

بسم الله الرحمن الرحيم

پردازش زبانهای طبیعی

جلسه ۴

احسان الدین عسگری

اسفند ۱۴۰۲

<http://language.ml/>

دانشکده مهندسی کامپیوتر

آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

asgari@berkeley.edu



جزء‌واژگی (meronymy) در مقابل holonymy

متفاوت

نوشتار

جزئی از یک شیء (یا برعکس)

معنا

- چرخ - دوچرخه (meronymy)

- لپتاپ - نمایشگر (holonymy)

- روابط واژگانی
 - فرم نرمال
 - ریشه و ظاهر
 - خطای نگارش
 - همنامی
 - چندمعنایی
 - هممعنایی
 - متضاد
 - زیرشمولی
 - جزء‌واژگی

کنایه، مجاز، دَگر نامی (metonymy)

متفاوت

نوشتار

معنای دیگر

معنا

— مذاکراتِ تهران و ۱+۵

— بر آستان جانان گر سر توان نهادن — گلبانگ سر بلندی بر آسمان توان زد

• روابط واژگانی

— فرم نرمال

— ریشه و ظاهر

— خطای نگارش

— هم‌نامی

— چندمعنایی

— هم‌معنایی

— متضاد

— زیرشمولی

— جزء‌واژگی

— کنایه

• روابط واژگانی

- فرم نرمال

- ریشه و ظاهر

- خطای نگارش

- هم‌نامی

- چندمعنایی

- هم‌معنایی

- متضاد

- زیرشمولی

- جزء‌واژگی

- کنایه

wordnet -

WordNet

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	<i>breakfast</i> ¹ → <i>meal</i> ¹
Hyponym	Subordinate	From concepts to subtypes	<i>meal</i> ¹ → <i>lunch</i> ¹
Instance Hyponym	Instance	From instances to their concepts	<i>Austen</i> ¹ → <i>author</i> ¹
Instance Hyponym	Has-Instance	From concepts to concept instances	<i>composer</i> ¹ → <i>Bach</i> ¹
Member Meronym	Has-Member	From groups to their members	<i>faculty</i> ² → <i>professor</i> ¹
Member Holonym	Member-Of	From members to their groups	<i>copilot</i> ¹ → <i>crew</i> ¹
Part Meronym	Has-Part	From wholes to parts	<i>table</i> ² → <i>leg</i> ³
Part Holonym	Part-Of	From parts to wholes	
Substance Meronym		From substances to their subparts	
Substance Holonym		From parts of substances to whole substances	
Antonym		Semantic opposition between lemmas	
Derivationally Related Form		Lemmas w/same morphological roots	

Wordnet	Lang	Synsets	Words	Senses	Core	Licence	Data	Citation
Albanet	als	4,673	5,988	9,599	31%	CC BY 3.0	als.zip (+xml)	cite:als; (bib)
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335	47%	CC BY SA 3.0	arb.zip (+xml)	cite:arb; (bib)
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,720	8,936	99%	CC BY 3.0	bul.zip (+xml)	cite:bul; (bib)
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%	wordnet	cmn.zip (+xml)	cite:cmn; (bib)
Chinese Wordnet (Taiwan)	qcn	4,913	3,206	8,069	28%	wordnet	qcn.zip (+xml)	cite:qcn; (bib)
DanNet	dan	4,476	4,468	5,859	81%	wordnet	dan.zip (+xml)	cite:dan; (bib)
Greek Wordnet	ell	18,049	18,227	24,106	57%	Apache 2.0	ell.zip (+xml)	cite:ell; (bib)
Princeton WordNet	eng	117,659	148,730	206,978	100%	wordnet	eng.zip (+xml)	cite:eng; (bib)

Open Multilingual Wordnet

Hebrew Wordnet	heb	5,448	5,325	6,872	27%	wordnet	heb.zip (+xml)	cite:heb; (bib)
Croatian Wordnet	hrv	23,120	29,008	47,900	100%	CC BY 3.0	hrv.zip (+xml)	cite:hrv; (bib)
IceWordNet	isl	4,951	11,504	16,004	99%	CC BY 3.0	isl.zip (+xml)	
MultiWordNet	ita	35,001	41,855	63,133	83%	CC BY 3.0	ita.zip (+xml)	cite:ita; (bib)
ItalWordnet	ita	15,563	19,221	24,135	48%	ODC-BY 1.0	ita.zip (+xml)	cite:iwn (bib)
Japanese Wordnet	jpn	57,184	91,964	158,069	95%	wordnet	jpn.zip (+xml)	cite:jpn; (bib)
Multilingual Central Repository	cat	45,826	46,531	70,622	81%	CC BY 3.0	cat.zip (+xml)	cite:cat; (bib)
Multilingual Central Repository	eus	29,413	26,240	48,934	71%	CC BY 3.0	eus.zip (+xml)	cite:eus; (bib)

<http://wordnetweb.princeton.edu/perl/webwn>

WordNet

synset	gloss
mark, grade, score	a number or letter indicating quality
scratch, scrape, scar, mark	an indication of damage
bell ringer, bull's eye, mark, home run	something that exactly succeeds in achieving its goal
chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug	a person who is gullible and easy to take advantage of
mark, stigma, brand, stain	a symbol of disgrace or infamy

• روابط واژگانی

- فرم نرمال

- ریشه و ظاهر

- خطای نگارش

- هم‌نامی

- چندمعنایی

- هم‌معنایی

- متضاد

- زیرشمولی

- جزء‌واژگی

- کنایه

wordnet -

WordNet

Wordnet	Lang	Synsets	Words	Senses	Core	Licence	Data	Citation
Albanet	als	4,675	5,988	9,599	31%	CC BY 3.0	als.zip (+xml)	cite:als; (.bib)
Arabic WordNet (AWN v2)	arb	9,916	17,785	37,335	47%	CC BY SA 3.0	arb.zip (+xml)	cite:arb; (.bib)
BulTreeBank Wordnet (BTB-WN)	bul	4,959	6,720	8,936	99%	CC BY 3.0	bul.zip (+xml)	cite:bul; (.bib)
Chinese Open Wordnet	cmn	42,312	61,533	79,809	100%	wordnet	cmn.zip (+xml)	cite:cmn; (.bib)
Chinese Wordnet (Taiwan)	qcn	4,913	3,206	8,069	28%	wordnet	qcn.zip (+xml)	cite:qcn; (.bib)
DanNet	dan	4,476	4,468	5,859	81%	wordnet	dan.zip (+xml)	cite:dan; (.bib)
Greek Wordnet	ell	18,049	18,227	24,106	57%	Apache 2.0	ell.zip (+xml)	cite:ell; (.bib)
Princeton WordNet	eng	117,659	148,730	206,978	100%	wordnet	eng.zip (+xml)	cite:eng; (.bib)
Open Multilingual Wordnet								
Hebrew Wordnet	heb	5,448	5,325	6,872	27%	wordnet	heb.zip (+xml)	cite:heb; (.bib)
Croatian Wordnet	hrv	23,120	29,008	47,900	100%	CC BY 3.0	hrv.zip (+xml)	cite:hrv; (.bib)
IceWordNet	isl	4,951	11,504	16,004	99%	CC BY 3.0	isl.zip (+xml)	
MultiWordNet	ita	35,001	41,855	63,133	83%	CC BY 3.0	ita.zip (+xml)	cite:ita; (.bib)
ItalWordnet	ita	15,563	19,221	24,135	48%	ODC-BY 1.0	ita.zip (+xml)	cite:iwn (.bib)
Japanese Wordnet	jpn	57,184	91,964	158,069	95%	wordnet	jpn.zip (+xml)	cite:jpn; (.bib)
Multilingual Central Repository	cat	45,826	46,531	70,622	81%	CC BY 3.0	cat.zip (+xml)	cite:cat; (.bib)
Multilingual Central Repository	eus	29,413	26,240	48,934	71%	CC BY 3.0	eus.zip (+xml)	cite:eus; (.bib)

<https://omwn.org/>

• روابط واژگانی

- فرم نرمال

- ریشه و ظاهر

- خطای نگارش

- هم‌نامی

- چندمعنایی

- متضاد

- زیرشمولی

- جزء‌واژگی

- کنایه

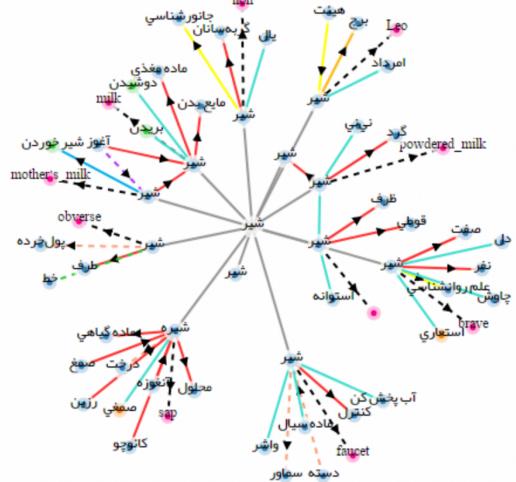
wordnet -

WordNet فارسی

آمار فارس نت

این جدول شامل آمار تفکیکی از کلیه اطلاعات موجود در شبکه واژگانی فارسنت می باشد :

#	نوع داده	فارسنت ۱	فارسنت ۲	فارسنت ۳	۲.۵	فارسنت ۴
۱	Words	۱۷۸۴۶	۳۰۲۲۲	۳۳۲۹۰	۱۰۰۰۰	۱۰۰۰۰
۲	Word–Senses	۲۴۴۸۰	۳۶۱۱۵	۳۹۷۳۵	بیش از	۱۰۰۰۰
۳	Sense–Relations	۳۶۰	۷۰۴۳	۱۹۰۲۱	بیش از	۳۰۰۰۰
۴	Synsets	۱۰۰۱۴	۱۹۳۹۸	۲۰۵۵۹	بیش از	۴۰۰۰۰
۵	Synset–Relations	۶۹۸۰	۳۶۸۴۸	۴۷۷۶۱	بیش از	۹۰۰۰۰
۶	Mapinngs to PWN	۶۹۵۰	۱۸۵۸۴	۱۹۳۵۷	بیش از	۳۰۰۰۰
۷	Words per Synset	۱.۷۸	۱.۰۵۶	۱.۶۲	بیش از	۲.۵
۸	Senses per Word	۱.۳۷	۱.۱۹	۱.۱۹	بیش از	۱.۵



<http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Statistics.jsp>

- روابط واژگانی
 - فرم نرمال
 - ریشه و ظاهر
 - خطای نگارش
 - همنامی
 - چندمعنایی
 - هممعنایی
 - متضاد
 - زیرشمولی
 - جزءواژگی
 - کنایه

WordNet

• روابط واژگانی

- فرم نرمال

- ریشه و ظاهر

- خطای نگارش

- هم‌نامی

- چندمعنایی

- هم‌معنایی

- متضاد

- زیرشمولی

- جزء‌واژگی

- کنایه

wordnet -

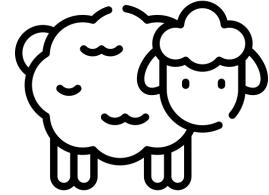
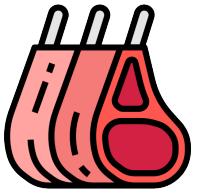
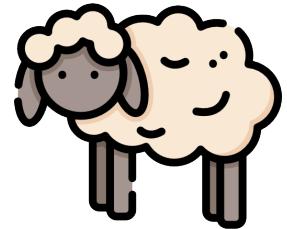
```
function SIMPLIFIED LESK(word, sentence) returns best sense of word
    best-sense  $\leftarrow$  most frequent sense for word
    max-overlap  $\leftarrow$  0
    context  $\leftarrow$  set of words in sentence
    for each sense in senses of word do
        signature  $\leftarrow$  set of words in the gloss and examples of sense
        overlap  $\leftarrow$  COMPUTEOVERLAP(signature, context)
        if overlap > max-overlap then
            max-overlap  $\leftarrow$  overlap
            best-sense  $\leftarrow$  sense
    end
    return(best-sense)
```

Figure 18.10 The Simplified Lesk algorithm. The COMPUTEOVERLAP function returns the number of words in common between two sets, ignoring function words or other words on a stop list. The original Lesk algorithm defines the *context* in a more complex way.

پیش پردازش



<http://language.ml/>



چرا پیش پردازش؟



?

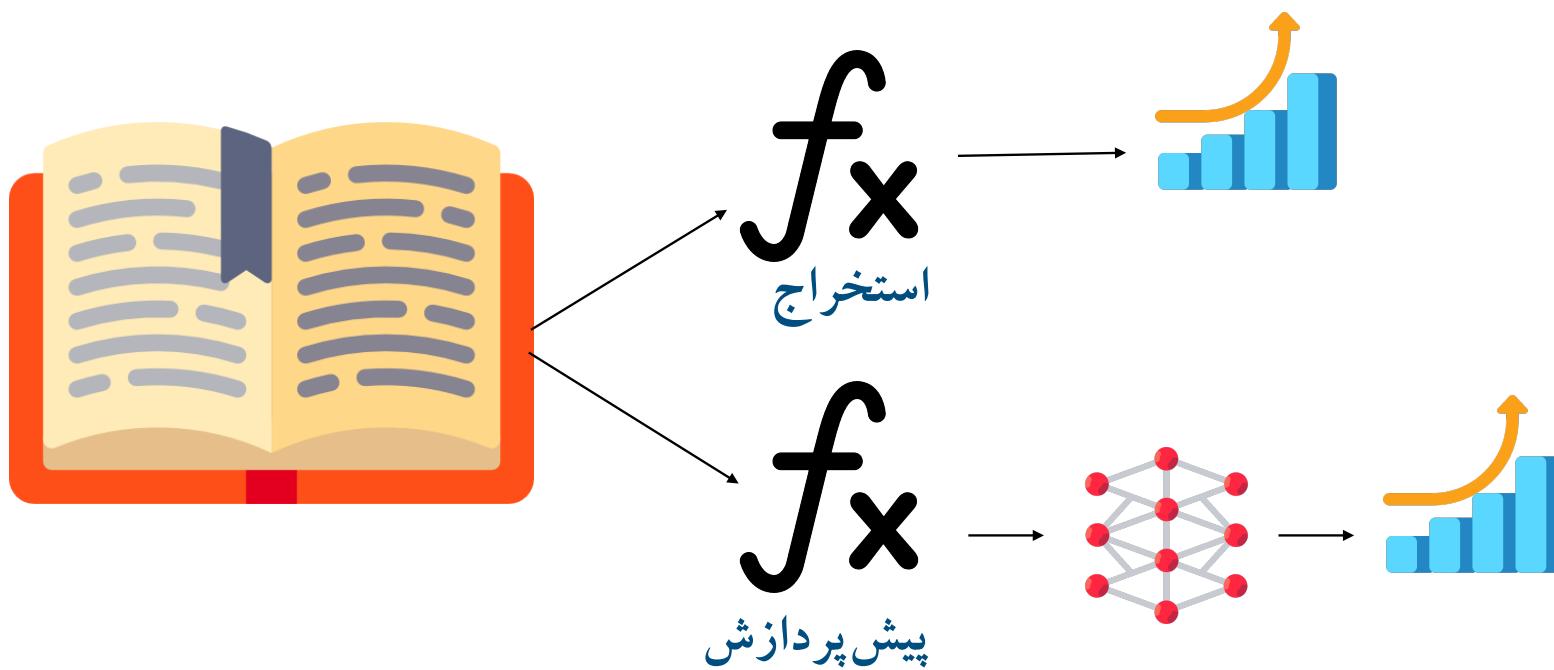


پیش پردازش در نگاه کلی یک استخراج اطلاعات است



استخراج اطلاعات در مقابل پیش‌پردازش

استخراج اطلاعات ساختار یافته از متن‌های بدون ساختار



كلمات زائد

بر اساس لیست لغات

– معمولاً با تحلیل بر اساس فرکانس –

— وابسته په مساله —

```
In [3]: from nltk import FreqDist
mp_freqdist = FreqDist(modified_text) # compute the frequency distribution
mp_freqdist.most_common(50) # show the top 50 (word, frequency) pairs

Out[3]: [(63067, 'و'), (47872, 'از'), (46694, 'کی'), (42763, 'به'), (36289, 'در'), (32266, 'تی'), (23587, 'ز'), (23171, 'را'), (18204, 'من'), (18154, 'بر'), (16817, 'بل'), (15968, 'ان'), (14924, 'با'), (13957, 'این'), (13552, 'چ'), (11375, 'او'), (11157, 'سر'), (10692, 'جان'), (9497, 'جو'), (9070, 'پین'), (8984, 'است'), (8589, 'هر'), (8287, 'تی'), (7787, 'یا')]
```

مراحل پیش پردازش

یکسانسازی‌های متنی به فرم کانونی آن که در معنا تغییری ایجاد نمی‌کند و فقط یک تغییر ظاهری است.

نمایش مالایی

دیگران هم بکنند آن چه مسیحا می‌کرد

ظاهر

فیض رُوح القدس ارباز مدد فرماید

بزرگ و کوچکی و حالات مختلف حروف

- برای بازیابی اطلاعات مفید
- اما برای
- تحلیل احساس
- زبان‌هایی مثل زبان آلمانی
- ترجمه
- استخراج اطلاعات