🤨

CoNLL-U

Задача №1

Условие

У вас есть файл формата CoNLL-U (`.conllu`), с которым вы уже начинали работать в начале прошлой пары.

Задание

Выполните задание с прошлой пары при помощи библиотеки `conllu`:

- Составьте список `lemmas` всех лемм (начальных форм) слов без пунктуации.
- Потом создайте список уникальных лемм `unique_lemmas`.
- Сделайте частотный словарь лемм `dict_lemmas`, где ключами будут уникальные леммы, а значениями — их количество в файле.

NB! Возможно, некоторые (или все) этапы задания будут сопровождаться тем, что `print()` будет отказываться печатать такие большие элементы. Не переживайте и выводите по 100-200 элементов, если подобное происходит.

Задача №2

Условие

У вас есть пять файлов формата CoNLL-U (`.conllu`) с разобранными первыми абзацами статьи из Википедии с заголовком “Язык” на разных идиомах:

1. русский
2. турецкий
3. арабский
4. индонезийский

5. баскский

Задание

1. Сначала посмотрите на русский язык. Составьте словарь `russian_tags`, где каждой части речи (`upos`) будет соответствовать словарь морфологических признаков (`feats`), внутри которого ключами будут признаки, а значениями — списки значений, принимаемых признакам. Например, неполный словарь может выглядеть так:

   ```
   {'NOUN': {'Number': ['Sing', 'Plur'], 'Case': ['Acc', 'Abl']}, 'VERB': {'Evident':
    ['Fh', 'Nfh']}}
   ```

   NB! Советуем сохранить полученный словарь в формате JSON (`.json`) и отсортировать в алфавитном порядке.

   Не переживайте, если сокращения кажутся совсем непонятными. Вот описание (к сожалению, на английском) для русских частей речи и морфологических признаков.

2. Проделайте подобную работу и для остальных языков, получив
   a. турецкий словарь `turkish_tags` (описание тегов)
   b. арабский словарь `arabic_tags` (описание тегов)
   c. индонезийский словарь `indonesian_tags` (описание тегов)
   d. баскский словарь `basque_tags` (описание тегов)

3. **Самая важная часть!**
   Посмотрите, какие части речи, морфологические признаки и их значения отсутствуют в русском языке, но присутствуют в других языках. Составьте список подобных отличий, где будут присутствовать все промежуточные “ключи” и отсутствующие признаки. Например, можно сделать список строк:

   ```
   ['CCONJ', 'NOUN Number Plur', , 'VERB Evident']
   ```

4. (*Если останется время*)
   Попробуйте вывести слова соответствующие значениям из списка в предыдущем задании. В этом вам поможет `.filter()`