Traitement Automatique du Langage Naturel

TP4 : Summarisation "résumé automatique"

Start notebook TP 4

The dataset (Arabic part) is included in the start notebook from WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization

you can also check thier github

For the model, you can use :

mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer

just load the mT5 from huggingface

or

AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization

just load the AraBART from huggingface

Using the start notebook provided in the link and the dataset attached to it, build and train your own Abstractive summarisation model for the Arabic language using either mT5 or AraBART model

you will be evaluated according to the performance of your model ROUGE metric

Summarization : creates a shorter version of a document or an article that captures all the important information. Along with translation, it is another example of a task that can be formulated as a sequence-to-sequence task. Summarization can be:

- **Extractive**: extract the most relevant information from a document.
- **Abstractive**: generate new text that captures the most relevant information.