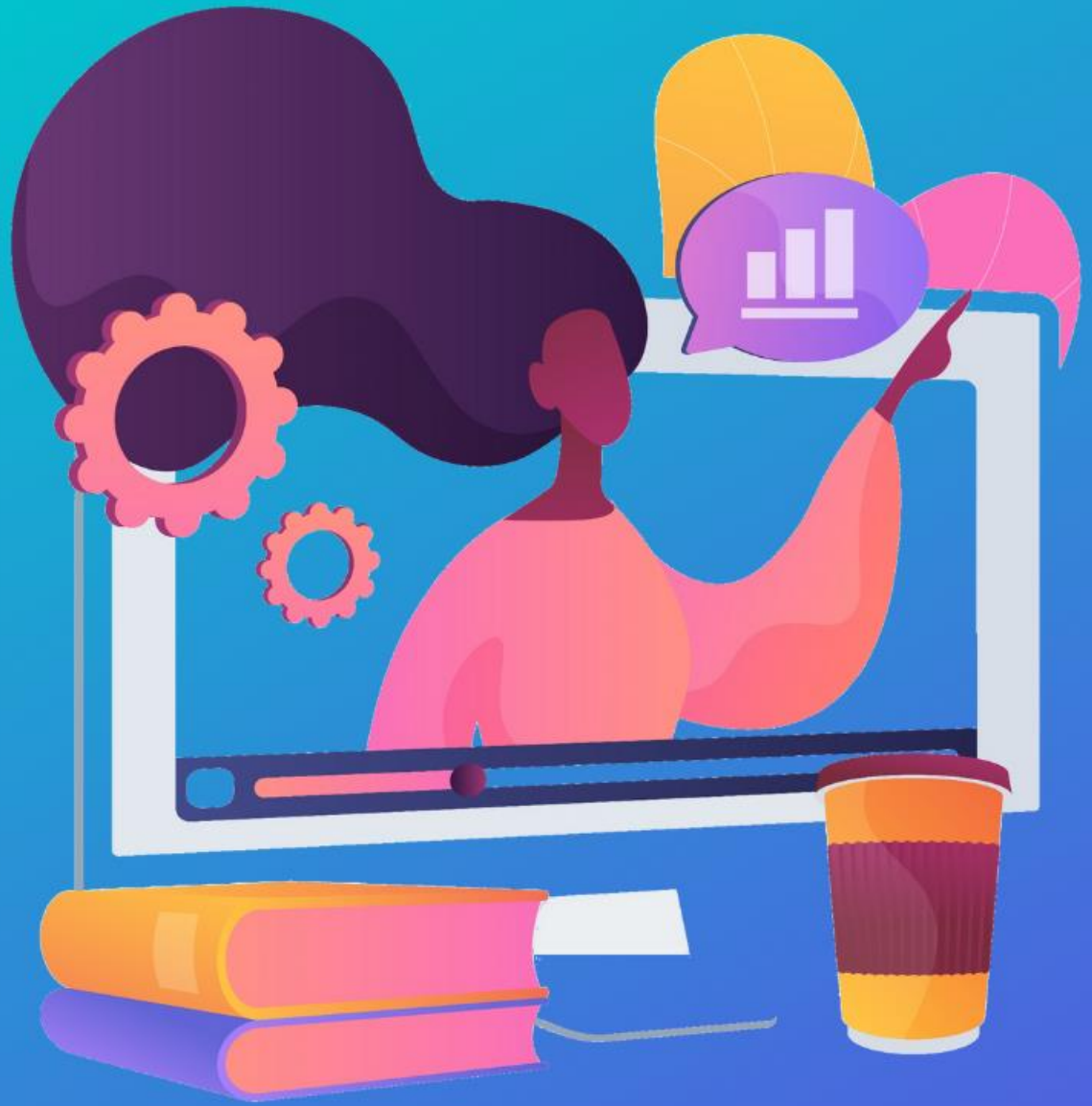


MBAZA NLP COMMUNITY

VIRTUAL TRAINING

22-24.06 | 3 - 5 PM

BASIC DATA WRANGLING WITH PANDAS
TEXT PRE-PROCESSING BASICS
NLP MODELING & ALGORITHMS

german cooperation
DEUTSCHE ZUSAMMENARBEIT

Implemented by
giz Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH

Digital Transformation Center Rwanda

FAIR FORWARD
Artificial Intelligence for all.

DIGITAL UMUGANDA

Why are we here?

Module 1
Basic Data Wrangling with Pandas

Any Data Science work you do in the future

Module 2
Text Pre-Processing Basics

Any NLP work you do in the future

Dataset cleaning challenge in July

Module 3
NLP Modelling & Algorithms

Your introduction to Machine Learning

Community activities on Chatbots and Voice

Review of Yesterday

Module 1
Basic Data Wrangling with Pandas

- **Jupyter Notebooks** provide a web-based interactive coding environment
- **Pandas** is an open-source Python library for data analysis & management
- It provides two basic data structures: `DataFrame` (two-dimensional) and `Series` (one-dimensional)
- `pd.read_csv()` allows you to import CSV files
- The `shape` attribute and `head()` method help you understand your data
- Pandas offers many methods to select data based on squared brackets:
  `data["column_name"][index_start:index_end]`
- Combine data by concatenation (`pd.concat()`) or via common identifier (`pd.merge()`)
- Basic operations: `.sort_values()` for sorting, `.drop_duplicates()` for dup removal
- Aggregation functions: `.mean()`, `.median()`, `.std()`, …
- Text functions: `.str.len()`, `.str.contains()`, `.str.lower()`, `.str.replace()`, …

Learning Outcomes of Today

Module 2
Text Pre-Processing Basics

You understand:

what NLP is good for

why text pre-processing is crucial for NLP

what regular expressions are good for

what parallel datasets are

You can:

use Pandas to cHaNgE cAsInG

~~remove~~ text parts

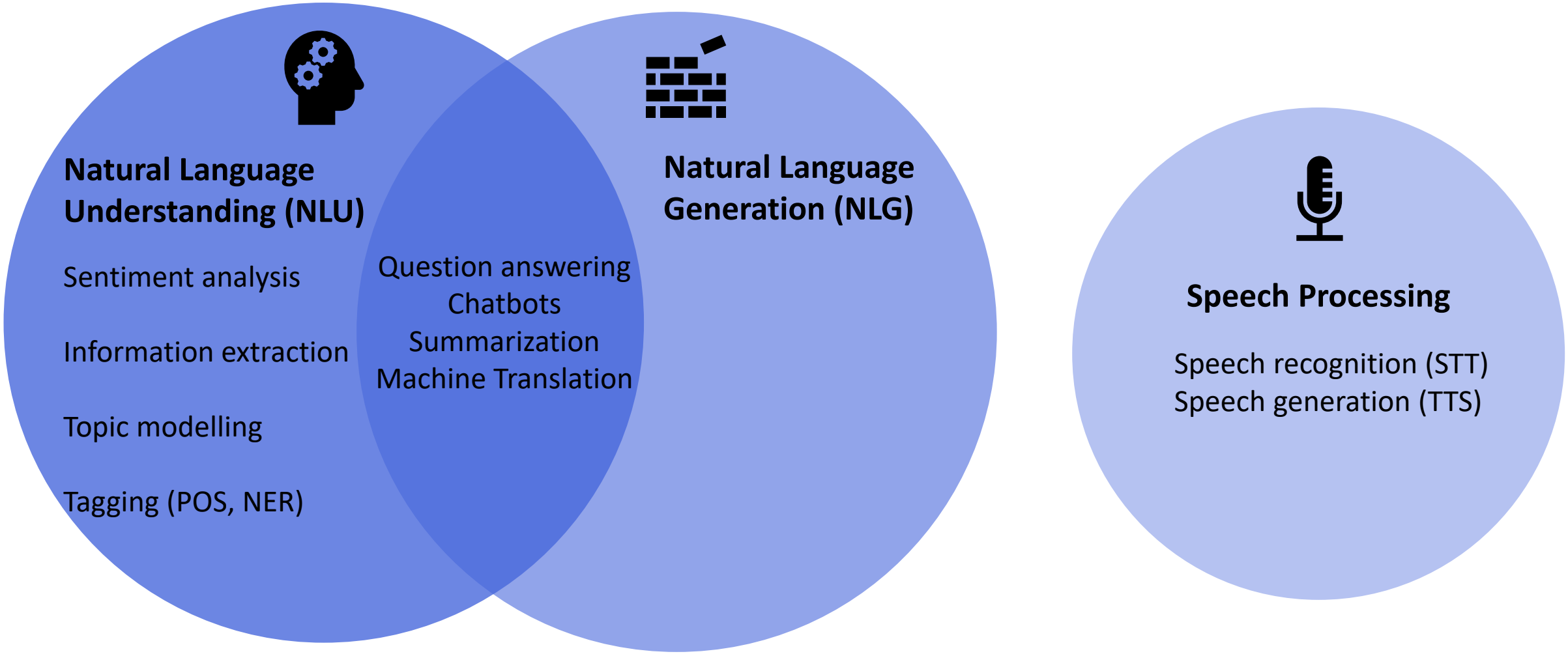
handle empty values

use regular expressions in Python

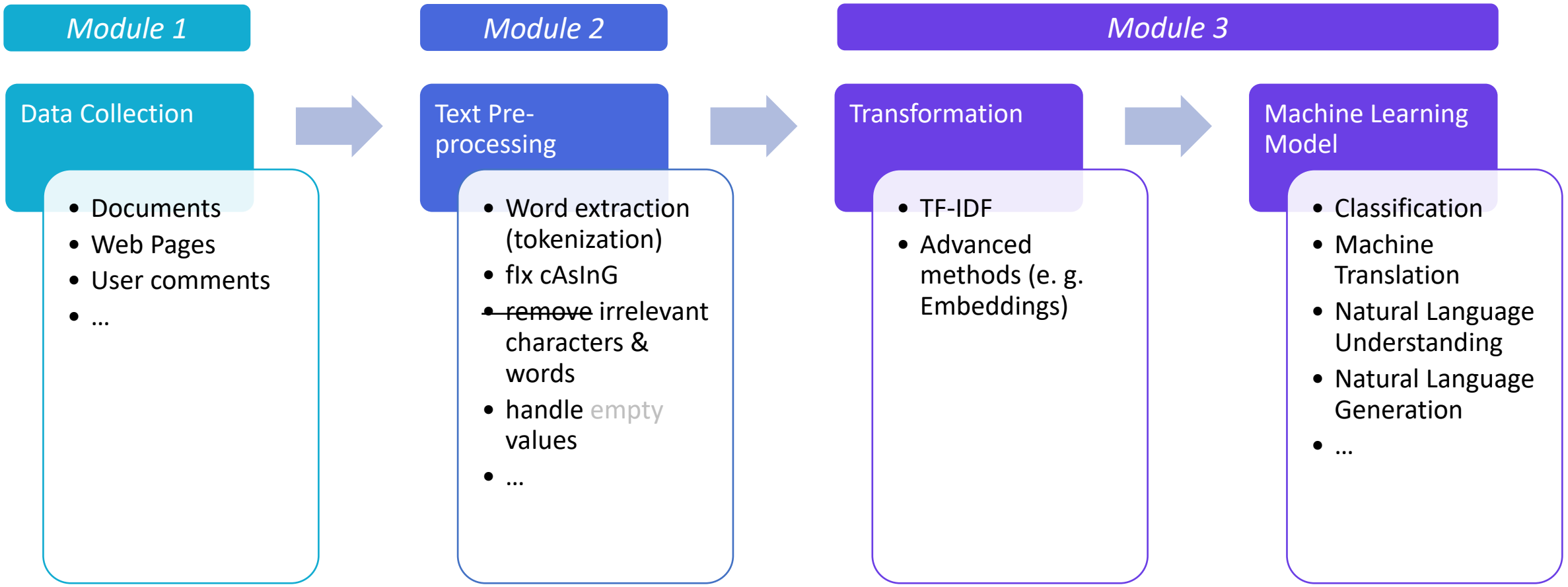
Practice on:

deal with parallel datasets and their challenges

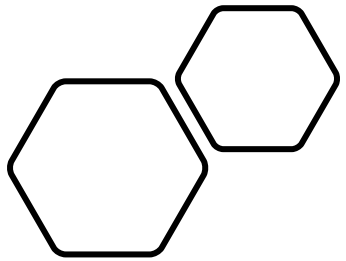
What is NLP & which tasks does it solve?



NLP Pipeline

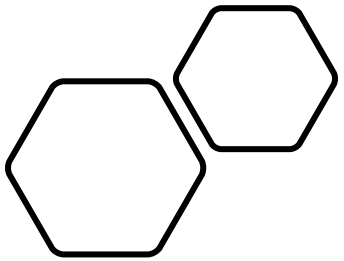


First Text Pre-Processing steps



1. Lower casing
2. Removing special characters
   - NOT alphabet or numeric
3. Removing certain words or parts of words (`replace()`)
4. Treating empty fields (NaN)

Regular Expressions (RegEx)



a sequence of characters that specifies a search pattern in text

1. Filtering HTML tags
2. Removing numerical values
3. Removing punctuations
4. Removing foreign and control characters
5. Using Regex with Pandas

Parallel Datasets

- Parallel datasets include the same text data, but in **two languages**
- This means that you in essence have to do **pre-processing twice**

Kinyarwanda

```
Nkoresha amasaha 2 ubusanzwe ntanga iki kiganiro ku banyeshuli bo mu mashuri yisumbu
Kandi byose byatangiriye mu ndenge imyaka 7 ishize ubwo nazaga hano TED.
Iruhande rwange hari hicaye umunyeshuli wo muri mashuri yisumbuye, w'umwangavu. Aturu
Yifuzaga kuba yagira ikintu cy'ingenzi yikora mu buzima bwe. yambajije akabazo korosh
Yaravuzengo, "ni iki umuntu yakora ngo akire?"
Nahise numva ntameze neza kuko ntari mpise nshobora kumusubiza igisubizo cyiza.
Nuko nyine mva mu ndenge nza hano TED
Narangiza ngatekereza, yegooo, ndi mu nzu irimo abantu benshi bageze ko byo bifuza(ba
Kuki ntababaza uko babigezeho ngo n'abana bato babimenye? naribajije.
Ngayo rero nguko dore imyaka 7 irashize, nabajije abantu 700 babigezeho uko babyumva
Freeman Thomas yaravuze ngo "Umuyobozi wanjye ni ubushake" abakora hano babikorera ul
Carol Coletta ati, "nanishyura umuntu kugirango akore ibyo nkora" ibanga ririmo ni i
Kora!
Ropert yarambiyengo "Byose ni ugukora cyane"
Nta kintu cy'ubuntu.
Kandi iyo ni yo soko y'ibyishimo"
Yavuze ibyishimo, Ropert?
Yego ntakindi yarengejeho.
Abakorera TED bakura ibyishimo muri uyu murimo.
Kandi barakora cyane.
Namenye ko ubuzima bwabo atari umurimo gusa.
Ni umurimo ariko bafata nk'umukino.
Byiza!
Alex aravuga ngo, "niba ushaka gukira, kiijugunyemo nk'ugwa mu mazi uvemo ari uko uk
Nta bufindo buhari; ni ugukora, ugakora, ukongera ugakora.
Ni no guhamya hamwe.
Norman yarambwiyengo,
"Bisaba gushyira imbaraga zawe zose ku kintu kimwe"
Kandi ushyiremo urutege.
David Galo ati.
"Ishyiremo imbaraga. ku mubiri, no mu myumvire, ugomba guhatiriza, ugasunika, ukanasu
Ugomba gukomeze ukashyiramo intege n'igihe wumva wabaye ikigwari ujijinganya.
Goldie Hawn aravuga ngo, "ntabwo ntagize gushidikanya ko ntabizi neza, ko ntari umuh
Sinumvaga ko nagera aho ngeze."
Ubu ntibyoroshye kwihaata, niyo mpamvu dufite ba Mama(ibitwenge)
"Mama yaransunitse"
Fasha abandi!
Sherwin Nuland aravugango " Byari iby'icyubahiro kuba dogiteri"
Ubu abana benshi barambwirango barashaka gutunga za miliyari ikintu cya mbere mbabwi
Ugomba guha abandi ikintu cy'agaciro.
```

English

```
This is really a two-hour presentation I give to high school students, cut down to t
And it all started one day on a plane, on my way to TED, seven years ago.
And in the seat next to me was a high school student, a teenager, and she came from
And she wanted to make something of her life, and she asked me a simple little quest
She said, "What leads to success?"
And I felt really badly, because I couldn't give her a good answer.
So I get off the plane, and I come to TED.
And I think, jeeze , I'm in the middle of a room of successful people!
So why don't I ask them what helped them succeed, and pass it on to kids?
So here we are, seven years, 500 interviews later, and I'm gonna tell you what really
Freeman Thomas says, "I'm driven by my passion." TEDsters do it for love; they don't
Carol Coletta says, "I would pay someone to do what I do." And the interesting thing
Work!
Rupert Murdoch said to me, "It's all hard work.
Nothing comes easily.
But I have a lot of fun."
Did he say fun?
Rupert? Yes!
TEDsters do have fun working.
And they work hard.
I figured, they're not workaholics.
They're work a frolics.
Good!
Alex Garden says, "To be successful put your nose down in something and get damn goo
There's no magic; it's practice, practice, practice.
And it's focus.
Norman Jewison said to me,
"I think it all has to do with focusing yourself on one thing."
And push!
David Gallo says, "Push yourself.
Physically, mentally, you've gotta push, push, push."
You gotta push through shyness and self-doubt.
Goldie Hawn says, "I always had self-doubts. I wasn't good enough; I wasn't smart en
I didn't think I'd make it."
Now it's not always easy to push yourself, and that's why they invented mothers.
"My mother pushed me."
Serve!
Sherwin Nuland says, "It was a privilege to serve as a doctor."
Now a lot of kids tell me they want to be millionaires. And the first thing I say to
"OK, well you can't serve yourself; you gotta serve others something of value.
```

let's get started!

Open the Colab link

Report back
Q & A

Summary

Module 2
Text Pre-Processing Basics

- Natural Language Processing (NLP) is an important field of AI and has many applications
- Basic tasks include **Natural Language Understanding (NLU)** and **Generation (NLG)**, as well as **speech processing**
- **Text Pre-Processing** is a crucial step in the NLP pipeline; it is done after data collection and before transforming text and using Machine Learning algorithms
- Basic text pre-processing steps include **cHaNgInG cAsInG**, ~~**removing**~~ **text parts**, and **handling** empty **values**
- **Regular expressions** are a flexible tool to selecting specific text part and dealing with it
- **Parallel datasets** include the same text data, but in two languages

Learning Outcomes of Today

Module 3
NLP Modelling & Algorithms

You understand:

what the scikit-learn library does
how the TF-IDF approach converts text to numeric information
which Machine Learning tasks exist
basics of how Neural Networks function
how to evaluate your model's performance
approaches to improve your model's performance

You can:

use Machine Learning for classifying news articles
use scikit-learn for TF-IDF text pre-processing
split your data into training and test sets
train a neural network with scikit-learn
generate an evaluation report with scikit-learn

Join the Mbaza NLP Community!

WhatsApp

https://chat.whatsapp.com/BRlxzsFiZgsLmK5SBT2XUo

Slack

https://join.slack.com/t/mbazanlpcommunity/shared_invite/zt-19ie5idhj-f0yWfOBgTKzs7VOKCcr_pw

GitHub

https://github.com/MBAZA-NLP

Hugging Face

https://huggingface.co/organizations/mbazaNLP/share/mUKyOkYpSRisRpspbfuwUvoQgWyfdiJYqU