Assignment NO.2 Solutions

NLP | Fall 1401 | Dr.Minayi

Teacher Assistant:

Farbod Davoodi

Student name : **Amin Fathi**

Student id : **400722102**

Problem 1

Text classification که به‌عنوان برچسب‌گذاری متن یا طبقه‌بندی متن شناخته می‌شود، فرآیند طبقه‌بندی متن به گروه‌های سازمان یافته است. با استفاده از پردازش زبان طبیعی ، طبقه‌بندی‌کننده‌های متن می‌توانند به طور خودکار متن را تجزیه و تحلیل کنند و سپس مجموعه‌ای از برچسب‌ها یا دسته‌های از پیش تعریف شده را بر اساس محتوای آن اختصاص دهند.

طبقه‌بندی متن از سه جز اصلی زیر تشکیل شده است:

۱. آماده‌سازی مجموعه داده‌ها: مرحله اول مرحله آماده‌سازی مجموعه داده است که شامل فرآیند بارگذاری مجموعه داده و انجام پیش‌پردازش اولیه است. سپس مجموعه داده به مجموعه‌های آموزشی و اعتبارسنجی تقسیم می‌شود.

۲ .مهندسی ویژگی: مرحله بعدی، مهندسی ویژگی است که در آن مجموعه داده خام به ویژگی‌های مسطح تبدیل می‌شود که می‌تواند در یک مدل یادگیری ماشین استفاده شود. این مرحله همچنین شامل فرآیند ایجاد ویژگی‌های جدید از داده‌های موجود است.

۳. آموزش مدل: مرحله نهایی، مرحله ساخت مدل است که در آن یک مدل یادگیری ماشین بر روی یک مجموعه داده برچسب‌گذاری شده آموزش داده می‌شود.

طبقه‌بندی متن در حال تبدیل شدن به بخش مهمی از کسب‌وکارها است؛ زیرا به شما امکان می‌دهد به راحتی بینش‌هایی از داده‌ها دریافت کنید و فرآیندهای تجاری را خودکار کنید. برخی از رایج ترین مثال‌ها و موارد استفاده برای طبقه بندی خودکار متن عبارتند از:

- تجزیه و تحلیل احساسات: فرآیند درک اینکه آیا یک متن معین به طور مثبت یا منفی در مورد یک موضوع خاص صحبت می‌کند (به عنوان مثال برای اهداف نظارت بر برند).
- تشخیص موضوع: وظیفه‌ی شناسایی موضوع یک قطعه از متن (مثلاً هنگام تجزیه و تحلیل بازخورد مشتری، بدانید که بررسی محصول در مورد سهولت استفاده، پشتیبانی مشتری یا قیمت‌گذاری است).
- تشخیص زبان: روش تشخیص زبان یک متن داده شده (به عنوان مثال، بدانید که آیا یک بلیط پشتیبانی دریافتی به زبان انگلیسی یا اسپانیایی برای مسیریابی خودکار بلیط‌ها به تیم مناسب نوشته شده است).

Problem 2

ترجمه ماشینی یا MT یا تفسیر روباتیک صرفاً رویه‌ای است که یک نرم‌افزار رایانه‌ای متنی را بدون مشارکت انسان از یک زبان به زبان دیگر ترجمه می‌کند. در سطح اساسی خود، ترجمه ماشینی جایگزینی مستقیم از کلمات اتمی در یک زبان مشخصه برای کلمات در زبان دیگر انجام می دهد.

با استفاده از روش‌های پیکره، ترجمه‌های پیچیده‌تری را می‌توان انجام داد، با در نظر گرفتن برخورد بهتر تضادها در گونه‌شناسی آوایی، تصدیق بیان و ترجمه اصطلاحات، درست مانند جداسازی عجایب. در حال حاضر، برخی از سیستم ها مانند یک مترجم انسانی قادر به انجام کار نیستند، اما در آینده این امکان نیز وجود خواهد داشت.

چهار نوع ترجمه ماشینی وجود دارد:

۱ .ترجمه ماشینی آماری ((Statistical Machine Translation (SMT)

با اشاره به مدل‌های آماری که به بررسی حجم عظیمی از محتوای دو زبانه بستگی دارد، کار می‌کند. انتظار دارد که مطابقت بین یک کلمه از زبان مبدأ و یک کلمه از زبان هدف را تعیین کند. یک نمونه واقعی آن Google Translate است. در حال حاضر، SMTبرای ترجمه پایه فوق‌العاده است، با این حال مهم‌ترین نقطه ضعف آن این است که به متن توجه نمی‌کند که به این معناست که ترجمه می‌تواند مرتباً اشتباه باشد یا می‌توان گفت، انتظار ترجمه با کیفیت عالی را نداشته باشید. انواع مختلفی از مدل‌های ترجمه ماشینی مبتنی بر آمار وجود دارد که عبارتند از: ترجمه مبتنی بر عبارت سلسله مراتبی، ترجمه مبتنی بر نحو، ترجمه مبتنی بر عبارت، ترجمه مبتنی بر کلمه.

۲ .ترجمه ماشینی مبتنی بر قانون ((Rule-based Machine Translation (RBMT)

RBMT اساساً اصول قواعد گرامری را ترجمه می‌کند. برای ایجاد جمله ترجمه شده، یک بررسی دستوری از زبان مبدأ و زبان هدف را هدایت می‌کند. اما، RBMTنیاز به ویرایش گسترده دارد و اتکای قابل توجه آن به فرهنگ لغت نشان می‌دهد.

۳ .ترجمه ماشینی ترکیبی ((Hybrid Machine Translation (HMT)

HMTترکیبی از RBMT و SMT است. از یک حافظه ترجمه استفاده می‌کند که بدون شک از نظر کیفیت موفق‌تر است. با این وجود، حتی HMT دارای معایب زیادی است که بزرگ‌ترین آن‌ها نیاز به ویرایش بسیار زیاد است و مترجمان انسانی نیز مورد نیاز خواهند بود. رویکردی مانند تولید قوانین آماری برای HMT وجود دارد.

۴ .ترجمه ماشین عصبی ((Neural Machine Translation (NMT)

NMT نوعی ترجمه ماشینی است که برای ساخت مدل‌های آماری با هدف نهایی ترجمه بر مدل‌های شبکه عصبی (بر اساس مغز انسان) متکی است. مزیت اساسی NMT این است که یک سیستم انفرادی ارائه می‌دهد که می‌تواند برای باز کردن متن منبع و

مقصد آماده شود. متعاقباً، به سیستم‌های خاصی که برای سایر سیستم‌های ترجمه ماشینی، به ویژه SMT، منظم هستند، متکی نیست.

Problem 3

a) ^[A-Z].*f$
b) ^.*4.*4.*4.*$
c) ^[1|3|5|7|9].*[a-z].*[0|2|4|6|8]$
d) ^[0-9]$|^[1-9][0-9][0-9]$|1000$|^(?=\d{2}$)1?2?3?4?5?6?7?8?9?$

Problem 4

```
[1] import re
```

```
[4] def validate_email(email):
      regex = re.compile(r'^[a-zA-Z1-9\.\_]+@[a-zA-Z1-9]+\.[a-zA-Z]{3}$')
      if re.match(regex,email):
        Res = True
      else:
        Res = False
      return Res

    print(validate_email('username@domain.tld'))
```

```
    True
```

```
[5] def validate_phone(number):
      regex = re.compile(r'^09[0-9]{9}$|^\+989[0-9]{10}$|^0098[0-9]{10}$')
      if re.match(regex, number):
        Res = True
      else:
        Res = False
      return Res
```