

بسم الله الرحمن الرحيم

# پردازش زبانهای طبیعی

جلسه ۳

احسان الدین عسگری

بهمن ۱۴۰۲

<http://language.ml/>

دانشکده مهندسی کامپیوتر

آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

asgari@berkeley.edu



|   |   |   |                       |
|---|---|---|-----------------------|
| 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                                |   |                       |



# ارزیابی

۱۲۰ نمره (برای تمرین‌ها)

۴ نمره پروژه پایانی

۲۰ نمره کوییزها (از هر مژول)

۲۰ نمره پایان‌ترم

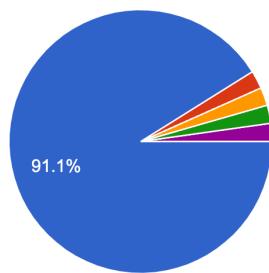
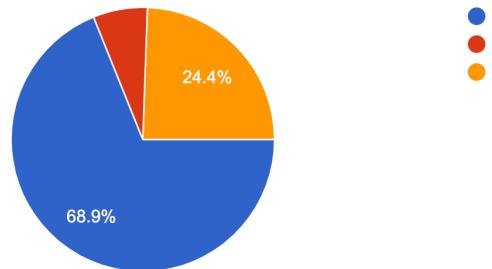
۰ تا سقف ۱ نمره حضور فعال و مستمر در کلاس

۰ تا سقف ۰.۵ نمره ارائه اختیاری صبحها

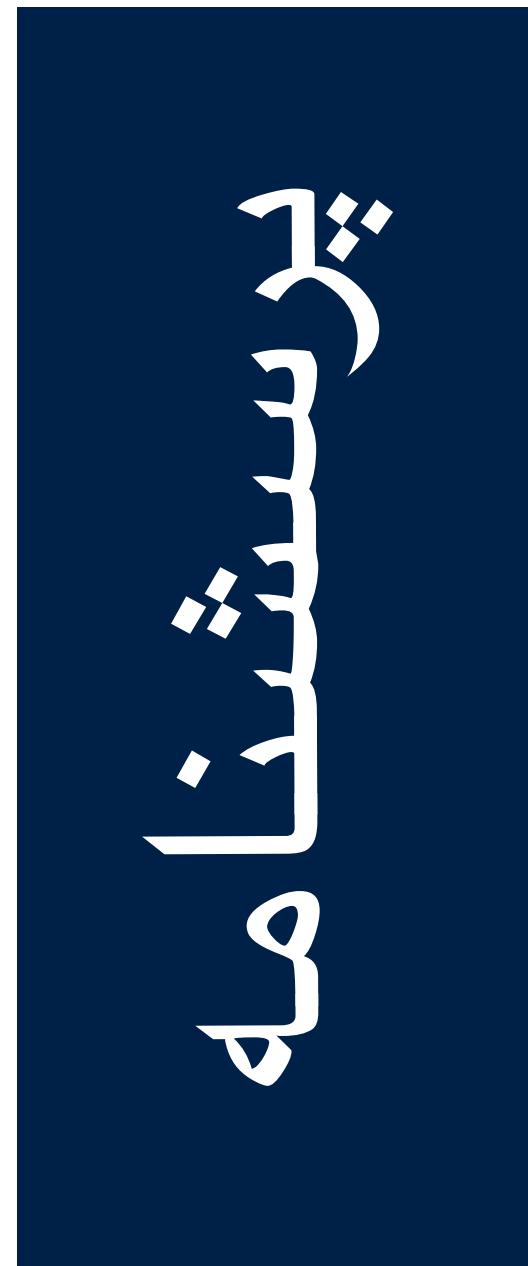
- البته توجه داشته باشید که ممکن است نمره‌های برخی بخش‌ها در انتهای ترم در نرمال‌سازی تغییراتی داشته باشند.



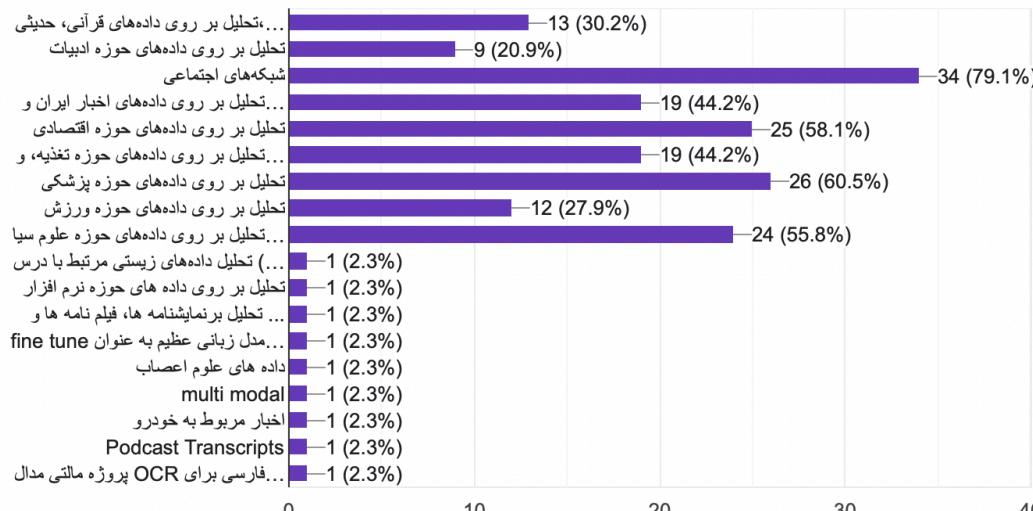
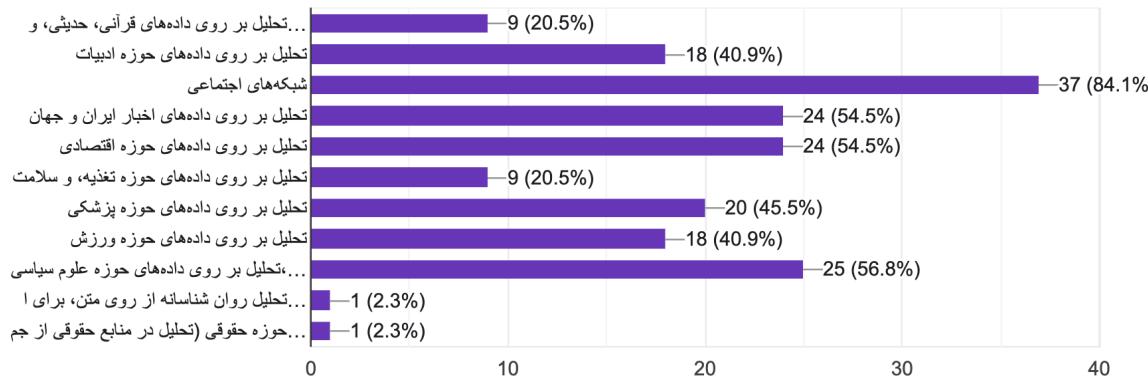
# مقطع و رشته تحصیلی



- مهندسی کامپیوٹر
- مهندسی برق
- ریاضی و علوم کامپیوٹر
- مهندسی مکانیک
- مهندسی عمران - مهندسی حمل و نقل



# نوع داده مورد علاقه



زبان

glk Gilaki  
mzn Mazandarani  
azb South Azerbaijani  
lrc Northern Luri  
psc Iranian Sign Language  
vmh Maraghei  
ntz Natanzi  
nyq Nayini  
sdh Southern Kurdish  
bqi Bakhtiâri  
Sistani  
Behbahani  
Birjandi



# توقعات از درس

آشنایی عمیق با حوزه پردازش زبان طبیعی

انجام پروژه‌های عملی و تحقیقاتی در این زمینه

آشنایی با روش‌های استفاده و تنظیم LLM‌ها

آشنایی با مدل‌های مالتی مدل

آمادگی برای پژوهش



• اسلایدها و ساختار درس در بخش‌هایی الهام گرفته و بومی‌سازی شده و بعضی با استفاده از مطالب دروس زیر است:

- Marti Hearst - Applied NLP Class / UC Berkeley
- David Bamman - NLP Class / UC Berkeley
- Christopher Manning - Deep NLP Class / Stanford
- Noah Smith - NLP Class / CMU
- Michael Collins - NLP Class / Columbia University

• کتابها

- Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd ed. Draft)
- Manning and Schuetze, Foundations of Statistical Natural Language Processing
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning



# انتخاب متن دلخواه برای پروژه بررسی متن



بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

- روابط واژگانی (lexical rel.)



کلمہ ۱

?

---

==

کلمہ ۲

یکسان

متفاوت

متفاوت

پکستان

نوشتار

معنا



- روابط واژگانی
  - فرم نرمال

## فرم نرمال شده در مقابل فرمهای گوناگون

مشابه

یکسان (از نظر کاربرد)

نوشتار

معنا

○ خانه‌ی دوست - خانه دوست

○ امریکا - آمریکا

○ <NUM> ۱۱۰ - صد و ده

○ فینگلیش

# ریشه (lemma) در مقابل ظاهر کلمه (surface form)

- روابط واژگانی
  - فرم نرمال
  - ریشه و ظاهر

معمولًا دارای اشتراک

دارای اشتراک

نوشتار

معنا

درخت - درختان

حمد - محمد

taught - teach

## فاصله ویرایشی

• روابط واژگانی

- فرم نرمال

- ریشه و ظاهر

- خطای نگارش

معمولاً دارای اشتراک

متفاوت - بی معنا

معنا

نوشتار

○ مدرسه - مساله

○ مشقات - مشکلات

○ صابون - سابون

هم آوا

Homophone

یکسان

تلفظ

یکسان

نوشتار

هم نام (homonym)

متفاوت

معنا

روابط واژگانی

- فرم نرمال

- ریشه و ظاهر

- خطای نگارش

- هم نامی

○ خوان - خان

○ شیر<sup>۱</sup> - شیر<sup>۲</sup>

○ مُلک - ملک

○ آهو<sup>۱</sup> - آهو<sup>۲</sup>

تشابه غیرتصادفی

یکسان

نوشتار و لفظ

متفاوت

معنا

ساعت<sub>۱</sub> - ساعت<sub>۲</sub> ○

دلتا<sub>۱</sub> - دلتا<sub>۲</sub> ○

دستزدن<sub>۱</sub> - دست زدن<sub>۲</sub> ○

## چند معنایی (polysemy)

• روابط واژگانی

- فرم نرمال

- ریشه و ظاهر

- خطای نگارش

- همنامی

- چندمعنایی

## هم معنایی (synonymy)

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

متفاوت

یکسان

نوشتار

معنا

عظیم - بزرگ

○

خانه - منزل

○

در - باب

○

Water - H<sub>2</sub>O

○

هم‌معنایی مطلق؟

شرط هم‌معنایی: به کار رفتن واژه هم معنا در تمام سیاق‌ها

## متضاد (antonyms)

متفاوت

مقابل هم از یک منظر

نوشتار

معنا

- کلماتی که در یک ویژگی معنایی مقابله هم هستند.

- ولی بسیار شبیه هم هستند.

کوتاه - بلند

○

کند - سریع

○

مرتفع - پست

○

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

## زیرشمولی (hypernymy) – فراشمول (hyponymy)

متفاوت

زیرشمول یا فراشمول

نوشتار

معنا

– وسیله نقلیه – اتوبوس

– انار – میوه

• روابط واژگانی

– فرم نرمال

– ریشه و ظاهر

– خطای نگارش

– هم‌نامی

– چندمعنایی

– هم‌معنایی

– متضاد

– زیرشمولی