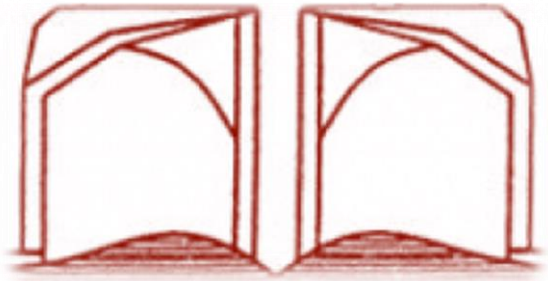

آشنایی با زبان‌شناسی رایانشی

واحدسازی و نرمال‌سازی

هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران – دانشکده علوم و فنون نوین

فهرست

واحدسازی و نرمال‌سازی

- برای پردازش متن، اولین کار واحدسازی و نرمال‌سازی متن ورودی است
  - کاربردهایی مانند ترجمه ماشینی، بازیابی اطلاعات، تبدیل متن به گفتار و ...

- واحدسازی (Tokenization)
  - تقطیع دنباله کاراکترهای تشکیل دهندهٔ متن به دنباله ای از کلمات

- نرمال‌سازی (Normalization)
  - یکدست سازی واحدهای متنی به طوری که قابل پردازش توسط ماشین باشد.

واحدسازی

- **ساده‌ترین راه برای واحدسازی متن و تقطیع آن به دنباله کلمات**
  - **واحدسازی بر اساس کاراکتر فاصله** (Space)
  - Token: **واحدهای متنی که با استفاده از فاصله از هم جدا می‌گردند.**

- **مشکل: همیشه کاراکتر فاصله مرز کلمات را مشخص نمی‌کند**
  - **کلماتی که از چند** token **تشکیل شده اند** (Multi Token Units)
    - می روند – کتاب ها – در حالی که – زبان شناسی
    - air force - St. Petersburg - go back - take off
  - **کلماتی که به هم چسبیده و یک** Token **را به وجود آورده‌اند** (Multi Unit Tokens)
    - ناشی از حروفی که شکل متصل و غیرمتصل آنها با هم یکسان است: ر، ز، ژ، و، ذ، د، آ، ا
    - درکتاب ⇦ در کتاب
    - مرا/زوی/کزین ⇦ من را/ از وی/که از این (برای POS tagging)
    - وبا ⇦ و با (یا «وبا»؟)

واحدسازی: مشکلات

- علائم نشانه‌گذاری (Punctuations) معمولاً به کلمه قبل از خود می‌چسبند.
  - علائم نقطه گذاری را هم مانند کاراکتر فاصله به عنوان مرز کلمات به حساب آوریم.
  - استثنا: اختصاراتی مانند B.B.C و اعداد ممیزدار مانند 2.3
- پسوندها و پیشوندها
  - در فارسی پسوندها و پیشوندها در بسیاری مواقع با فاصله از کلمه اصلی نوشته می‌شوند
  - مثال: رفته ام – بر می گردد – دانش آموز
- کلمات مرکب
  - بین اجزای کلمات مرکب معمولاً فاصله گذاشته می‌شود
  - دوچرخه سوار – برون مرزی

واحدسازی: راه حل‌ها

- استفاده از علائم نشانه‌گذاری
  - تعیین قطعی مرز کلمات با علائم «،»، «؛»، «:»، «!» و «؟»
  - نقطه «.»: علاوه بر مرز در سرنام‌ها و علائم اختصار هم می‌آید
- استفاده از قواعد املایی برای پسوندها و پیشوندها
  - مثال: چسباندن تمام پیشوندهای "می" به اول کلمات و پسوندهای "ها" به آخر کلمات
  - رفع موارد استثنا: مشخص کردن تمام کلمات مُجازی که وندها به آنها می‌چسبند (با استفاده از یک پیکره متنی بزرگ)
- استفاده از یک پیکره متنی بزرگ برای مشخص نمودن MUTهای پرکاربرد و تشکیل یک جدول از آنها برای تقطیع
- استفاده از یک پیکره متنی تقطیع شده برای مشخص کردن کلمات چندقسمتی و تشکیل قواعد املایی برای آنها
  - استفاده از جدول حاوی لیست کلمات مرکب

نرمال‌سازی: مشکلات زبان فارسی . . .

- وجود encodingهای مختلف برای بعضی از کاراکترها
  - مانند "ی" و "ک"
- روش‌های مختلف چسبیدن وندها به کلمات اصلی
  - می روند - می‌روند – میروند
  - زبان شناسی - زبان‌شناسی – زبانشناسی
- روش‌های مختلف اتصال اجزای کلمات مرکب
  - برون مرزی - برون‌مرزی - برونمرزی
- کلمات چندامﻻیی
  - مسئولیت – مسوولیت – مسؤولیت
  - گفتگو – گفت و گو – گفت وگو - گفت‌و گو - گفت‌وگو
  - پاییز – پائیز

نرمال‌سازی: مشکلات زبان فارسی . . .

- فرم‌های نوشتاری مختلف کلمهٔ «درحالی‌که» در پیکرهٔ متنی فارسی

صورت نوشتاری	تعداد
در_حالي_كه	22,845
درحالي^كه	2,210
درحالي_كه	2,057
در_حاليكه	1,818
در_حالي^كه	1,220
درحاليكه	359
در^حالي^كه	24
در^حالي_كه	4
در^حاليكه	2

- ابهام در تبدیل فاصله به نیم‌فاصله
  - اسب سواری یکی از تفریحات انسان‌هاست
  - او اسب سواری را با تازیانه زد
  - من با اسب سواری کردم

نرمال‌سازی: راه حل‌ها . . .

- **درست کردن فاصله بین متن و علائم نشانه‌گذاری**
  - **علائم «.»، «،»، «؛»، «:»، «!» و «؟» به کلمات قبلی می‌چسپند و با کلمه بعدی به اندازه یک واحد فاصله (Space)، فاصله دارند**
    - مثال: « من،تو ، او ضمایر شخصی فارسی هستند . » ⇦ « من، تو، او ضمایر شخصی فارسی هستند. »
    - <عبارت+ علامت+ فاصله+ عبارت>
    - استثنا: علامت «.» در سرنام‌ها و اختصارها
      - مثال: « در سال پنجم هـ . ق » ⇦ « در سال پنجم هـق »
  - **علائم {}، []، () و «» به عبارت داخل خود می‌چسپند**
    - مثال: «دانشجویان اینجا(دانشگاه تهران) در ...» ⇦ «دانشجویان اینجا (دانشگاه تهران) در ...»
    - <عبارت+ فاصله+ علامت ابتدا+ عبارت داخل + علامت انتهایی+ فاصله+ عبارت>

نرمال‌سازی: راه حل‌ها . . .

- درست کردن فاصله بین متن و اعداد
  - اعداد باید با کلمات قبل و بعد از خود یک واحد فاصله داشته باشند
    - مثال: « دیروز ۶مرد و ٤زن مدل گرفتند. » ⇦ «دیروز ۶ مرد و ٤ زن مدل گرفتند »
  - در اعداد اعشاری و تاریخ علامت «ممیز» باید یک‌دست و اصلاح شود
    - مثال: « معدلش در ۱۲/۱۱/۱۳۹۳برابر با ۱۷.۵ است. » ⇦ «معدلش در ۱۲/۱۱/۱۳۹۳برابر با ۱۷/۵ است. »

نرمال‌سازی: راه حل‌ها . . .

- درست کردن فاصله بین کلمات و پسوندها
  - پسوندها: اشتقاقی یا صرفی
  - <کلمه+ پسوند اشتقاقی+ پسوند صفت+نشانه جمع+پسوند ملکی(یا ی نکره)>
    - پسوند صفت: مانند «تر» و «ترین»
    - نشانه جمع: مانند «ها»
    - پسوند ملکی: مانند «م، ت، ش، مان، تان، شان»
    - یک کلمه نمی تواند هم شامل «ی نکره» و هم پسوند ملکی باشد
  - کلماتی که به «ا» یا «و» ختم می‌شوند، قبل از پسوند ملکی یک «ی» میانجی قرار می‌گیرد
    - مثال: عصایم، سبویم
  - برخی پسوندهای اشتقاقی می‌توانند نقش اسم هم داشته باشند
    - مثال: «سیر» در «گرم سیر» یا «سیر≠گرسنه»
  - کلماتی که چند پسوند دارند، همه آنها با نیم‌فاصله از کلمه (یا سایر پسوندها) جدا می‌شوند
    - مثال: « جزوه ها یمان » ⇦ « جزوه‌هایمان »
    - مثال: « حیله گر ها » ⇦ « حیله‌گرها »
    - مثال: « زشت ترین ها » ⇦ « زشت‌ترین‌ها »

نرمال‌سازی: راه حل‌ها . . .

- درست کردن فاصله بین کلمات و پیشوندهای اشتقاقی
  - <پیشوند اشتقاقی+ کلمه>: با نیم‌فاصله از کلمه بعد قرار می‌گیرند
    - مثال: « بی ادب » ⇦ « بی‌ادب »
    - مثال: « نا مرد » ⇦ « نامرد »

نرمال‌سازی: راه حل‌ها . . .

- **درست کردن فاصله در افعال**
  - **پسوندهای فعل ماضی نقلی: با نیم‌فاصله از کلمه قبل (فعل) قرار می‌گیرند**
    - همه پسوندهای «ام»، «ای»، «است»، «ایم»، «اید» و «اند»
      - گاهی «است» را استثنا می‌کنند و آن را با فاصله جدا می‌کنند
    - مثال: « گرفته ام » ⇦ « گرفته‌ام »
  - **پسوند «می» در ماضی استمراری و مضارع اخباری: با نیم‌فاصله از جز فعلی قرار می‌گیرد**
    - استثنا: «می به معنی شراب» ⇦ اگر کلمه بعد از «می» اسم بود، فاصله را حذف نکنید
      - قطعی نیست: «می» در «می خورد» می‌تواند هم به معنای شراب باشد و هم جز فعلی
  - **اصلاح پسوندهایی مانند «بر»**
    - مثال: « بر خواهد گشت » ⇦ « برخواهدگشت »

نرمال‌سازی: راه حل‌ها . . .

- درست کردن نشانه جمع «ها»
  - هم به صورت متصل و هم به صورت مجزا ظاهر می‌شود
  - مواردی که «ها» باید جدا نوشته شود
    - قرار گرفتن بعد از کلمات بیگانه نامانوس مانند «فرمالیست‌ها»
    - مشخص کردن کلمه به منظور برجسته کردن آن و یا آموزش: « دبستانی‌ها»
  - در نرمال‌سازی بهتر است همه جا «ها» را با نیم‌فاصله جدا کنید

نرمال‌سازی: راه حل‌ها . . .

- یک‌دست‌سازی املای کلمات چندامللایی با استفاده از جدول حاوی کلمات (Lookup Table) بر اساس شیوه نگارش استاندارد
  - مسوولیت ← مسئولیت
- حذف علائم نشانه‌گذاری (diacritics)
  - فنّاوری ← فناوری
  - کاملاً ← کاملا
- درست کردن کاراکترهایی که چند کد دارند
  - ي ← ی
  - ک ← ک
  - ؤ ← و
  - ئ ← ی
  - أ ← ا
  - ۀ ← ه

نرمال‌سازی: راه حل‌ها

- سایر
  - حذف «ی» از اخر کلماتی مانند «کلمه ی»
  - حذف «ء» از کلماتی مانند «املاء»
  - چسپاندن پیشوند «هم» در کلماتی مانند «هم چنین»
  - چسپاندن پیشوند «این» و «آن» در کلماتی مانند «این جا» و «آن جا»

- نیم فاصله
  - کاراکتر zero-width non-joiner (ZWNJ)
  - یونی کد: U+200C
  - در HTML: &#8204 یا &zwnj

نرمال‌سازی: موارد تکمیلی

- در بعضی کاربردها مانند تبدیل متن به گفتار، موارد دیگری نیز باید نرمال‌سازی شوند
  - بعضی از اختصارات
    - (ع) ← علیه السلام
  - اعداد
    - ۲۳ ← بیست و سه
    - ۲.۳ ← دو ممیز سه دهم
    - ۱/۱۰۰۰ ← یک هزارم
    - ۶۶۶۱۱۵۲۵
      - اگر شماره تلفن باشد: شصت و شش شصت و یک پانزده بیست و پنج
      - اگر مبلغ باشد: شصت و شش میلیون و شش صد و یازده هزار و پانصد و بیست و پنج
  - تاریخ
    - ۸۹/۸/۱۵ ← پانزدهِ هشتِ هشتاد و نه