

دانشکده علوم و فنون نوین

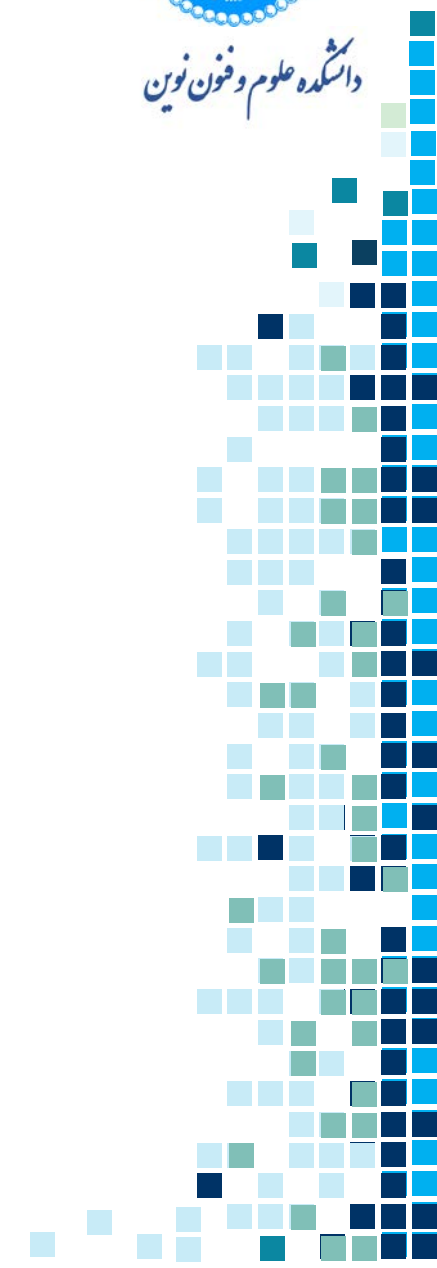
آشنایی با زبان‌شناسی رایانشی

Introduction to Computational Linguistics

هادی ویسی

h.veisi@ut.ac.ir

نیم‌سال دوم ۱۴۰۰-۱۴۰۱



معرفی درس . . .

□ زمان و مکان

- ◀ شنبه و دوشنبه، ساعت ۱۰:۰۰ الی ۱۱:۳۰
- ◀ در سامانه eLearn یا در دانشکده علوم و فنون نوین

□ وب سایت

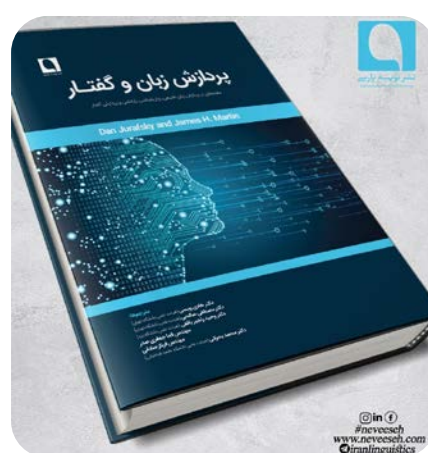
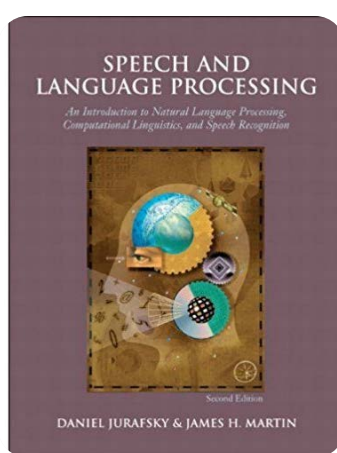
- ◀ dsp.ut.ac.ir

□ هدف

- ◀ مروری بر مفاهیم پایه زبان‌شناسی رایانشی و کاربردهای آنها
- ◀ مروری بر پردازش زبان طبیعی (متن و گفتار)
- ◀ فعالیت‌های تمرینی با رویکرد کاربردی

معرفی درس . . .

منابع □



◀ هادی ویسی، مصطفی صالحی، وحید رنجبر بافقی، الما جعفری صدر، فرناز صادقی، محمد بحرانی، پردازش زبان و گفتار: مقدمه‌ای بر پردازش زبان طبیعی، زبان‌شناسی رایانشی و پردازش گفتار، نویسه پرسی، ۱۴۰۰

Daniel Jurafsky, James Martin, Speech and Language Processing, 2nd Edition, Prentice Hall, 2009.

◀ James Allen, Natural Language Understanding, ۱۹۹۵.

◀ Ruslan Mitkov, The Oxford Handbook of Computational Linguistics, 2003.

◀ Igor Bolshakov, Alexander Gelbukh, Computational Linguistics, Models, Resources, Applications, 2004.

معرفی درس: ارزیابی . . .

□ تمرین (برای هر موضوع)

- ◀ همفکری و همکاری در یافتن پاسخ سوالها توصیه می‌شود
 - ◀ در صورت کپی بودن یکی یا چند مورد از پاسخها، کل نمره آن تمرین در نظر گرفته نمی‌شود.
 - ◀ تمرین‌های دارای پیاده‌سازی، باید هم شامل کدها و هم شامل گزارش مربوطه باشد
 - ◀ تاخیر در تحویل
 - ◀ ارسال پاسخ حداکثر تا ساعت ۲۳:۵۹ مهلت تعیین شده
 - ◀ هر یک ساعت تاخیر در ارسال پاسخها (از یک ثانیه تا ۶۰ دقیقه!)، کسر یک درصد نمره به عنوان جریمه تاخیر
 - ◀ ارسال پاسخ تمرینها
 - ◀ تنها به صورت الکترونیکی و به ایمیل استاد درس است.
 - ◀ همه فایل‌های مرتبط با یک تمرین را در یک فایل فشرده شده
 - ◀ فرمت نام‌گذاری فایل ارسالی: `CL_YourFamilyName_YourStNo_HW#`
- `CL_Ahmadi_830496001_HW1.rar` = پاسخ تمرین شماره ۱ توسط خانم/آقای احمدی با شماره دانشجویی ۰۰۱

◀ وزن تمرین‌های مختلف با هم برابر نیست



معرفی درس: ارزیابی . . .

□ آزمونک (کويز)

◀ در کلاس‌های غیرحضورى نداريم، در صورت حضورى بودن خواهيم داشت!

□ امتحان میان‌ترم

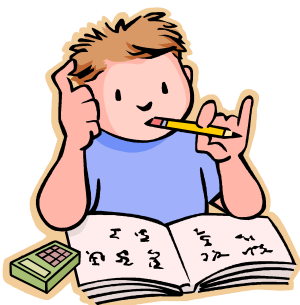
◀ مطالب تدریس شده تا تاریخ اعلام شده برای این امتحان

◀ زمان: دوشنبه ۱۴۰۱/۰۲/۰۵ ساعت ۱۰:۰۰

□ امتحان پایان‌ترم

◀ شامل کلیه مطالب تدریس شده، از جمله مطالب میان‌ترم

◀ زمان: مطابق برنامه دانشگاه



معرفی درس: ارزیابی . . .

□ پروژه (نمره اضافی)

- ◀ پروژه کاربردی دارای پیاده‌سازی در Python یا سایر زبان‌های برنامه‌نویسی
- ◀ عدم نیاز به نوآوری: پیاده‌سازی یک پژوهش (مقاله، پایان‌نامه و ...) موجود
- ◀ علاوه بر کد برنامه، گزارش مکتوب (به صورت تایپ شده)، داده‌ها و منابع هم تحویل گرفته می‌شود
- ◀ در صورت مساعد شدن شرایط کرونایی، تحویل به صورت حضوری است
- ◀ آخرین زمان پیشنهاد موضوع: روز دوشنبه ۱۴۰۱/۰۲/۰۵
- ◀ تحویل پروژه: اولین هفته بعد از آخرین امتحان پایان‌ترم
- ◀ موضوع الزاما مرتبط با مباحث درس باشد

◀ برخی موضوع‌های پیشنهادی

- ◀ پیاده‌سازی یک ریشه‌یاب برای فارسی
- ◀ برجسب‌زنی اجزای کلام فارسی با روش‌های یادگیری ماشین
- ◀ تحلیل احساس در متون فارسی



معرفی درس: ارزیابی . . .

□ حضور و مشارکت در کلاس (نمره اضافی)

◀ مشارکت فعال در بحث‌های کلاس

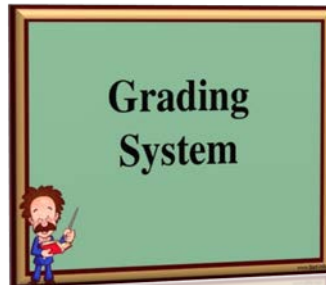
□ بازنگری نمره‌ها و برگه‌ها

◀ در زمان تحویل پروژه درس (به صورت حضوری)

◀ اولین دوشنبه بعد از آخرین امتحان پایان ترم



معرفی درس: ارزیابی . . .



عنوان	وزن	توضیح
تمرین	۵۵%	بعد از هر موضوع (وزن تمرین‌ها برابر نیست)
آزمونک (کويز)	؟	در شرایط غیرحضورى کلاس‌ها نداریم
امتحان میان‌ترم	۲۰%	دوشنبه ۱۴۰۱/۰۲/۰۵ ساعت ۱۰:۰۰
امتحان پایان‌ترم	۲۵%	از کل مطالب درس، مطابق برنامه دانشگاه در صورت داشتن آزمونک وزن این آزمون ۲۰% خواهد بود
پروژه (نمره اضافی)	۱۰%	موضوع اختیاری، تعیین موضوع تا شنبه ۱۴۰۱/۰۲/۰۵ تحويل پروژه: اولین دوشنبه بعد از آخرین امتحان پایان‌ترم
حضور و مشارکت کلاس (نمره اضافی)	۵%	پاسخ دادن به سوالات حین تدریس و مشارکت در بحث‌های کلاس

معرفی درس . . .



دستیار آموزشی □

◀ امیدرضا بهرامیان

◀ or.bahramian@gmail.com

◀ برگزاری کلاس‌های حل تمرین و رفع اشکال

◀ به ویژه راهنمایی پیاده‌سازی و کدنویسی

معرفی درس: سرفصل‌ها . . .

- معرفی زبان‌شناسی رایانشی
- معرفی نمونه‌هایی از کاربردهای پردازش زبان طبیعی
- واحدسازی (Tokenization)
- نرمال‌سازی (Normalization)
- ◀ مشکلات زبان فارسی و راه‌حل‌ها

- تحلیل مورفولوژی (Morphological Analysis)
 - ◀ انواع مورفولوژی
 - ◀ عبارت باقاعده (Regular Expression)
 - ◀ اتوماتای محدود (FSA: Finite State Automata)
 - ◀ ریشه‌یابی بدون واژگان (Porter Stemmer)
 - ◀ خطایاب املائی
- ◀ روش Minimum Edit Distance

معرفی درس: سرفصل‌ها . . .

□ مدل‌سازی زبانی (Language Modeling)

◀ شمارش کلمات و قانون Zipf

◀ مدل n-gram

◀ هموارسازی (Smoothing)

◀ ارزیابی مدل‌های زبانی

□ برچسپ‌زنی اجزای کلام (POS: Part-of-Speech tagging)

◀ کاربردها

◀ روش‌ها

◀ مبتنی بر قاعده

◀ آماری و مبتنی بر مدل مخفی مارکوف (HMM)

معرفی درس: سرفصل‌ها . . .

□ گرامر و تجزیه نحوی (Parsing)

- ◀ انواع گرامر
- ◀ الگوریتم‌های تجزیه
 - ◀ تجزیه CKY
 - ◀ تجزیه Earley

□ آواشناسی و مبانی پردازش سیگنال

□ مروری بر کاربردها

- ◀ بازشناسی گفتار
 - ◀ پیچیدگی‌ها
 - ◀ روش‌ها
- ◀ سنتز گفتار
- ◀ ترجمه ماشینی

معرفی درس

زمان بندی

متناسب شرایط و سطح کلاس، و همچنین تغییرات پیش بینی نشده در زمان بندی، ممکن است سرفصل مطالب و یا زمان بندی های کلاس مقداری تغییر داشته باشد

توضیحات	موضوع	تاریخ	هفته
	معرفی درس، معرفی زبان شناسی رایانشی و کاربردهای آن	۱۴۰۰/۱۱/۲۵ و ۲۳	۱
	واحدسازی و نرمال سازی، ساخت واژه (مورفولوژی)	۱۴۰۰/۱۱/۳۰ ۱۴۰۰/۱۲/۰۲	۲
	ساخت واژه (مورفولوژی): عبارتهای منظم	۱۴۰۰/۱۲/۰۹ و ۰۷	۳
تمرین	ساخت واژه (مورفولوژی): ماشین/مبدل حالت محدود	۱۴۰۰/۱۲/۱۶ و ۱۴	۴
	خطایاب املایی و Minimum Edit Distance	۱۴۰۰/۱۲/۲۳ و ۲۱	۵
	مدل سازی زبانی	۱۴۰۰/۱۲/۲۸	۶
تمرین	مدل سازی زبانی با N-Gram	۱۴۰۱/۰۱/۱۵	۷
	هموارسازی و ارزیابی مدل سازی زبانی	۱۴۰۱/۰۱/۲۲ و ۲۰	۸
	برچسپ زنی اجزای کلام	۱۴۰۱/۰۱/۲۹ و ۲۷	۹
تمرین	آزمون میان ترم	۱۴۰۱/۰۲/۰۵ و ۰۳	۱۰
	برچسپ زنی اجزای کلام و مدل مخفی مارکوف (HMM)	۱۴۰۱/۰۲/۱۲ و ۱۰	۱۱
	برچسپ زنی اجزای کلام و مدل مخفی مارکوف (HMM)	۱۴۰۱/۰۲/۱۹ و ۱۷	۱۲
تمرین	انواع گرامر و تجزیه نحوی	۱۴۰۱/۰۲/۲۶ و ۲۴	۱۳
	گرامر و تجزیه نحوی CKY و Early	۱۴۰۱/۰۲/۳۱	۱۴
		۱۴۰۱/۰۳/۰۲	
	آواشناسی و مبانی پردازش سیگنال	۱۴۰۱/۰۳/۰۹ و ۰۷	۱۵
تمرین	آواشناسی و مبانی پردازش سیگنال	۱۴۰۱/۰۳/۱۶ و ۱۴	۱۶
	مروری بر کاربردها (بازشناسی)	۱۴۰۱/۰۳/۲۳ و ۲۱	۱۷
		۱۴۰۱/۰۲/۰۳: شهادت حضرت علی	
		۱۴۰۱/۰۲/۱۲: عید فطر	
		۱۴۰۱/۰۳/۱۴: رحلت امام خمینی (ره)	