🤨

CoNLL-U 2.0

Задача №1

Условие

У вас есть файл формата CoNLL-U (`.conllu`) на языке суахили. Это крупный африканский язык, относящийся к семье банту.

Для языков этой семьи характерно наличие большого количества именных классов (по сути, аналог рода у существительных в русском, только их может быть несколько десятков). В суахили, например, в районе 15 классов, выражающихся различными приставками:

Class	Class prefix	Example word	Concord (subject, object)	Referential Concord	Possessive Concord	‘Meaning’
1	m	mtu ‘person’	a/yu	ye	wa	People
2	wa	watu ‘people’	wa	o	wa	
3	m	mti ‘tree’	u	o	wa	Trees, plants
4	mi	miti ‘trees’	i	yo	ya	
5	ji/∅	jicho ‘eye’	li	lo	la	Round things, liquids, masses, augmentatives
6	ma	macho ‘eyes’	ya	yo	ya	
7	ki	kiti ‘chair’	ki	cho	cha	Artefacts, tools, manner, diminutives
8	vi	viti ‘chairs’	vi	vyo	vya	
9	n/∅	ndege ‘bird’	i	yo	ya	Animals, loanwords
10	n/∅	ndege ‘birds’	zi	zo	za	
11	u	ubao ‘board’	u	o	wa	Long things, abstracts
15	ku	kuimba ‘to sing’	ku	ko	kwa	Infinitives
16	*(pa)*	*mahali ‘place’*	*pa*	*po*	*pa*	*Locatives*
17	*(ku)*		*ku*	*ko*	*kwa*	
18	*(mu)*		*m*	*mo*	*mwa*	

Задание

Составьте список словоформ каждого класса (обратите внимание, что вам нужны только существительные) и запишите все в словарь.

Для каждого класса найдите его приставку, отвечающую за именной класс существительного (если такая вообще есть). Сравните полученные вами результаты с таблицей выше и описанием на Википедии.

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()`

5. (*Убедитесь сперва, что всё предыдущее выполнено!)*
   Могут быть такие ситуации, что у нас есть много текстов на разных языках. Может быть даже уже разобранных в `conllu`. А может, у нас сложная система, где сперва для текста определяется язык (*это само по себе интересно!! как вообще можно определять язык текста? а что если в тексте отрывки на разных языках? безумие! но методы бывают…*), а потом тексты размечаются в `conllu`.

   Так вот, постарайтесь как можно больше узнать про языки по нескольким размеченным текстам. Разберитесь с языками выше, используя те же файлы `.conllu`. Некоторые идеи:

   a. Как часто употребляются слова каждой части речи?
      т.е. слова которые мы считаем глаголами, существительными, прилагательными, и т.д. (“*Считаем*” сказано потому, что не всегда очевидно, какой части речи слово в языке, ср. причастия!)

```
{"Арабский": {'NOUN': 31, 'VERB': 141, ...}, "Турецкий": {...}, ...}
```

Сделайте выводы: в каких языках распространена модификация прилагательными или наречиями, а в каких нет? В каких языках часто встречаются “специальные” части речи, редкие или вообще отсутствующие в других языках? В каких языках существительных сильно больше глаголов, а в каких, наоборот, глаголов больше, чем существительных?

Как часто выражаются на словах *каждой части речи* (!) различные грамматические категории и грамматические значения? Обратите внимание, здесь полезно считать как *относительные значения*. (вдруг в каком-то языке глаголы вообще редки? - Это вы из предыдущего задания всё могли узнать!). А чтобы можно было понять, насколько грамматическая категория или значение важны для языка, эти относительные значения могут считаться как $\frac{\text{частота грамматического\_значения}}{\text{частота части\_речи}}$

(либо заводите два отдельных словаря)

```
# словарь долей грам. значений для каждой категории
{'NOUN': {'Number': {'Sing': 0.67, 'Plur': 0.33}, 'Case': {'Acc': 0.85, 'Abl': 0.15}}, 'VE
RB': {'Evident': {'Fh': 0.41, 'Nfh': 0.59}}}

# словарь долей категорий в части речи
{'NOUN': {'Number': 1.0, 'Case': 1.0}, 'VERB': {'Evident': 0.6}}
```

Сделайте выводы: Какие категории частей речи в языках *обязательны* (т.е. присутствуют на каждом слове определённой части речи), а какие нет (т.е. выражаются не на всех словах этой части речи? (Обратите внимание на `VerbForm` в русском! У всех ли слов это есть? А какая причина?) Обнаружились ли различия между языками?

> Наконец, установите частотность разных порядков слов в языках. Порядок слов это порядок S (подлежащего), O (объекта / дополнения) и V (глагола-сказуемого). Какие в языках самые частые порядки слов? Есть ли во всех языках какой-то наиболее предпочтительный? Есть ли наиболее предпочтительный порядок слов в русском?
>
> Обратите внимание: подлежащее, сказуемое и дополнение можно определить из разметки! Нужно использовать её части `dep` и `rel`.