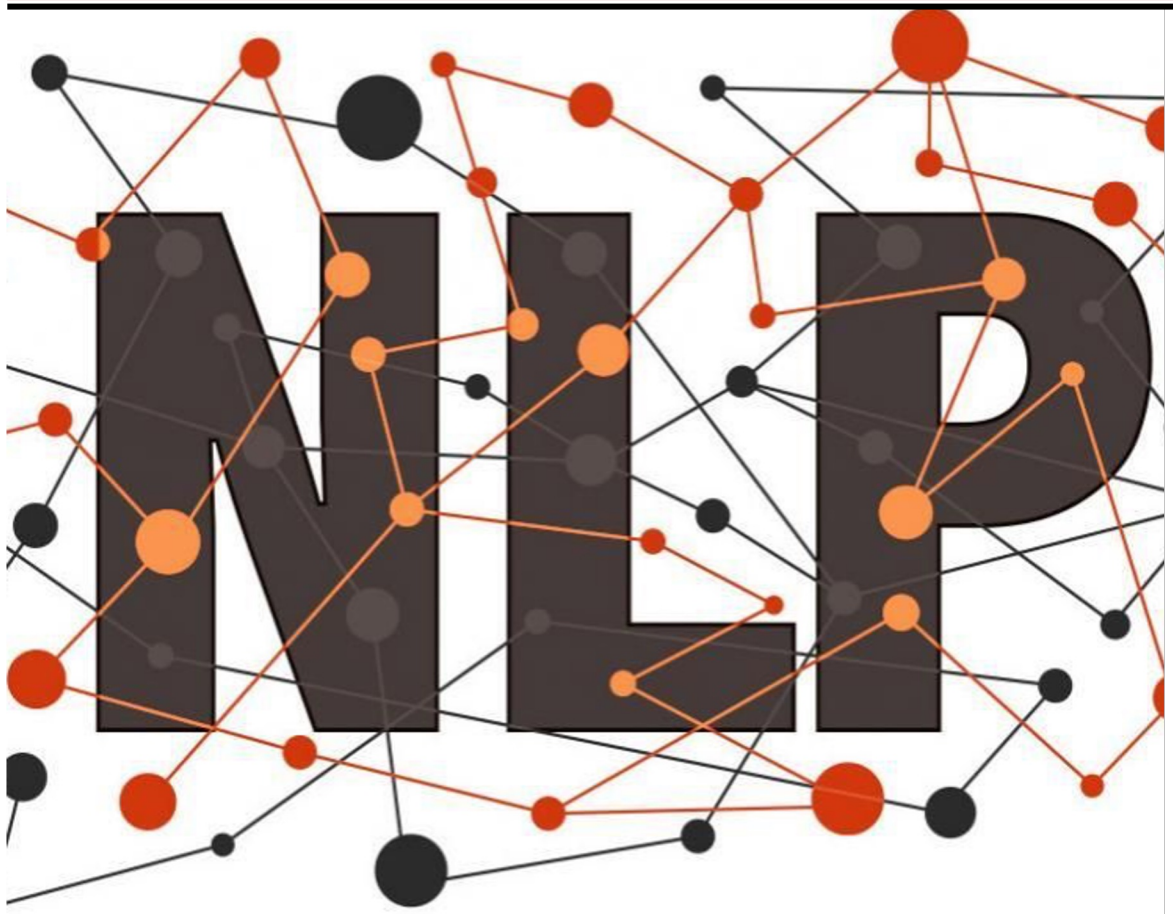
Lots of articles on medium.com; towardsdatascience.com

The internet is your best friend when it comes to programming

There is a wild constellation of resources out there, especially for NLP

I will be offering projects on NLP;
if this is something you are interested
in please talk to myself/Adam

Next steps of the day



- You will be exploring some basic python-based NLP code
- You will learn to preprocess data
- You will identify the optimal number of topics in a dataset
- You will then separate the data into those topics and define them

https://github.com/dragos-gruia/MSc-TranslationalNeuroscience-Module3/blob/main/Day7/Day7.ipynb