به نام خدا

درس مبانی پردازش زبان و گفتار

کارگاه ابزار دادماتولز

سینا علی‌نژاد

۹۹۵۲۱۴۶۹

- برای ماژول normalizer مورد زیر را یافتم که نتوانسته عدد ۱۲ را با کلمه‌ی "عدد" جایگزین کند.

```
from dadmatools.normalizer import Normalizer

normalizer = Normalizer(
    full_cleaning=False,
    unify_chars=True,
    refine_punc_spacing=True,
    remove_extra_space=True,
    remove_puncs=False,
    remove_html=False,
    remove_stop_word=False,
    replace_email_with="<EMAIL>",
    replace_number_with="عدد",
    replace_url_with="",
    replace_mobile_number_with=None,
    replace_emoji_with=None,
    replace_home_number_with=None
)

text = "من دیروز12 مدرسه رفتم"
print('input text : ', text)
print('output text when replace emails and remove urls : ', normalizer.normalize(text))
```

```
input text :  من دیروز12 مدرسه رفتم
output text when replace emails and remove urls :  من دیروز12 مدرسه رفتم
```

- برای ماژول lemmatize فعل "فهمیدم" را دادم و خروجی نیز همان "فهمیدم" شد. در حالیکه باید فهمید/فهم میشد.

```
nlp_lem('فهمیدم')
```

```
{'sentences': [{'id': 1,
   'tokens': [{'id': 1,
     'text': 'فهمیدم',
     'upos': 'NOUN',
     'xpos': 'N_SING',
     'feats': 'Number=Sing',
     'head': 0,
     'deprel': 'root',
     'lemma': 'فهمیدم'}]}],
 'lang': 'persian'}
```

- برای عملیات part of speech tagging جلمه "ناگهان باران بارید" را دادم و برای کلمه‌ی ناگهان پیشبینی نادرست انجام شد. باید ADV بدهد اما Noun داد.

```
nlp_pos('ناگهان باران بارید')
```

```
{'sentences': [{'id': 1,
   'tokens': [{'id': 1,
     'text': 'ناگهان',
     'upos': 'NOUN',
     'xpos': 'N_SING',
     'feats': 'Number=Sing',
     'head': 0,
     'deprel': 'root'},
    {'id': 2,
     'text': 'باران',
     'upos': 'NOUN',
     'xpos': 'N_SING',
     'feats': 'Number=Sing',
     'head': 1,
     'deprel': 'nsubj'},
    {'id': 3,
     'text': 'بارید',
     'upos': 'NOUN',
     'xpos': 'N_SING',
     'feats': 'Number=Sing',
     'head': 1,
     'deprel': 'root'}]}],
 'lang': 'persian'}
```

- برای عملیات named entity recognition در جمله ای که از عبارت "دکتر احمدی" استفاده کردم، نتوانست متوجه شود که دکتر احمدی با همدیگر یک entity است و فقط کلمه‌ی "احمدی" را به عنوان شخص شناخت. برای دیگر عناوین مثل پروفسور سمیعی نیز به همین صورت بود.

```
nlp_ner('دکتر احمدی، رئیس دانشگاه تهران، در کنفرانس هوش مصنوعی که در هتل آزادی برگزار شد، سخنرانی کرد.')
```

```
{'sentences': [{'id': 1,
   'tokens': [{'id': 1, 'text': 'دکتر', 'ner': 'O'},
    {'id': 2, 'text': 'احمدی', 'ner': 'S-PER'},
    {'id': 3, 'text': '،', 'ner': 'O'},
    {'id': 4, 'text': 'رئیس', 'ner': 'O'},
    {'id': 5, 'text': 'دانشگاه', 'ner': 'B-ORG'},
    {'id': 6, 'text': 'تهران', 'ner': 'E-ORG'},
    {'id': 7, 'text': '،', 'ner': 'O'},
    {'id': 8, 'text': 'در', 'ner': 'O'},
    {'id': 9, 'text': 'کنفرانس', 'ner': 'O'},
    {'id': 10, 'text': 'هوش', 'ner': 'O'},
    {'id': 11, 'text': 'مصنوعی', 'ner': 'O'},
    {'id': 12, 'text': 'که', 'ner': 'O'},
    {'id': 13, 'text': 'در', 'ner': 'O'},
    {'id': 14, 'text': 'هتل', 'ner': 'B-LOC'},
    {'id': 15, 'text': 'آزادی', 'ner': 'E-LOC'},
    {'id': 16, 'text': 'برگزار', 'ner': 'O'},
    {'id': 17, 'text': 'شد', 'ner': 'O'},
    {'id': 18, 'text': '،', 'ner': 'O'},
    {'id': 19, 'text': 'سخنرانی', 'ner': 'O'},
    {'id': 20, 'text': 'کرد', 'ner': 'O'},
    {'id': 21, 'text': '.', 'ner': 'O'}]}],
 'lang': 'persian'}
```

- برای گراف وابستگی، بهش جمله "حراست دانشگاه سمنان" رو دادم. اینجا باید بدین شکل عمل بکنه که کلمه "حراست" ریشه هست و کلمه "دانشگاه" وابسته به کلمه "حراست" هست و کلمه "سمنان" وابسته به کلمه "دانشگاه" هست. اما خروجی یک اشتباه دارد و آن هم روی کلمه "سمنان" است که آن را وابسته کلمه "حراست" می‌داند.

```
nlp_dep('حراست دانشگاه سمنان')
```

```
{'sentences': [{'id': 1,
   'tokens': [{'id': 1,
     'text': 'حراست',
     'upos': 'NOUN',
     'xpos': 'N_SING',
     'feats': 'Number=Sing',
     'head': 0,
     'deprel': 'root'},
    {'id': 2,
     'text': 'دانشگاه',
     'upos': 'NOUN',
     'xpos': 'N_SING',
     'feats': 'Number=Sing',
     'head': 1,
     'deprel': 'nmod:poss'},
    {'id': 3,
     'text': 'سمنان',
     'upos': 'NOUN',
     'xpos': 'N_SING',
     'feats': 'Number=Sing',
     'head': 1,
     'deprel': 'nmod:poss'}]}],
 'lang': 'persian'}
```

- در جمله زیر، نتوانسته متوجه شود "دیشپ" همان "دیشب" است. همچنین "دیرنت" همان "دیدنت" است.

```
nlp_spell('دیشپ گذفتار بودم نتونستم به دیرنت بیام')
```

```
1it [00:00,  2.76it/s]
{'spellchecker': {'orginal': 'دیشپ گذفتار بودم نتونستم به دیرنت بیام',
  'corrected': 'پیش گرفتار بودم نتونستم به دیرنت بیام',
  'checked_words': [('پیش' ,'دیشپ'), ('گرفتار' ,'گذفتار')]},
 'sentences': [{'id': 1,
   'tokens': [{'id': 1, 'text': 'دیشپ'},
    {'id': 2, 'text': 'گذفتار'},
    {'id': 3, 'text': 'بودم'},
    {'id': 4, 'text': 'نتونستم'},
    {'id': 5, 'text': 'به'},
    {'id': 6, 'text': 'دیرنت'},
    {'id': 7, 'text': 'بیام'}]}],
 'lang': 'persian'}
```

- در جمله زیر، در دو جا نتوانسته تشخیص دهد کلمه نیاز به کسره دارد. یکی در عبارت "کلید در" روی کلمه "کلید" و دیگری در عبارت "حراست دانشگاه علم و صنعت" روی کلمه "دانشگاه"

```
nlp_kasreh('کلید در را برداشتم و به حراست دانشگاه علم و صنعت تحویل دادم.')
```

```
{'sentences': [{'id': 1,
   'tokens': [{'id': 1, 'text': 'کلید', 'kasreh': 'O'},
    {'id': 2, 'text': 'در', 'kasreh': 'O'},
    {'id': 3, 'text': 'را', 'kasreh': 'O'},
    {'id': 4, 'text': 'برداشتم', 'kasreh': 'O'},
    {'id': 5, 'text': 'و', 'kasreh': 'O'},
    {'id': 6, 'text': 'به', 'kasreh': 'O'},
    {'id': 7, 'text': 'حراست', 'kasreh': 'S-kasreh'},
    {'id': 8, 'text': 'دانشگاه', 'kasreh': 'O'},
    {'id': 9, 'text': 'علم', 'kasreh': 'O'},
    {'id': 10, 'text': 'و', 'kasreh': 'O'},
    {'id': 11, 'text': 'صنعت', 'kasreh': 'O'},
    {'id': 12, 'text': 'تحویل', 'kasreh': 'O'},
    {'id': 13, 'text': 'دادم', 'kasreh': 'O'},
    {'id': 14, 'text': '.', 'kasreh': 'O'}]}],
 'lang': 'persian'}
```

- در جمله زیر، پس از بیان اینکه کتاب مذکور بسیار محبوب است و نگارشش خوب بوده، نظر خودم را گفتم که با آن ارتباط نگرفتم. پس باید احساس منفی داشته باشد ولی مدل احساس مثبت را با احتمال ۷۷ درصد پیش بینی کرده است.

```
[28] nlp_sent('با اینکه کتاب بسیار خوب نگارش شده بود و از نظر خیلی‌ها شاهکاری در ادبیات بود، نتوانستم با آن ارتباط بگیرم.')['sentiment']

[{'label': 'positive', 'score': 0.7725887298583984}]
```

- در عبارت "خیلی خستم" نتوانسته است کلمه "خستم" را به شکل رسمی آن یعنی "خسته‌ام" یا "خسته هستم" تغییر دهد.

```
nlp_itf('خیلی خستم نمیتونم کار کنم.')
```

```
{'itf': ' خیلی خستم نمی\u200cتونم کار بکنم .',
 'sentences': [{'id': 1,
   'tokens': [{'id': 1, 'text': 'خیلی'},
    {'id': 2, 'text': 'خستم'},
    {'id': 3, 'text': 'نمیتونم'},
    {'id': 4, 'text': 'کار'},
    {'id': 5, 'text': 'کنم'},
    {'id': 6, 'text': '.'}]}],
 'lang': 'persian'}
```