

Course Presentation & Logistics

Natural Language Processing

Master in Applied Artificial Intelligence

Jerónimo Arenas-García, Lorena Calvo-Bartolomé, Jesús Cid-Sueiro

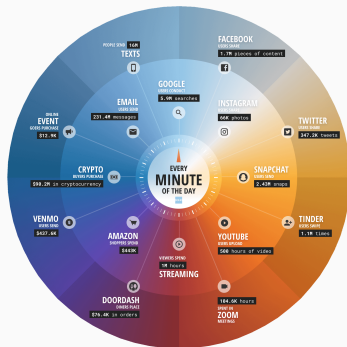
November, 2023

Department of Signal Theory and Communications
Universidad Carlos III de Madrid

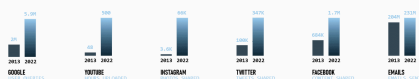
Part I

NLP Introduction

Motivation I



DATA NEVER SLEEPS 1.0 VS. 10.0



GLOBAL INTERNET POPULATION GROWTH



As of April 2022, the internet reaches 63% of the world's population, representing roughly 5 billion people. Of this total, 4.65 billion - over 93 percent - were social media users. According to Statista, the total amount of data predicted to be created, captured, copied and consumed globally in 2022 is 97 zettabytes, a number projected to grow to 181 zettabytes by 2025.

To succeed in an increasingly digital world where the volume of data created keeps accelerating, businesses need the right tools to put that data to work right where work gets done. Domo gives you the power to rapidly unlock value from all your data, regardless of where it lives, and drive actions across your organization that will improve business outcomes. Every click, swipe, share, or like tells a story, and Domo helps you do something powerful with it.

LEARN MORE AT DOMO.COM

SOURCES

Global Media Insights, eMarketer, Hootsuite, eMarketer, Matthew Woodward.co.uk, Web Intelligence, Statista.com, Intel® Business of Apps, Gartner, Statista, Statista, Data Never Sleeps 1.0



<https://www.domo.com/data-never-sleeps>

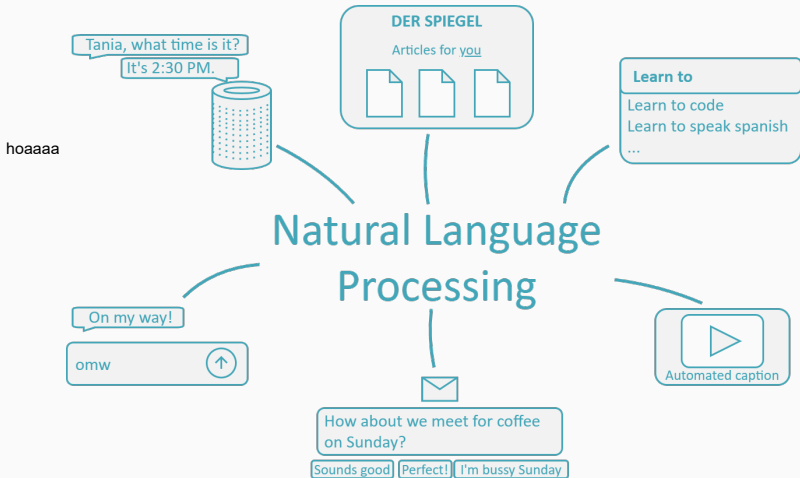


Large amounts of unstructured text data are generated every second

- » **We can no longer use the common approach to understand the text and this is where NLP comes in**

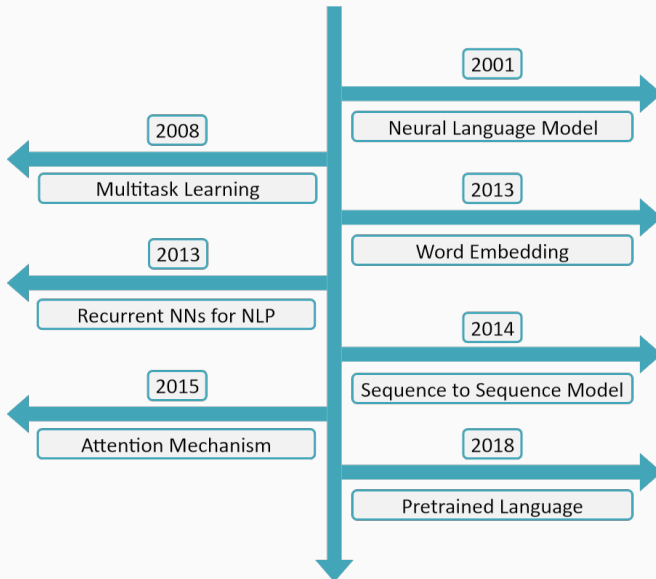
What is NLP?

AI's discipline concerned with giving computers the ability to understand written and spoken human languages in the same manner humans do



Modified from 

A walk-through of recent developments in NLP



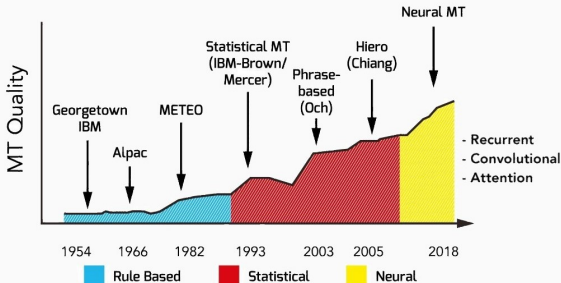
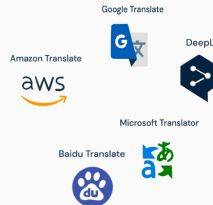
Some NLP applications

Treffen wir uns am Sonntag
zum Kaffee?

Machine translation

¿Nos vemos el domingo para
tomar un café?

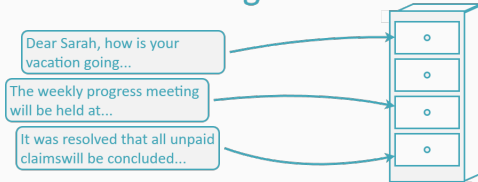
The application of computers to
the task of translating texts from
one natural language to another



Source: IconicTranslation, modified from [\[link\]](#)

Some NLP applications

Text categorization



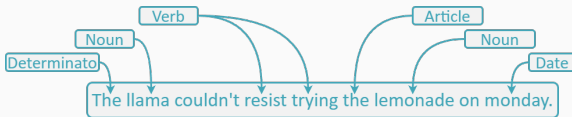
Process of automatically analyzing text and then assigning a set of pre-defined tags or categories based on its content.

- Spam Filtering
- Sentiment Analysis
- Alert detection
- Automating customer walkthroughs in support systems
- ...

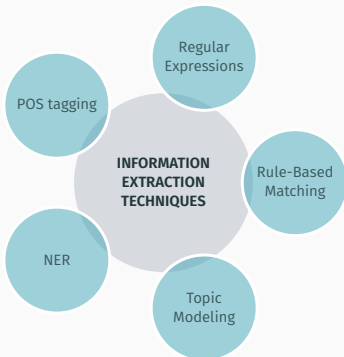


Source: Singapore Institute of Manufacturing Technology (SIMTech). Free demo [link](#)

Information extraction



Process of extracting information from unstructured textual sources to enable finding entities (names, places, events, dates, etc.) that can be applied for a variety of purposes, e.g., identify keywords, classifying text items, etc.



Automate tasks with keyword extraction:

Test with your own text

Elon Musk has shared a photo of the spacesuit designed by SpaceX. This is the second image shared of the new design and the first to feature the spacesuit's full-body look.

Extract Text

Results

TAG	VALUE
KEYWORD	second image
KEYWORD	spacesuit
KEYWORD	body look
KEYWORD	new design
KEYWORD	photo

Free demo [↗](#)

Some NLP applications

Summarization

Pink ponies and purple giraffes roamed the field. Cotton candy grew from the ground as a chocolate river meandered off to the side. What looked like stones in the pasture were actually rock candy. Everything in her dream seemed to be perfect except for the fact that she had no mouth.



A little girl had a dream that she was in a candy factory.

Extractive

From sentences within the text.

Abstractive

Possible to contain words not explicitly present in the text.

Text summarization is the problem of reducing the number of sentences and words of a document without changing its meaning.

Spaces: anaxagoras7 **gauvags-text-summarizer** 1 file 36 [Reviewing](#)

App Files and versions Community

Hugging Face Text Summarizer

Let Hugging Face models summarize texts for you. Note: Shorter articles generate faster summaries. This summarizer uses bart-large-cnn model by Facebook, pegasus by Google and distilbart-cnn-12-6 by Salesforce. You can compare these models against each other on their performances. Sample Text Input is provided!

Text

It went through such rapid contortions that the little bear was forced to change his hold on it so many times he became confused in the darkness, and could not, for the life of him, tell whether he held the sheep right side up, or upside down. But that point was decided for him a moment later by the animal itself, who, with a sudden twist, jabbed its horns so hard into his lowest ribs that he gave a grunt of anger and disgust.

sshleifer/distilbart-cnn-12-6: Summary 6/81

The bear was forced to change his hold on the sheep so many times he became confused in the darkness. He could not, for the life of him, tell whether he held the sheep right side up, or upside down. But that point was decided for him by the animal itself, who jabbed its horns so hard into his ribs that he gave a grunt of anger and disgust.

facebook/bart-large-cnn: Summary

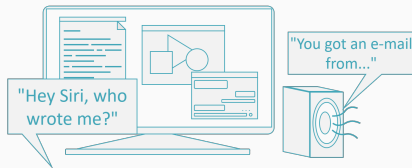
The sheep went through such rapid contortions that the little bear was forced to change his hold on it so many times he became confused in the darkness. He could not, for the life of him, tell whether he held the sheep right side up, or upside down. But that point was decided for him a moment later by the animal itself, who, with a sudden twist, jabbed its horns so hard into his lowest ribs.

google/pegasus-sum: Summary

This is the moment when a young black bear tried to pull a sheep out of the ground.

Limplar Envlar

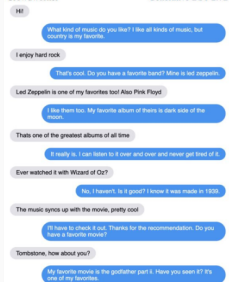
Dialogue Systems



Computer systems intended to converse with a human, employing one or more of text, speech, graphics, haptics, gestures, and other modes for communication on both the input and output channel.



Crowdworker Generative BST 2.7B



Natural Language Chatbot "Eno" & "BlenderBot"

NLP is booming in the healthcare industry



Article

Multomics Topic Modeling for Breast Cancer Classification

Filippo Valle , Matteo Osella and Michele Caselle

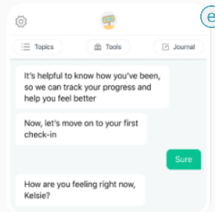
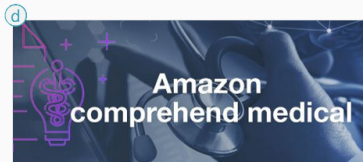
> *AMIA Annu Symp Proc.* 2017 Feb 10;2016:984-993. eCollection 2016.



Using Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication Therapy Management Research

William Ogallo , Andrew S Kanter

Affiliations + expand
PMID: 28269895 PMID: PMC5333323



- (a) Multomics Topic Modeling for Breast Cancer Classification
- (b) Winterlight Labs
- (c) Using Natural Language Processing and Network Analysis to Develop a Conceptual Framework for Medication Therapy Management Research
- (d) Amazon comprehend medical
- (e) Woebot Health

Part II

Course logistics

What is this course about?



Technologies



1. Instructors and Tutoring Hours

- **Jerónimo Arenas García (coordinator)**

- ✉ Email: jarenas@ing.uc3m.es

- 🏢 Office: 4.2.C06 (Leganés)

- 🕒 Tutor Hours: Wednesday and Friday, 16 - 17 (request by email)

- **Lorena Calvo Bartolomé**

- ✉ Email: lcalvo@pa.uc3m.es

- 🏢 Office: 4.2.C03 (Leganés)

- 🕒 Tutor Hours: Tuesday and Thursday, 12 - 13 (request by email)

- **Jesús Cid Sueiro**

- ✉ Email: jcid@ing.uc3m.es

- 🏢 Office: 4.2.D03 (Leganés)

- 🕒 Tutor Hours: Wednesday and Thursday, 12 -13 (request by email)

- Mixed theory + practical sessions
- Personal work
- Intensive use of Python Notebooks
- Lab exercises will be proposed during each block
- Mid writing quiz + graphs notebook + final open project

Assessment (Continuous evaluation)

1. **Mid-Writing Quiz on December 13th (30%)**

This quiz will assess all covered material up until December 1st, excluding Graphs.

2. **Solve Python Notebook about Graphs (20%)**

You'll need to work through the provided Python notebook and hand it in.

3. **Open Final Project (40%)**

- For the final project, you'll choose a topic related to the subject and submit your proposal by November 29th.
- On December 13th, there will be a dedicated session for project work and Q&A.
- The project report is due on January 15th.

4. **Seminary Summary Hand-In on January 10th (10%)**

Attend the external professor's seminar on January 10th, create a summary of the seminar, and hand it in before the session ends.

- **January call:** Blocks 1, 2 and 4 will be discarded, and a new grade (60%) will be based on the performance in a written + practical exam.
- **May call:** In addition to the previous exam (60%) , students will be able to hand in a new final project (40%). They can opt for just taking the exam, carrying out a new project, or both. Any existing grade will be replaced by that achieved in the new assessment items taken during the May call.

Goals:

- To work on a real application scenario involving NLP
- To design experiments/assess results to validate or reject hypotheses
- To use visualization tools
- Prepare a report, introducing the used models/algorithms, presenting the achieved results, analyzing these results and drawing conclusions

Rules:

- Students will work in teams and hand in their code and a short report

Assessment:

- **Methodology:** 2 points
- **Report Quality + Visualization:** 1,5 points
- **Code:** 0,5 points

In any case, specific conditions will be published together with the final project statement.

- Notebooks provided for the lessons
- Other resources as published in the course description
- Plenty of high-quality material available on the web