

بسم الله الرحمن الرحيم

پردازش زبانهای طبیعی

جلسه ۲

احسان الدین عسگری

بهمن ۱۴۰۲

<http://language.ml/>

دانشکده مهندسی کامپیوتر

آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

asgari@berkeley.edu

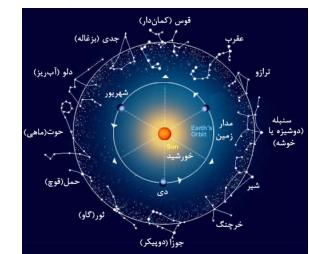
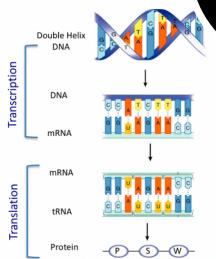


تعریف زبان

Chomsky (1959: 137) “*A language is a collection of sentences of finite length all constructed from a finite alphabet (or, where our concern is limited to syntax, a finite vocabulary) of symbols.*”



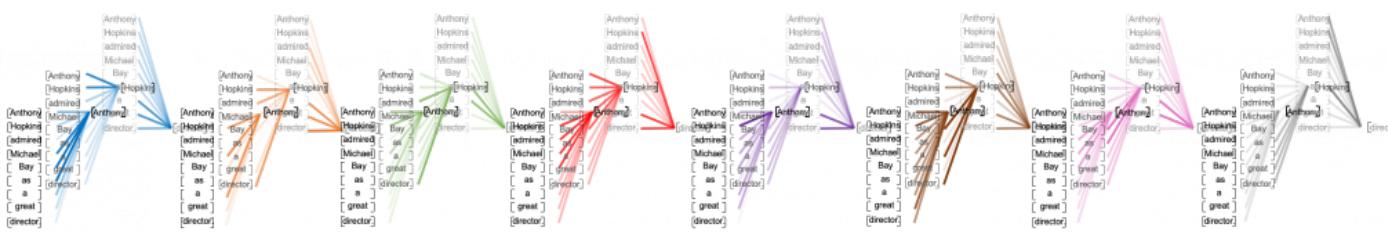
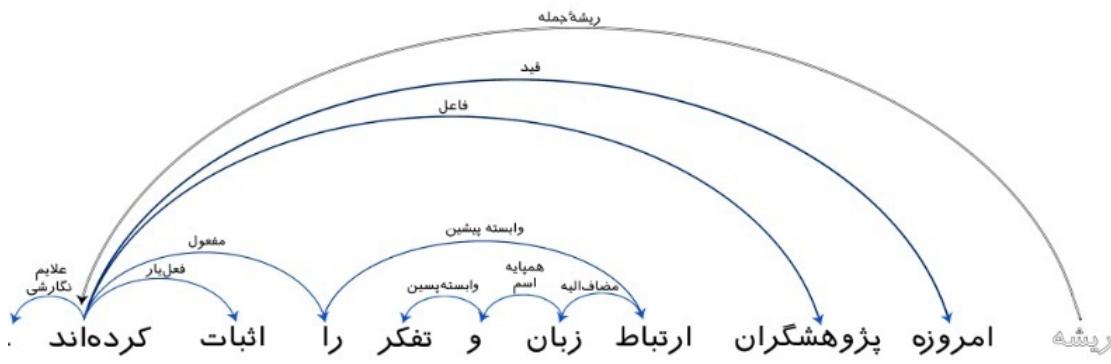
مجموعه توصیفاتی از حقایق: توصیفاتی در محور زمان از المان‌های تکرار پذیر که با هم ارتباطات تنگاتنگ ساختاری و معنایی دارند: ارتباطات درختی و شبکه‌ای.



تعريف زبان



- مثال ارتباطات درختی و شبکه‌ای



گو

زبانهای ایرانی

Persian & dialects	58%
Azeri & dialects	26%
Kurdish	9%
Luri	2%
Balochi	1%
Arabic	1%
others	3%

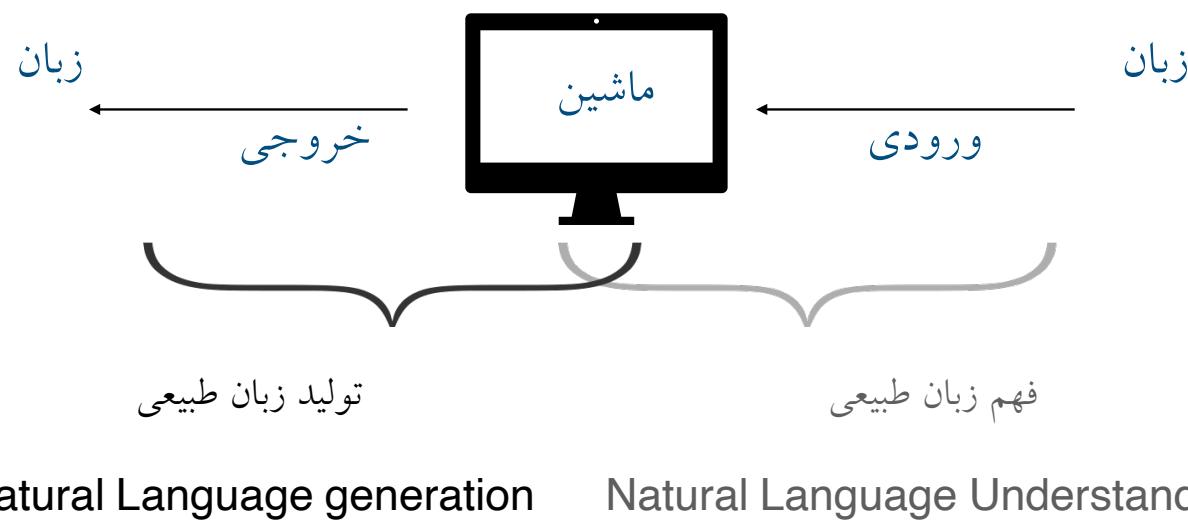


گویا

— پردازش زبان طبیعی



پردازش زبان طبیعی (NLP)



چگونه فهمی؟ پردازشی که به نمایش قابل فهم برای ماشین بیانجامد.



پردازش زبان طبیعی (NLP)

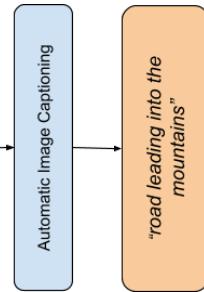
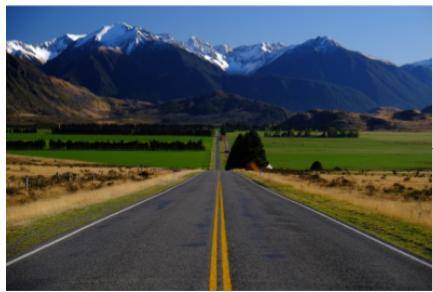
- مطالعه علمی زبان با دید رایانشی
- کار با مدل‌های محاسباتی پدیده‌های مختلف زبانی با انگیزه علمی یا فنی.

www.aclweb.org

- فرق میان NLP و Computational Linguistics ؟



کاربردهای NLP



Contextual Spelling Error

Deer Mr. Theodore: **Spelling Error**

I am exceeedingly interested in this po
and employment background are appri

While working toward my degree, I wa
small firm. I increased my call volume
success. I will completes my degree in
employment in early June.

Grammar Error

A screenshot of a text editor window. It shows a letter addressed to "Mr. Theodore" with several errors highlighted: "Deer" (spelling), "exceeedingly" (spelling), "po" (likely "position"), and "completes" (spelling). Below the letter, a sentence about employment is shown with a "Grammar Error" callout pointing to the word "in".

- معرفی زبان پردازش زبان طبیعی
- درس NLP
- جلسه بعد

• معرفی

• زبان

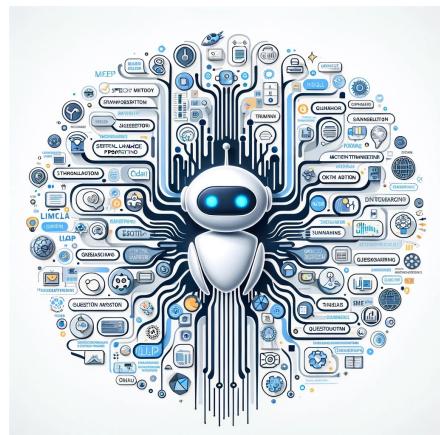
پردازش زبان طبیعی •

NLP درس •

جلسه بعد •

NLP سایر مسائل کاربردی

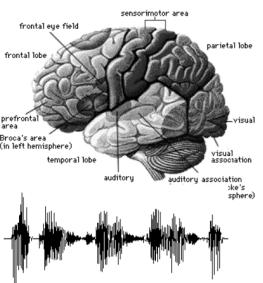
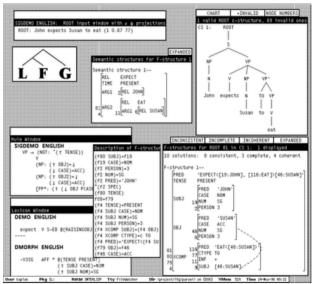
- Spelling Suggestions
- Grammar Checking
- Synonym Generation
- Information Extraction
- Text Categorization
- Summarization
- Essay Scoring
- Automated Customer Service
- Text generation
- Machine Translation
- Question Answering
- Improving Web Search Engine results
- Automated Metadata Assignment
- Online Dialogs



- معرفی زبان
- پردازش زبان طبیعی
- درس NLP
- جلسه بعد

پردازش زبان طبیعی - زمینه میان رشته‌ای

- علوم کامپیوتر
- یادگیری ماشین: مدل‌ها آماری یا شبکه‌های عصبی
- زبان‌شناسی
- علوم انسانی و اجتماعی
- رشته‌های هم مرز
- زبان‌شناسی
- پردازش گفتار (صوت)
- علوم شناختی
- یادگیری ماشین



چالش‌های پردازش زبان و متن

- معرفی
- زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

○ پدیده‌ی بسیار پیچیده

○ ابهام در لایه‌های مختلف

۱) درسطح تشخیص گفتار – 

موذنا از آن بگو یا موذنا اذان بگو

تشخیص گفتار: تشخیص کلمات تلفظ شده



چالش‌های پردازش زبان و متن

- ابهام در لایه‌های مختلف
 - پدیده‌ی بسیار پیچیده
- ۲) در سطح صرف - 

که با این درد اگر دربند درمانند درمانند

صرف: علم چگونگی ساخت واژگان

- معرفی
- زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

چالش‌های پردازش زبان و متن

- ابهام در لایه‌های مختلف
 - پدیده‌ی بسیار پیچیده
- ۳) سطح نحو – 

و آن دگر شیر است کَادم می خورد..

نحو: دستور زبان

- معرفی
- زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

چالش‌های پردازش زبان و متن

- پدیده‌ی بسیار پیچیده
 - ابهام در لایه‌های مختلف
- ۳) سطح معناشناسی - 

ای دمت عیسی دم از دوری مزن / من غلام آن که دوراندیش نیست

معناشناسی: بررسی معنای کلمات و عبارات در زبان.

- معرفی
- زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

چالش‌های پردازش زبان و متن

- پدیده‌ی بسیار پیچیده
 - ابهام در لایه‌های مختلف
- ٤) در سطح گفتمان - 

وَعَلَمَ آدَمَ الْأَسْمَاءَ كُلَّهَا. ثُمَّ عَرَضَهُمْ عَلَى الْمَلَائِكَةِ. فَقَالَ أَنْبِئُونِي بِاسْمَاءِ هَؤُلَاءِ إِنْ كُنْتُمْ صَادِقِينَ

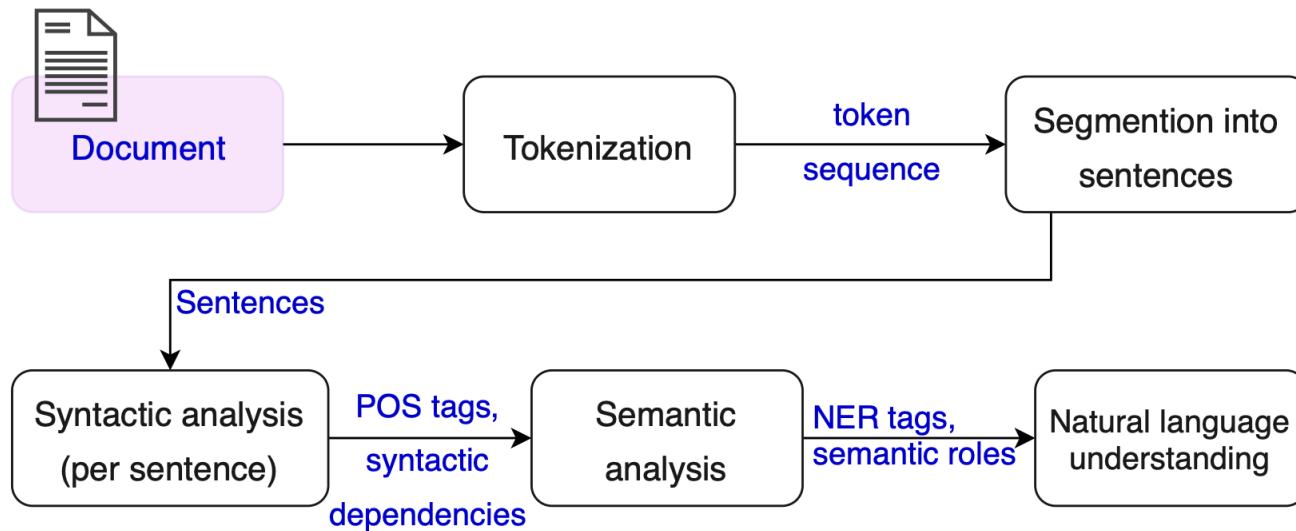
گفتمان: بررسی معنا در بیش از یک جمله.

- معرفی
- زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

- معرفی زبان

- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

NLP خط‌لوله (پایپلاین) سنتی



NLP مسائل پایه‌ای

	در سطح کلمات	در سطح عبارات	در سطح جمله	در سطح متن
(Segmentation) تقطیع	(tokenisation) تقطیع کلمات	تقطیع عبارات (chunking)	(Sentence Boundary Detection) تشخیص جملات	بخش‌بندی متن (TextTiling)
(Syntax) دستور زبان	ساخت‌شناسی لغت Morphology و Stemming و Lemmatization	استخراج اطلاعات (Information extraction)	پارسینگ (parsing)	پارسینگ ساختار متن
(Semantics) معناشناسی	شباهت کلمات Word Similarity	استخراج اطلاعات (Information extraction)	تشخیص احساس و نظر پاسخ به پرسش رفع ابهام کلمه	خلاصه‌سازی متن ترجمه ماشینی

- معرفی
- زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

• معرفی زبان

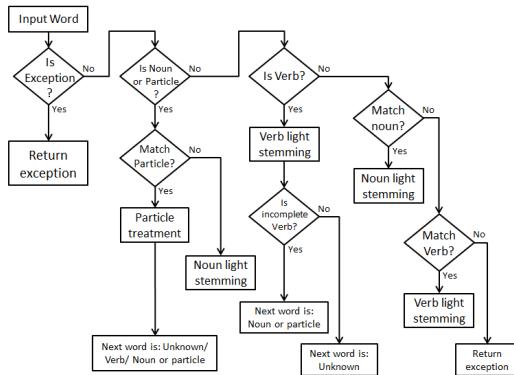
• پردازش زبان طبیعی

• درس NLP

• روابط واژگانی

NLP روشهای

- Heuristic



- Rule-based

```
VALID_EMAIL_REGEX = /\A[\w+\.-]+@[a-z\d\.-]+\.[a-z]+\z/i
validators_email = [
    {name: "email", regexp: VALID_EMAIL_REGEX, validate: true, length: {maximum: 255}, REGEX },
    {name: "url", regexp: /https?:\/\/(www\.)?[-\w\d\.]+\.\w+([-|\w\d])?/, validate: true, length: {maximum: 255}, REGEX },
    {name: "ip", regexp: /\b((25[0-5]|2[0-4][0-9]|1[0-9]{2}|[1-9]?[0-9])\.(25[0-5]|2[0-4][0-9]|1[0-9]{2}|[1-9]?[0-9])\.(25[0-5]|2[0-4][0-9]|1[0-9]{2}|[1-9]?[0-9])\.(25[0-5]|2[0-4][0-9]|1[0-9]{2}|[1-9]?[0-9])\b/, validate: true, length: {maximum: 255}, REGEX }
]

RegExp: \A[\w+\.-]+@[a-z\d\.-]+\.[a-z]+\z
Sample: example@jetbrains.com|
```

NLP روشهای

- Probabilistic models
- Naive Bayes, Logistic regression, HMM, CRF

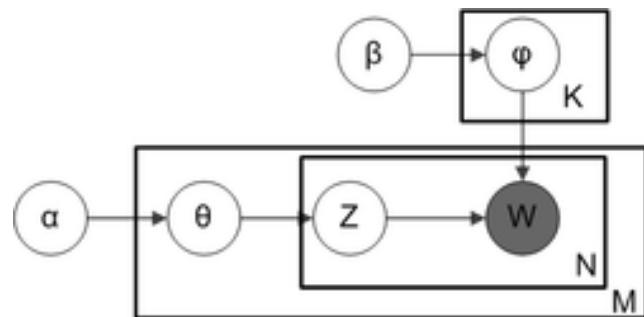
$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{\sum_y P(Y = y)P(X = x|Y = y)}$$

- معرفی زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

NLP روشهای

- معرفی
- زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

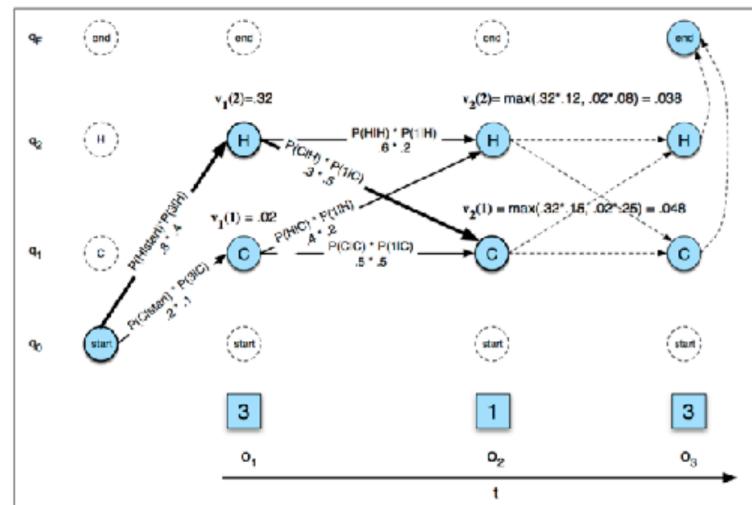
- Latent variable models (specifying probabilistic structure between variables and inferring likely latent values)



- معرفی زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

NLP روشهای

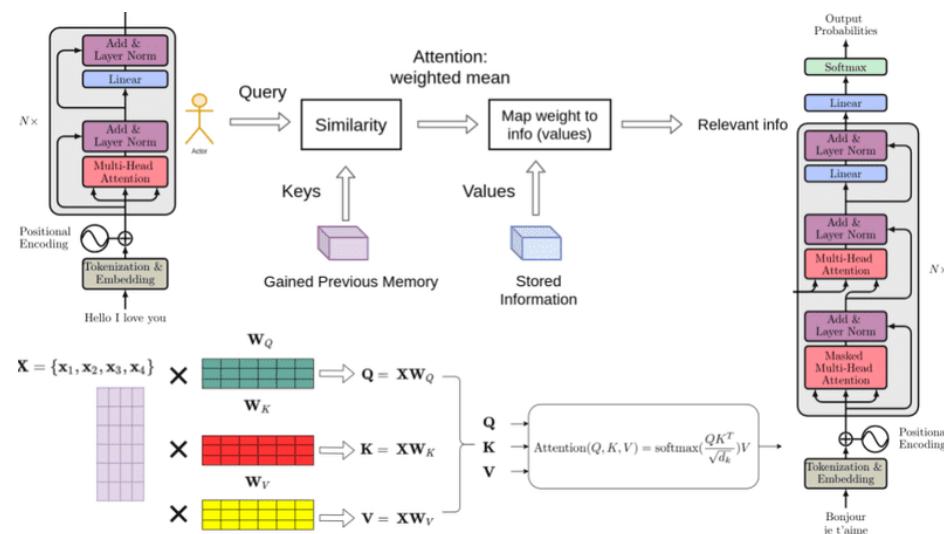
- Dynamic programming
- Viterbi algorithm, CKY



NLP روشهای

- معرفی زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

- Deep learning approaches



NLP - درس

آدرس سایت درس

<https://language.ml/courses/nlp14022/index.html>



هدف‌گیری درس

- آشنایی با مسائل مهم پردازش زبان طبیعی
- تجربه عملی کار با داده‌های متنی و پیش‌پردازشها
- تئوری روش‌های اصلی SOTA
- پرداختن به پردازش زبان‌های ایرانی و مشارکت در توسعه بنچمارک
- در پروژه‌ها) پرداختن به مسائل نو و کاربردی



Introduction to NLP challenges	Language	HW1- Text	HW2-Parsi-IO Bot
	Challenges of Language Processing		
	Lexical Relations - Word Net		
Preprocessing / Rule-based NLP	NLP Preprocessing		
	Rule-based NLP		
Statistical Language Model and Distributional Semantics	Tokenization	HW3 - Text/Token Classification	HW4 - Text Generation
	n-gram language model		
Word Vectors to Multi-head Attention	Deep/Representation learning (4)		
	Attention mechanism		
	Transformer		
Encoder Transformers and Fine-tuning	Encoder model		
	Classification/Token Classification model (2)		
Decoder Transformers and Prompt tuning	Decoder model		
	Prompt Tuning		
Encoder-Decoder Transformers	Encoder-decoder model		
	Tranlation models (2)		
Advanced NLP	PEFT (2)	Starting the final project including advanced topic	
	Multilingual NLP		
	Multimodal NLP		



پیش‌نیاز تمرینهای عملی

- Python3
- jupyter notebook / colab.
- Pytorch
- Huggingface
- nltk/spacy
- scikit-learn

- معرفی زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

کوئرا

<https://quera.org/course/16769/>
رمز nlpsp4023

پردازش زبان طبیعی

احسانالدین عسگری

دستیاران آموزشی

امید قهرودی

سارا کریمی

محمدسینا پاکسرشت

ریحانه زهراei

حمیدرضا امیرزاده

محراب مرادزاده

علی درخشش

abolfazl malekahmadi

محمدحسین سامتی

سارا آذرنوش

پر迪س زهراei

ahmadian_hadis

Seyedemad Zolhavarieh

• معرفی

• زبان

• پردازش زبان طبیعی

• درس NLP

• روابط واژگانی

- معرفی زبان
- پردازش زبان طبیعی
- درس NLP
- روابط واژگانی

انتخاب متن دلخواه برای پروژه بررسی متن

