بسم الله الرحمن الرحیم

پردازش زبانهای طبیعی

جلسه ۵ و ۶ (عمده جلسه ۶ از روی نوتبوک)

احسان الدین عسگری

اسفند ۱۴۰۲

http://language.ml/
دانشکده مهندسی کامپیوتر
آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

asgari@berkeley.edu

مراحل پیش پردازش

توکنایزیشن

نرمالایزیشن فیض روح القدس اگر باز مدد فرماید دیگران هم بکنند آن چه مسیح می کرد

ظاهر فیض رُوح القدس ار باز مـــــدد فرماید دیـــگران هم بکنند آن چه مَسیحا می کرد

مراحل پیش پردازش

توکنایزیشن	فیض روح‌القدس اگر باز مدد فرماید	دیگران هم بکنند آن‌چه مسیح می‌کرد
نرمالایزیشن	فیض روح القدس اگر باز مدد فرماید	دیگران هم بکنند آن چه مسیح می کرد
ظاهر	فیض رُوح القدس ار باز مـــــدد فرماید	دیـــگران هم بکنند آن چه مَسیحا می کرد

تقلیل فرم (stemming)

- خرد کردن لغت به فرم کوچک
- حذف پسوندها و پیشوندها
- کلمات با ریشه یکسان به یک کلمه واحد

خانوار - خانواده

ریشه‌یابی (lemmatization)

- بردن کلمه به ریشه
  - افعال
  - اسامی

با استفاده از دیکشنری یا پارسر ساخت لغت

مرد – گرد

?

شست – قوی

کاربرد ادات سخن، جزء کلام، یا Part of speech

- محدود کردن معنا (مرد؟ فعل یا اسم)
- چگونگی تلفظ (قوی؟ صفت یا اسم)
- ریشه یابی

آنچه در POS نهفته

- معنا
- قواعد نحوی
- ریشه‌یابی
- و ترکیبی از آنها

چه چالشهایی؟

مراحل پیش پردازش

ادات سخن

```
In [9]: tagger = POSTagger(model='resources/postagger.model')

        for x, y in tagger.tag(tokens):
            print(F"{x} ---> {y} ")
```

```
فیض <--- Ne
روح <--- Ne
القدس <--- N
ار <--- CONJ
باز <--- ADV
مدد <--- N
فرماید <--- V
دیگران <--- PRO
هم <--- CONJ
بکنند <--- V
آن <--- PRO
چه <--- DET
مسیحا <--- N
می‌کرد <--- V
```

توکنایزیشن فیض روح‌القدس اگر باز مدد فرماید دیگران هم بکنند آن‌چه مسیح می‌کرد

نرمالایزیشن فیض روح القدس اگر باز مدد فرماید دیگران هم بکنند آن چه مسیح می کرد

ظاهر فیض رُوح القدس ار باز مـــــدد فرماید دیـــگران هم بکنند آن چه مَسیحا می کرد

مراحل پیش پردازش

ادات سخن

```
In [9]: tagger = POSTagger(model='resources/postagger.model')

for x, y in tagger.tag(tokens):
    print(F"{x} ---> {y} ")
```

```
Ne <--- فیض
Ne <--- روح
N <--- القدس
CONJ <--- ار
ADV <--- باز
N <--- مدد
V <--- فرماید
PRO <--- دیگران
CONJ <--- هم
V <--- بکنند
PRO <--- آن
DET <--- چه
N <--- مسیحا
V <--- میکرد
```

ریشه‌یابی و تقلیل فرم

```
In [8]: tokens = word_tokenize(normalized)

stemmer = Stemmer()
lemmatizer = Lemmatizer()

for tok in tokens:
    print(F"{tok} ---stem--> ", stemmer.stem(tok))
    print(F"{tok} ---lemma--> ", lemmatizer.lemmatize(tok))
```

```
فیض  <--stem--- فیض
فیض  <--lemma--- فیض
روح  <--stem--- روح
روح  <--lemma--- روح
القدس  <--stem--- القدس
القدس  <--lemma--- القدس
ار  <--stem--- ار
ار  <--lemma--- ار
باز  <--stem--- باز
باز  <--lemma--- باز
مدد  <--stem--- مدد
مدد  <--lemma--- مدد
فرماید  <--stem--- فرماید
فرماید#فرما  <--lemma--- فرماید
دیگر  <--stem--- دیگران
دیگر  <--lemma--- دیگران
ه  <--stem--- هم
هم  <--lemma--- هم
بکنند  <--stem--- بکنند
بکنند#کن  <--lemma--- بکنند
آن  <--stem--- آن
آن  <--lemma--- آن
چه  <--stem--- چه
چه  <--lemma--- چه
مسیحا  <--stem--- مسیحا
مسیحا  <--lemma--- مسیحا
میکرد  <--stem--- میکرد
کرد#کن  <--lemma--- میکرد
```

توکنایزیشن فیض روح‌القدس اگر باز مدد فرماید دیگران هم بکنند آن‌چه مسیح می‌کرد

نرمالایزیشن فیض روح القدس اگر باز مدد فرماید دیگران هم بکنند آن چه مسیح می کرد

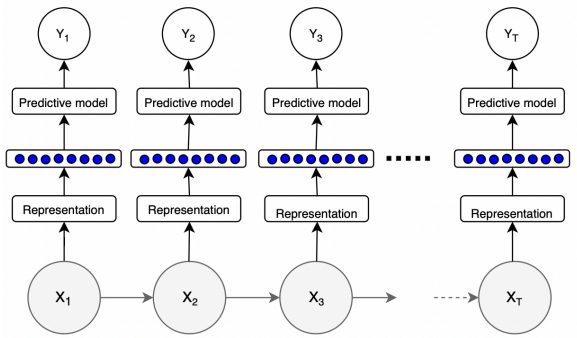
ظاهر فیض رُوح القدس ار باز مـــــدد فرماید دیـــگران هم بکنند آن چه مَسیحا می کرد

http://language.ml/

موجودیت‌های نامدار

- PER (Person شخص): «مولا امیرالمومنین علیه السلام»
- LOC (Location مکان): «حافظیه شیراز»
- ORG (Organization موسسه): «دانشگاه صنعتی شریف»
- GPE (Geo-Political Entity موقعیت ژئوپولیتیک): «جمهوری اسلامی ایران»

- معمولا چند کلمه‌ای است.
- شامل موارد متعدد دیگری نیز می‌شود:
  - زمان، تاریخ، قیمت



چالشهای موجودیت‌های نامدار

- دشواریهای تقطیع

- ابهام موجودیت‌های نامدار
  - ایران واکسن بیشتری تولید کرد.
  - مسابقات جام ملت‌های آسیا در ایران برگزار خواهد شد.

تگ‌های BIO

شیخ	B- PER
بهایی	I-PER
معمار	O
و	O
دانشمند	O
برجسته	O
ایران	B-GPE
در	O
...	...

دیتاست‌های فارسی

پیما (۷۱۴۵ جمله)

Label	#
Organization	16964
Money	2037
Location	8782
Date	4259
Time	732
Person	7675
Percent	699

آرمان (۷۶۸۲ جمله)

Label	#
Organization	30108
Location	12924
Facility	4458
Event	7557
Product	4389
Person	15645

https://hooshvare.github.io/docs/datasets/ner