روش‌های آماری در پردازش زبان طبیعی
(۸۳-۰۵-۱۶۲)

Statistical Methods in Natural Language Processing

هادی ویسی

h.veisi@ut.ac.ir

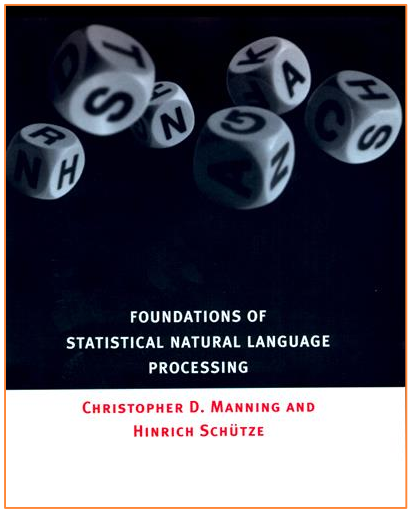
دانشگاه تهران – دانشکده علوم و فنون نوین

معرفی درس ...

- زمان و مکان
  - شنبه و دوشنبه، ساعت ۱۱:۰۰ الی ۱۲:۳۰
  - دانشکده علوم و فنون نوین- ساختمان ۲۳
- وب سایت
  - http://dsp.ut.ac.ir/courses/statistical-nlp-fall1394

- هدف
  - مرور روش‌های یادگیری ماشین در پردازش زبان طبیعی (با تاکید بر روش‌های آماری)
    - مفاهیم پایه آمار و احتمال، نظریه اطلاعات و روش‌های تخمین
    - مفاهیم یادگیری ماشین
    - روش‌های مدل‌سازی آماری
    - مرور نمونه کاربردها
  - فعالیت‌های تمرینی با رویکرد کاربردی

معرفی درس ...



- منابع

- Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.
- Daniel Jurafsky, James Martin, Speech and Language Processing, 2nd Edition, Prentice Hall, 2009.
- Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.
- 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 پاسخ سوالات یک تمرین ، 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
      - در صورت داشتن یک روز تاخیر در ارسال پاسخ‌ها (از یک ثانیه تا ۲۴ ساعت!)، ۲۵٪ نمره آن تمرین به عنوان جریمه تاخیر
      - در صورت تاخیر دو روزه ۵۰٪ نمره مربوطه به عنوان جریمه تاخیر
      - پس از دو روز نمره‌ای در نظر گرفته نمی‌شود
    - وزن تمرین‌های مختلف با هم برابر نیست

معرفی درس ...

- ارزیابی . . .
  - آزمونک (کویز)
    - از مطالب هر بخش
    - ممکن است بدون اطلاع قبلی باشد
  - امتحان میان‌ترم
    - نداریم!
  - امتحان پایان‌ترم
    - شامل کلیه مطالب تدریس شده
    - زمان: ؟؟

معرفی درس ...

- ارزیابی . . .
  - پروژه: یک مورد
    - پروژه کاربردی دارای پیاده‌سازی در MATLAB، Python یا سایر زبان‌های برنامه‌نویسی
    - علاوه‌بر کد برنامه، گزارش مکتوب (به صورت تایپ شده) تحویل گرفته می‌شود
    - تحویل حضوری
    - آخرین زمان تعیین موضوع پروژه: روز دوشنبه ۲۷/۰۷/۱۳۹٤
    - تحویل پروژه: ۵ روز بعد از آخرین امتحان پایان‌ترم
  - مقاله (نمره اضافی)
    - مقاله آماده انتشار مورد قبول است
    - به هر قیمتی مقاله ننویسید!
  - بازنگری نمره‌ها و برگه‌ها
    - در زمان تحویل پروژه درس (به صورت حضوری)

معرفی درس ...

- ارزیابی

عنوان	وزن	توضیح
تمرین	٪۴۰	بعد از هر موضوع (وزن تمرین‌ها برابر نیست)
امتحان کوتاه (کویز)	٪۱۵	ممکن است بدون اطلاع قبلی باشد
امتحان پایان‌ترم	٪۳۰	از کل مطالب درس روز ؟؟/۱۰/۱۳۹۴ ساعت ؟؟:۰۰
پروژه	٪۱۵	موضوع اختیاری تعیین موضوع تا دوشنبه ۲۷/۰۷/۱۳۹۴ تحویل پروژه: ۵ روز بعد از آخرین امتحان پایان‌ترم
مقاله (نمره اضافی)	٪۱۵	مقاله آماده انتشار مورد قبول است

معرفی درس ...

- سرفصل‌ها . . .
  - مروری بر مبانی آمار و احتمال
    - احتمال (توام، شرطی)، امید ریاضی
    - قانون بیز
    - متغیر تصادفی
    - توابع توزیع
  - مروری بر نظریه اطلاعات و آنتروپی
  - مروری بر روش‌های تخمین
    - کمینه میانگین مربعات خطا (MMSE)
    - تخمین بیشینه شباهت (MLE)
    - تخمین بیز (Bayesian)
  - مروری بر مفاهیم و اصول یادگیری ماشین
  - مدل مخفی مارکوف (HMM)
    - کاربرد در برچسب‌زنی اجزای کلام (POS: Part-of-Speech tagging)
  - تبدیل متن به بردار ویژگی

معرفی درس ...

- سرفصل‌ها
  - روش‌های خوشه‌بندی
    - روش k-میانگین و الگوریتم امید-بیشینه (EM)
  - روش بیز (ساده)
  - شبکه عصبی مصنوعی
    - مبانی و مفاهیم
    - شبکه عصبی پرسپترون چندلایه (MLP)
    - شبکه عصبی احتمالاتی (PNN)
    - شبکه عصبی خودسازمان‌ده (SOM)
  - گرامر مستقل از متن احتمالاتی (PCFG)
  - نمونه کاربردها
    - دسته‌بندی متون
    - مروری بر بازیابی اطلاعات در متن
    - مروری بر بازشناسی گفتار
    - مروری بر ترجمه ماشینی

معرفی درس

- زمان‌بندی

هفته	تاریخ	موضوع	توضیحات
۱	۲۱و۲۳/۰۶/۱۳۹۴	معرفی درس و مروری بر مبانی آمار و احتمال	
۲	۲۸و۳۰/۰۶/۱۳۹۴	مروری بر نظریه اطلاعات و روش‌های تخمین	
۳	۰۴و۰۶/۰۷/۱۳۹۴	مروری بر مفاهیم یادگیری ماشین	تمرین، کویز
۴	۱۱و۱۳/۰۷/۱۳۹۴	مدل مخفی مارکوف (HMM)	
۵	۱۸و۲۰/۰۷/۱۳۹۴	کاربرد مدل مخفی مارکوف در POS Tagging	تمرین، کویز
۶	۲۵و۲۷/۰۷/۱۳۹۴	تبدیل متن به بردار ویژگی	اعلام موضوع پروژه
۷	۰۲و۰۴/۰۸/۱۳۹۴	خوشه‌بندی	
۸	۰۹و۱۱/۰۸/۱۳۹۴	روش E-M و روش بیز ساده	
۹	۱۶و۱۸/۰۸/۱۳۹۴	شبکه عصبی مصنوعی	تمرین، کویز
۱۰	۲۳و۲۵/۰۸/۱۳۹۴	شبکه عصبی پرسپترون چندلایه	
۱۱	۳۰/۰۸/۱۳۹۴ ۰۲/۰۹/۱۳۹۴	شبکه عصبی احتمالاتی شبکه عصبی خودسازمان‌ده	
۱۲	۰۷و۰۹/۰۹/۱۳۹۴	گرامر مستقل از متن احتمالاتی (PCFG)	تمرین، کویز
۱۳	۱۴و۱۶/۰۹/۱۳۹۴	دسته‌بندی متون	
۱۴	۲۱و۲۳/۰۹/۱۳۹۴	بازیابی اطلاعات در متن	
۱۵	۲۸و۳۰/۰۹/۱۳۹۴	مروری بر بازشناسی گفتار	تمرین، کویز
۱۶	۰۵و۰۷/۱۰/۱۳۹۴	مروری بر ترجمه ماشینی	