Task 1: Tokenization, lemmatization, POS-tagging, and NER

Data

- Kazakh TreeBank https://github.com/UniversalDependencies/UD_Kazakh-KTB

  Tyers, Francis M., and Jonathan Washington. "Towards a Free/Open-source Universaldependency Treebank for Kazakh." PROCEEDINGS OF THE INTERNATIONAL CONFERENCE" TURKIC LANGUAGES PROCESSING" TurkLang-2015. 2015.

  Makazhanov, Aibek, et al. "Syntactic annotation of kazakh: Following the universal dependencies guidelines. a report." *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE" TURKIC LANGUAGES PROCESSING" TurkLang-2015*. 2015.

- KazNERD https://github.com/IS2AI/KazNERD/tree/main/KazNERD

  Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. KazNERD: Kazakh Named Entity Recognition Dataset. LREC’2022.

Tasks

1. Stanza
   a. Install Stanza library.
   b. Read Stanza documentation, familiarize yourself with Stanza’s tokenizers, lemmatizers, POS-taggers, and named entity recognizers.
2. Lemmatization and POS-tagging
   a. Collect sentences from the **test** subset of the KTB (*# text* fields), join them using double newlines (*\n\n*).
   b. Define a pipeline with a tokenizer, lemmatizer, and POS-tagger. (Use `tokenize_no_ssplit=True`, see details.) Run the pipeline.
   c. Collect lemmas and POS tags from the KBT annotations and those produced by the Stanza pipeline. Check whether sentences in the KTB and Stanza annotations have the same number of words. Analyze and report discrepancies. Calculate lemmatization and POS tagging accuracy scores for the sentences with the matching number of words.
   d. Report results, analyze errors.
3. Named Entity Recognition (NER)
   a. Generate a document from the KazNERD **test** subset: words separated by spaces, sentences separated by newlines, see details.
   b. Define a new pipeline with a tokenizer and NER (use `tokenize_pretokenized=True`).
   c. Collect ground truth annotations from KazNERD and annotations generated by Stanza. Use https://huggingface.co/spaces/evaluate-metric/seqeval for evaluation.
   d. Report results, analyze errors.
4. Possible add-on
   a. Evaluate Stanza on a small collection of noisy social media and/or fiction texts (~10 sentences of each genre).