Arabic NLP Libraries

Prof. Gheith Abandah

Outline

- PyArabic
- Camel Tools
- AraBERT
- Exercise

PyArabic

- Is an Arabic text processing library that provides capabilities for Arabic character and **token manipulation**, including normalization, segmentation, and more. It's particularly **useful for preprocessing steps**.
  - Installation: `pip install pyarabic`
  - Project Description
  - Package Documentation

Camel Tools

- Is a suite of Arabic natural language processing tools specifically designed for Arabic. It includes utilities for **preprocessing**, **morphological analysis**, **POS tagging**, **named entity recognition**, **sentiment analysis**, and more.
  - Installation: `pip install camel-tools`
  - Online documentation
  - The Guided Tour

Outline

- PyArabic
- Camel Tools
- AraBERT
- Exercise

AraBERT

- AraBERT is a pre-trained language model specifically **designed for Arabic**.
- It is part of the **BERT** (Bidirectional Encoder Representations from Transformers) family of models, which are based on the Transformer architecture and are widely used for natural language processing tasks.
- AraBERT has been **trained on a large corpus of Arabic text**, making it well-suited for a variety of Arabic NLP tasks.

Key Features of AraBERT

- **Pre-trained on Arabic Text**: AraBERT is trained on a large dataset of Arabic text, which includes diverse sources such as Wikipedia, news articles, and other web texts. This makes it particularly effective for understanding and generating Arabic language.
- **Transformer Architecture**: Like BERT, AraBERT uses the Transformer architecture, which allows it to consider the context of words in a sentence more effectively than traditional models.
- **State-of-the-Art Performance**: AraBERT achieves state-of-the-art performance on several Arabic NLP benchmarks and tasks.

Variants

Model	HuggingFace Model Name	Size (MB/Params)	Pre-Segmentation	DataSet (Sentences/Size/n Words)
AraBERTv0.2-base	bert-base-arabertv02	543MB / 136M	No	200M / 77GB / 8.6B
AraBERTv0.2-large	bert-large-arabertv02	1.38G 371M	No	200M / 77GB / 8.6B
AraBERTv2-base	bert-base-arabertv2	543MB 136M	Yes	200M / 77GB / 8.6B
AraBERTv2-large	bert-large-arabertv2	1.38G 371M	Yes	200M / 77GB / 8.6B
AraBERTv0.1-base	bert-base-arabertv01	543MB 136M	No	77M / 23GB / 2.7B
AraBERTv1-base	bert-base-arabert	543MB 136M	Yes	77M / 23GB / 2.7B

Applications of AraBERT

- **Sentiment Analysis**: Understanding the sentiment of Arabic texts such as tweets, reviews, and comments.
- **Named Entity Recognition**: Identifying and classifying named entities in Arabic texts.
- **Text Classification**: Classifying Arabic texts into predefined categories.
- **Question Answering**: Building question-answering systems that can understand and answer questions posed in Arabic.

AraBERT Documentation

- https://github.com/aub-mind/arabert

Exercise

- Using Hugging Face Transformers library, find an mT5 model that can accurately summarize Arabic paragraphs, e.g.:

```
arabic_text = """
احتضنت كلية الهندسة في الجامعة الأردنية الأمس المهرجان التكنولوجي الوطني الحادي عشر، وقد كان إلمهرجان مناسبة تكنولوجية وطنية فريدة، حيث شارك طلبة وأساتذة من مختلف الجامعات الأردنية ومختصين من الصناعة في فعاليات المهرجان التي تضمنت معرضًا لمشاريع الطلبة ومسابقات لمشاريع الطلبة على 9 محاور لتطبيقات التكنولوجيا في نواحي الحياة المختلفة، ومعارض للشركات الراعية، ومحاضرات علمية ومهنية. والمهرجان هو فعالية سنوية تتنقل بين الجامعات الأردنية أعيد إحيائها بعد توقف، ويشرف عليها فرع الأردن لمجمع المهندسين الكهربائيين والإلكترونيين ولجنة توجيهية وطنية، ونظمته في الجامعة الأردنية هذا ألعام لجنة تنظيمية خاصة من كلية الهندسة. أنا سعيد بإعادة إحياء المهرجان، وأشكر المنظمين له من فرع الأردن والجامعة ألأردنية واللجنة التوجيهية واللجنة التنظيمية، وأهنئ الفائزين في المسابقات ومنهم من طلبتي في قسم هندسة الحاسوب.
"""
```

Summary

- PyArabic
- Camel Tools
- AraBERT
- Exercise