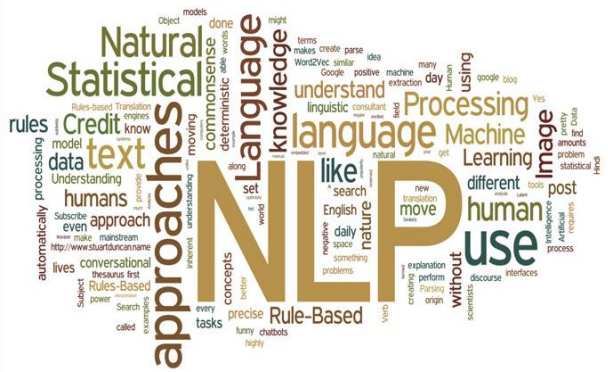


به نام خدا

گروه هوش مصنوعی، دانشکده مهندسی کامپیوتر

بخش پنجم

استخراج اطلاعات و تشخیص موجودیت نامدار

IE & NER

حمیدرضا برادران کاشانی

سرفصل مطالب

- مقدمه ای بر استخراج اطلاعات
- مراحل اصلی استخراج اطلاعات
- شناسایی موجودیت های نامدار (NER)
  - تعریف و کاربرد NER
  - مراحل اصلی NER
  - ویژگی ها برای مساله NER
  - کدگذاری موجودیت های نامدار

Hamidreza Baradaran Kashani

مقدمه ای بر استخراج اطلاعات

- سیستمهای استخراج اطلاعات (IE) چه کاری انجام می دهند؟
  - یافتن و درک روابط موجود در متون
  - جمع آوری اطلاعات از قسمت های مختلف متن
  - تولید یک فرمت ساختار یافته از اطلاعات مرتبط موجود در متن: مثلا ایجاد یک پایگاه دانش
- هدف سیستمهای IE چیست؟
  - کمک مستقیم به انسان با سازماندهی کردن اطلاعات
  - مفید برای استنتاج بیشتر اطلاعات توسط کامپیوترها:
    - ✓ پردازش اطلاعات در قالبی ساختاریافته مثلا پایگاه دانش به مراتب برای یک سیستم کامپیوتری ساده تر از درک یک متن زبان طبیعی است.

Hamidreza Baradaran Kashani

استخراج اطلاعات

- سیستمهای استخراج اطلاعات (IE) اطلاعات شفاف و واقعی را استخراج می کنند:
- چه کسی چه کاری را چه زمانی برای چه کسی انجام داد؟ (*Who did what to whom when*)

- *مثال:*
- استخراج اطلاعات مختلف یک شرکت از گزارش هایشان، مثلا میزان سود و درآمد، اعضای شرکت و محل دفاتر مرکزی شرکت و ...

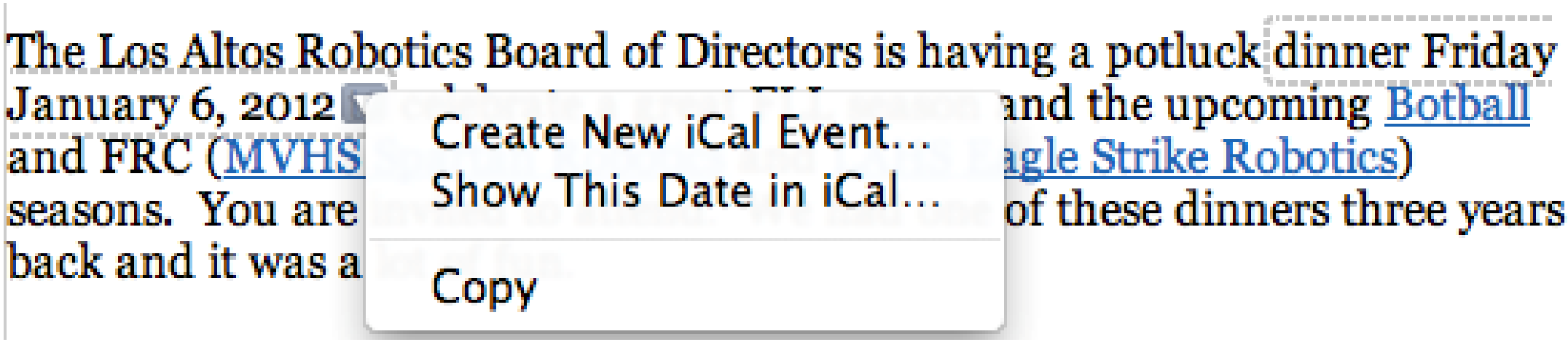
The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.

headquarters("BHP Biliton Limited", "Melbourne, Australia")

Hamidreza Baradaran Kashani

استخراج اطلاعات در سطوح پایین

- استخراج اطلاعات در سطوح پایین در کاربردهای مختلفی استفاده می شود مثل:
  - Google mail و شاخص گذاری وب (web indexing)



مراحل اصلی استخراج اطلاعات

❖ فرآیند استخراج اطلاعات:

1) تحلیل داده های ساختارنیافته مانند روزنامه ها، مقالات علمی و صفحات وب و ...

2) استخراج موجودیت ها (Entities)، رخدادها (Events) و روابط میان آنها (Relations)

3) ساخت یک پایگاه دانش (Knowledge base) با استفاده از اطلاعات و موجودیت های استخراج شده

✓ مثلا کدام شرکت چه محصولاتی را دارد؟ یا کدام مقالات علمی در ارتباط با موضوع خاصی هستند؟

Hamidreza Baradaran Kashani

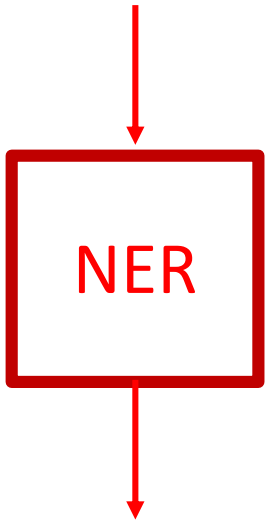
تشخیص موجودیت های نامدار (NER)

❖ تفاوت بین Entity و Named Entity ؟

❖انواع موجودیت های نامدار:

- اشخاص (Person)
- مکان ها (Locations)
- سازمانها (Organizations)
  - ✓ وزارتخانه ها، شرکتها
  - ✓ مجلات و روزنامه ها
  - ✓ تیم های ورزشی
- ✓ زمانها (Times)
- ✓ کشورها (Geo-political entities or GPE)

Jim bought 300 shares of Acme Corp. in 2006



[Jim]Person bought 300 shares of [Acme Corp.] Organization in [2006]Time

Hamidreza Baradaran Kashani

کاربردهای NER

❖ **سیستم های استخراج اطلاعات (IE)**

✓ عضو کلیدی در سیستمهای IE ، NER است.

❖ **سیستم های پرسش و پاسخ (QA)**

✓ پاسخ ها اغلب موجودیت های با نام هستند.

❖ **ذخیره سازی و بازیابی اطلاعات**

✓ موجودیت ها را میتوان اندیس گذاری کرد.

✓ بسیار مفید در بهینه سازی فرآیند جستجو است.

❖ **استخراج محتوای متن و خلاصه سازی متون**

Hamidreza Baradaran Kashani

مراحل اصلی NER

❖ **دو کار اصلی انجام شده در مساله NER:**

۱- آشکارسازی (Detection) موجودیت های نامدار (NE) در متن

۲- دسته بندی (Classification) NE های آشکار شده به یکی از انواع موجودیت ها

Brazilian football legend Pele’s condition has improved, according to a Thurseday evening statement from a Sao Paulo hospital.

Hamidreza Baradaran Kashani

مراحل اصلی NER

❖ **دو کار اصلی انجام شده در مساله NER:**

۱- آشکارسازی (Detection) موجودیت های نامدار (NE) در متن

۲- دسته بندی (Classification) NE های آشکار شده به یکی از انواع موجودیت ها

Brazilian football legend Pele's condition has improved, according to a Thurseday evening statement from a Sao Paulo hospital.

Hamidreza Baradaran Kashani

مراحل اصلی NER

❖ **دو کار اصلی انجام شده در مساله NER:**

۱- آشکارسازی (Detection) موجودیت های نامدار (NE) در متن

۲- دسته بندی (Classification) NE های آشکار شده به یکی از انواع موجودیت ها

Brazilian football legend [PERSON Pele]'s condition has improved, according to a [TIME Thurseday evening] statement from a [LOCATION Sao Paulo] hospital.

Hamidreza Baradaran Kashani

رویکرد یادگیری ماشین برای مساله NER

- **مرحله یادگیری (Training)**
  - جمع آوری مجموعه ای از اسناد متنی
  - تخصیص برچسب صحیح موجودیت نامدار به هر توکن
  - استخراج ویژگی (Feature) از کلمات موجود در متن
  - آموزش یا یادگیری سیستم طبقه بند (Classifier) با ویژگی های استخراجی

- **مرحله آزمون (Test)**
  - تهیه مجموعه ای از اسناد متنی بدون برچسب به عنوان مجموعه تست
  - اجرای مدل آموزش دیده بر روی توکن های تست
  - برگرداندن خروجی بصورت موجودیت های آشکار شده و برچسب گذاری شده

Hamidreza Baradaran Kashani

ویژگی ها برای مساله NER

❖ ویژگی های سطح کلمه

- استخراج از خود کلمه و کلمات مجاور
- برچسب نحوی مثل POS
- ویژگی های املایی
  - مثل بزرگ بودن تمام حروف کلمه، بزرگ بودن تنها حرف اول کلمه، شامل بودن عدد، وجود علائم نگارشی
  - پیشوند و پسوند و ریشه کلمات، طول کلمات

Hamidreza Baradaran Kashani

ویژگی ها برای مساله NER

❖ ویژگی های مبتنی بر دیکشنری یا لیست

- استفاده از لیست هایی از اسامی خاص مشهور، سازمانها و مکانها و مخفف های آنها.
  - برای مثال حضور یا عدم حضور موجودیت ها در لیست های هر یک از انواع موجودیت ها به عنوان ویژگی در نظر گرفته می شود.

❖ ویژگی های سطح سند و پیکره

- این ویژگی ها بر اساس محتوا و ساختار سند تعریف می شوند.
  - تعداد رخدادهای کلمه در سند و تعداد رخدادهای کلمه با حروف کوچک و بزرگ در سند و محل قرارگیری کلمه در سند و ....

Hamidreza Baradaran Kashani

کدگذاری موجودیت های نامدار

❖ دو نوع کدگذاری برای NER

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

IOB : Inside, Outside, Begin

دو نکته مهم:

۱- یک موجودیت ممکن است بیش از یک توکن باشد.

۲- روش کدگذاری IOB روش رایج برای برچسب گذاری NE ها است.

Hamidreza Baradaran Kashani

مثال کدگذاری IOB بر روی متن فارسی

❖ داده Arman NER:

NEs are categorized into six classes:

1- person,
2- organization (such as banks, ministries, embassies, teams, nationalities, networks and publishers),
3- location (such as cities, villages, rivers, seas, gulfs, deserts and mountains),
4- facility (such as schools, universities, research centers, airports, railways, bridges, roads, harbors, stations, hospitals, parks, zoos and cinemas),
5- product (such as books, newspapers, TV shows, movies, airplanes, ships, cars, theories, laws, agreements and religions),
And
6- event (such as wars, earthquakes, national holidays, festivals and conferences); other are the remaining tokens.

Hamidreza Baradaran Kashani

مثال کدگذاری IOB بر روی متن فارسی

❖ داده Arman NER:

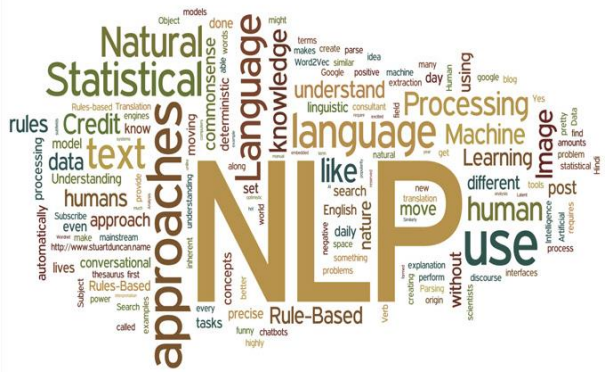
```
سعید B-pers
پورصمیمی I-pers
که O
فیلم O
عروس B-pro
آتش I-pro
با O
بازی O
او O
هم اکنون O
روی O
پرده O
سینما های O
کشور O
است O
```

```
کاشان B-loc
: O
کاروانسرای B-fac
میرپنج I-fac
کاشان I-fac
در O
شمار O
آثار O
ملی O
کشور O
به O
ثبت O
رسید O
```

```
استاندار O
اردبیل B-loc
گفت O
: O
به O
مناسبت O
هفته B-event
دولت I-event
0 O
طرح O
عمرانی O
و O
تولیدی O
در O
```

```
دکتر O
اکبر B-pers
میرعرب I-pers
در O
همایش B-event
بررسی I-event
و I-event
پیشگیری I-event
از I-event
بیماری I-event
ایدز I-event
در O
همدان B-loc
به O
خبرنگاران O
گفت O
```

Hamidreza Baradaran Kashani



با تشکر از توجه شما