به نام خدا

دانشکده مهندسی کامپیوتر

درس پردازش زبان‌های طبیعی
نیم‌سال دوم ۰۱-۰۲
استاد: دکتر عسگری

تمرین دوم **محمدمهدی قره‌گوزلو، ۴۰۱۲۰۶۸۳۶**

توضیحات کلی

انتخاب داده

با توجه به اطلاعات نگارش شده در داکیومنت تمرین دوم بایستی با استفاده از عبارات منظم اطلاعات مورد نیاز را از متون ارائه شده استخراج کنیم. مشخصا در ترک *بررسی متن ویزیت* ۱۰ موجودیت مختلف را باید پیدا کنیم.
برای این کار من داده‌های ترک ۲ را انتخاب کردم که تا حد خوبی دارای ساختار مناسب برای عبارات منظم است و کار را راحت‌تر می‌کند.

پیش پردازش

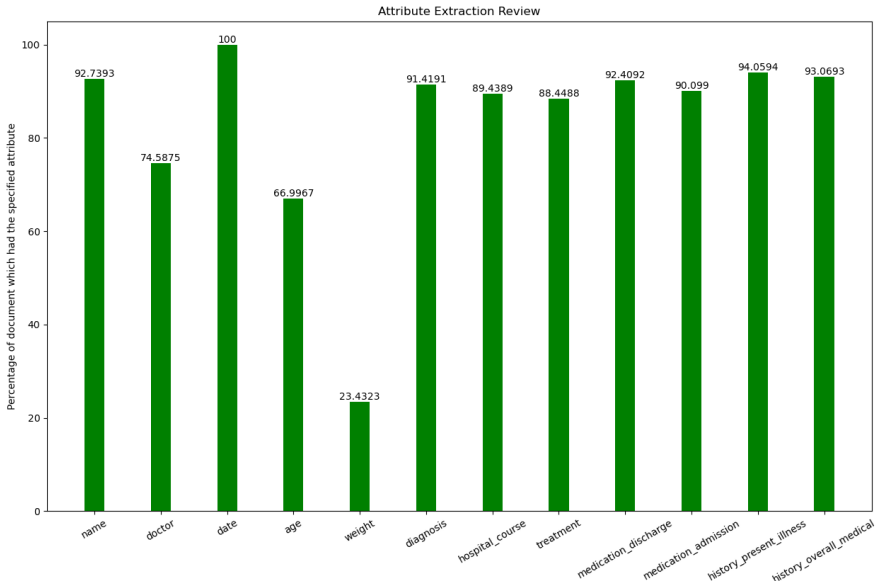
با توجه به اینکه عبارات منظم بر اساس الگوهای مشخص و جداکننده تعریف می‌شوند، بعضی تغییرات یکسان سازی در متن ممکن است استخراج اطلاعات را مشکل کند. مثلا بیشتر داروها با حرف بزرگ انگلیسی شروع می‌شوند و کوچک کردن حروف پیدا کردن آنها را سخت می‌کند. یا در یک نمونه دیگر برای پیدا کردن نام دکتر معمولا به دنبال عبارت *.Dr* می‌گردیم و اگر علایم نگارشی را حذف کرده باشیم تشخیص اینگونه الگوها نیز مشکل می‌شود. بنابراین من کل متن را پیش پردازش نمی‌کنم اما در مراحل مختلف بر اساس نیاز برخی پردازش‌ها مانند حذف علایم یا کوچک کردن حروف را انجام خواهم داد.

نگاهی کلی به داده‌ها

هر سند از قسمت‌های مختلفی مانند داروهای تجویزشده تا بیماری‌های مراجعه کننده تشکیل شده است. برخی از اسناد فاقد اطلاعات خواسته شده می‌باشند یا ممکن است الگوهای تعریف شده توسط نتوانسته باشد آنها را پیدا کند. در نمودار پیش‌رو می‌توانید یک دید کلی نسبت به وضعیت اطلاعات بدست آمده داشته باشید.
توجه کنید که روش‌های قاعده محور مانند عبارات منظم ممکن است دارای دقت بالایی نباشند، مثلا در لیست داروهای استخراج شده شاید بعضی از آنها جا بمانند یا برخی نوشته‌های بی ارتباط دیگر به عنوان دارو شناسایی شوند. نمودار فوق صرفا یک دید سطح بالا ارایه می‌دهد و لزوما بیانگر دقت نیست.

نوع خروجی

اطلاعات استخراج شده هر بیمار در یک فایل *json* به همان نام سند بیمار ذخیره می‌شود، این فایل‌ها در دایرکتوری *track2_results* قرار دارند. در ادامه مثال‌هاتیی را مشاهده خواهید کرد، این اطلاعات از سند ۱۰۰۰۳۹ بدست آمده



شکل ۱: درصد تعداد اسناد با اطلاعات یافت شده نسبت به تعداد کل اسناد

استخراج اطلاعات

نام بیمار

تقریبا در همه اسناد نام بیمار با استفاده از عبارت *Attending* مشخص شده است. بنابراین برای پیدا کردن نام بیمار دنبال این عبارت می‌گردیم.

```
(attending:|mr.|mrs.|ms.)\s*\[\*{2}(.+?)\*{2}\]
```

البته نام می‌تواند با استفاده پیشوندهایی مانند *Mr, Ms, Mrs* نیز مشخص شود پس اینها را نیز در الگوی مورد نظر شامل می‌کنیم. البته که بیشتر نام‌ها از طریق *Attending* یافت می‌شوند. همچنین تمامی نام افراد در این اسناد حذف شده (احتمالا به دلیل رعایت حریم شخصی) و به جای هر نام یک عبارت محصور شده در براکت و ستاره قرار داده شده است که این را نیز در الگو بایستی در نظر گرفت.

```
"Patient Name": "First Name3 (LF) 3918"
```

نام دکتر

برای بدست آوردن نام دکتر می‌توان از پیشوند *Dr.* استفاده کرد. همچنین عباراتی مانند *Doctor, Surgen, Phyician* نیز بیانگر آمدن نام پزشک هستند.

```
(dr\.{0,1}|doctor|physician|surgen)s{0,1}\s*\[\*{2}(.+?)\*{2}\]
```

من سعی داشتم تا بتوانم از لکسیکان وردنت برای پیدا کردن معادل کلمات دکتر استفاده کنم تا آنها را در الگو شامل کنم که البته معادل خاصی بدست نداد.

```
"Doctor Name": "Doctor not found"
```

تاریخ ویزیت

در ابتدای هر سند دو تاریخ درج شده است، تاریخ مراجعه و ترخیص. برای به دست آوردن تاریخ مراجعه از الگوی زیر استفاده می‌کنیم.

```
admission date:\s*\[\*{2}(.+?)\*{2}\]
```

توجه کنید که تاریخ مراجعه با عبارت *Admission Date* شروع شده و مقدار آن نیز داخل براکت و دو ستاره قرار می‌گیرد.

```
"Admission Date": "2174-4-18"
```

بیماری های تشخیص داده شده

در هر سند قسمتی به عنوان *Discharge Diagnosis* وجود دارد که بیماری‌های تشخیص داده شده در این قسمت ذکر می‌شوند. این بخش معمولا پیش از قسمت *Discharge Condition* قرار می‌گیرید. بنابراین تمام خطوطی بین دو عبارت فوق را در مرحله اول انتخاب می‌کنیم. البته ممکن است در بعضی سندها عبارت پایانی وجود نداشته باشد، بنابراین از ۴ خط خالی به عنوان نشانگر پایانی الگو نیز استفاده می‌کنیم.

```
discharge diagnosis:(.*?)(\n{4,}|discharge)
```

در مرحله دوم متن انتخاب شده را بر اساس کاراکتر *newline* توکنایز کرده و اگر در هر خط علامت ویرگول , وجود داشت طبق آن نیز توکنایز می‌کنیم. دلیل توکنایز کردن دو مرحله‌ای این است که اسناد موجود اکثرا به دو شکل نام بیماری‌ها را لیست کرده‌اند. در یک روش هر بیماری در یک خط قرار می‌گیرد و در روش دیگر در هر خط چند بیماری نوشته شده و توسط ویرگول جدا می‌شوند. بنابراین استفاده از توکنایزرهای پیشفرض مانند *nltk sentence tokenizer* و *nltk word tokenizer* نتیجه‌ی خوبی نمی‌دهند و برای جدا کردن نام بیماری‌ها باید همان طور که گفتم عمل کرد.

```
"Diagnosed Diseases": [
        "Abdominal Pain",
        "Acute on chronic renal failure",
        "Systolic Heart failure",
        "ALL",
        "History of embolic stroke"
    ]
```

سن بیمار

معمولا سن بیمار شامل یک عدد است که بلافاصله با عبارتی مانند *years old* همراه می‌شود. به الگوی زیر توجه بفرمایید.

```
(\d+)\s*(year old|years old|yo|y\/o)
```

به جز عبارت ذکر شده، اصطاحات مخفف دیگری نیز مانند *yo, y/o* نیز ممکن است نشانگر سن بیمار باشند.

```
"Age": "38"
```

وزن بیمار

این ویژگی در اکثر اسناد وجود ندارد. من دستی تعداد خوبی از آنها را بررسی کردم اما اکثرا فاقد وزن بیمار هستند. به هر ترتیب، معمولا وزن یک عدد است که دنبالش پسوندهای مانند ... *Kg, Lbs,* می‌آید.

```
(\d+\.{0,1}\d*)\s*(kg|kilo|kilogram|pound|lb)s{0,1}
```

گاهی ممکن است در متن خود وزن نباشد اما تغییر وزن ذکر شده باشد، مثلا فلان بیمار در هفته گذشته ۳ کیلو کاهش وزن داشته است. الگوی فوق تمامی این موارد را انتخاب می‌کند، سپس در مرحله بعد بزرگترین عدد را به عنوان وزن انتخاب می‌کنیم. البته که روش *Naive* ای است. ولی خب فکر روش‌های قاعده محور معمولا در همین حد کارا هستند.

```
"Weight": "3.0"
```

نوع مشکل و محل آسیب دیدگی

ویژگی‌های ذکرشده شماره ۴ و ۶ و ۷ کمی گیج کننده هستند. از آن جهت که پیشتر نام بیماری‌ها را استخراج کرده ایم (شماره ۴) اما در شماره ۶ نیز اسم بیماری خواسته شده. از طرفی فرق بیماری و آسیب دیدگدگی نیز مبهم است. محل بیماری اما تعریف مشخصی دارد و برای آن می‌توان یک راه حل پیش پا افتاده ارائه داد. ابتدا یک دیکشنری مختصر از اعضای بدن تشکیل دادم.

```
body_parts =
{"head" :
    "eye ear nose mouth forehead eyebrow lip cheek chin tongue tooth
       jaw deaf blind brain",
"upper_body" :
    "shoulder arm chest back elbow wrist hand finger thumb neck ribs
       abdomen waist",
"lower_body" :
    "hip leg knee thigh foot calf ankle toe buttocks groin heel shin
       ",
"digestive_system" :
    "stomach liver intenstine kidney bladder pancreas spleen
       esophagus gallbladder bowel rectum rectal renal",
"respiratory_system" :
    "lungs throat thrachea bronchi diaphragm alveoli nostrils larynx
       pharynx pleura",
"circulatory_system" :
    "heart vein capillaries aorta artery hypertension strock vascular
       "}
```

سپس بیماریهای تشخیص داده شده در مرحله ۴ را با استفاده از دیکشنری فوق دسته بندی کردم. مثلا عبارت *Heart* مربوطه به سیستم عروقی بدن است.

```
"Disease Location": [
      "circulatory_system",
      "head",
      "digestive_system"
   ]
```

درمان و دارو

دو قسمتی اصلی مربوط به تجویز دارو در اسناد وجود دارد. یکی برای زمان بستری شدن به نام *Medications on Admission* و دیگری برای زمان ترخیص به نام *Discharge Medications* . ابتدا متن این دو قسمت را انتخاب کرده سپس در آنها به دنبال دارو می‌گردیم.

```
medications on admission:(.*?)(\n{4,}|discharge)
discharge medications:(.*?)(\n{4,}|discharge)
```

نام داروها معمولا با حرف بزرگ شروع می‌شود، ممکن است شامل چند کلمه جدا باشد سپس دوز دارو به عدد و در نهایت واحد آن نوشته می‌شود. البته الگوی مورد اشاره بسیار محدود است و تعداد زیادی از داروها را کشف نمی‌کند. بنابراین آنرا ساده‌تر در نظر می‌گیریم.

```
[a-z][a-z\- ]+)\s[\d\.\-]*\s*
(g|mg|mgs|mcg|unit|U|QAM|QPM|%|prn|MDI|daily|weekly|monthly|qd|once|
    twice)
```

اگر الگوی فوق دارویی پیدا نکرد یعنی داروها بدون واحد نوشته شده اند و از الگوی زیر استفاده می‌کنیم.

```
([a-z][a-z\- ]+)\s[\d\.\-]+
```

اگر بازهم چیزی پیدا نشد یعنی حتی دوز دارو ذکر نشده و صرفا اسم داروها نوشته شده، در این حالت بر اساس ویرگول توکنایز کرده و به مرحله بعد می‌فرستیم. در انتها برای تمیز شدن نتایج برخی پردازش متنی مانند حذف کلمات پرتکرار که ممکن است به عنوان دارو تشخیص داده شده باشند را انجام می‌دهیم (*By mouth, inhale ...*) و نتیجه به خروجی ارسال می‌شود.

```
"Medication on Admission": [
        "Carvedilol",
        "Morphine",
        "Valsartan",
        "Torsemide",
        "Multivitamin",
        "Albuterol",
        "Lorazepam",
        "Warfarin",
        "Ondansetron",
        "tid",
        "Pentamidine",
        "Colace"
    ]
```

```
"Medication on Discharge": [
        " fluticasone-salmeterol",
        " docusate sodium",
        " lorazepam",
        " albuterol sulfate",
        " Zofran",
        " metoprolol succinate",
        " hr",
        " morphine",
```

```
        " dicyclomine",
        " allopurinol",
        " sulfamethoxazole-trimethoprim",
        " acyclovir",
        " torsemide",
        " simethicone",
        " as needed abdominal pain or",
        " mycophenolate mofetil",
        " prednisone",
        " magnesium hydroxide",
        " morphine",
        " warfarin"
    ]
```

همچنین در اسناد بخشی وجود دارد که در نوع عمل جراحی را ذکر می‌کند، البته که لزوما همه بیماران جراحی نشده اند. این بخش با عبارت *Major Surgical or Invasive Procedure* مشخص می‌شود. به طور مثال:

```
"Operation Type": [
        "Upper GI series with small bowel follow through",
        "Right heart catheterization",
        "IR guided paracentesis"
    ]
```

تاریخچه بیماری

ابتدا بنا داشتم از با استفاده از عباراتی مانند *History of, Prior use, Previous, Past...* گذشته بیمار را استخراج کنم. اما خود اسناد دو بخش مخصوص سابقه بیمار را مشخص کرده‌اند. یکی از آنها سابقه سلامتی کلی بیمار را نوشته که با عبارت *Past Medical History* مشخص شده. دیگری سابقه بیماری فعلی مراجعه کننده را نوشته و دارای تگ *History of Present Illness* می‌باشد. بنابراین احتمالا دقیق‌ترین حالت انتخاب همین بخش‌ها به عنوان پیشینه‌ی بیماری مراجعه کننده خواهد بود.

```
"Medical History": " ONCOLOGIC HISTORY: ALL: - initially presented in
    [**2172-8-5**] ..."
```

اطلاعات بیشتر

در سند اطلاعات زیادی وجود دارد، شاید مناسب‌ترین آنها قسمت *Social History* باشد.

```
{
    "Patient Name": "First Name3 (LF) 3918",
    "Doctor Name": "Doctor not found",
    "Admission Date": "2174-4-18",
    "Diagnosed Diseases": [
        "Abdominal Pain",
        "Acute on chronic renal failure",
        "Systolic Heart failure",
        "ALL",
        "History of embolic stroke"
    ],
    "Age": "38",
    "Weight": "3.0",
    "Disease Type": "",
    "Operation Type": [
        "Upper GI series with small bowel follow through",
        "Right heart catheterization",
        "IR guided paracentesis"
    ],
    "Disease Location": [
        "digestive_system",
        "head",
        "circulatory_system"
    ],
    "Medication on Admission": [
        " Carvedilol",
        " Morphine",
        " Valsartan",
        " Torsemide",
        " Multivitamin",
        " Albuterol",
        " Lorazepam",
        " Warfarin",
        " Ondansetron",
        " tid",
        " Pentamidine",
        " Colace"
    ],
    "Medication on Discharge": [
        " fluticasone-salmeterol",
        " docusate sodium",
        " lorazepam",
        " albuterol sulfate",
        " Zofran",
        " metoprolol succinate",
        " hr",
```

```
        " morphine",
        " dicyclomine",
        " allopurinol",
        " sulfamethoxazole-trimethoprim",
        " acyclovir",
        " torsemide",
        " simethicone",
        " as needed abdominal pain or",
        " mycophenolate mofetil",
        " prednisone",
        " magnesium hydroxide",
        " morphine",
        " warfarin"
    ],
    "Medical History": " ONCOLOGIC HISTORY: ALL: - initially presented
        in [**2172-8-5**] ... REMOVED TO MAKE ROOM ",
    "Illness History": " 38 yo F w/ h/o ALL in remission s/p cord
       transplant in [**1-13**], ... REMOVED TO MAKE ROOM ",
    "More Info": " Smoke: never EtOH: Occasional in past, none
       currently Drugs: Never Lives/ ... REMOVED TO MAKE ROOM "
}
```