به نام خدا



دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس پردازش زبان طبیعی

استاد ممتازی

تمرین سوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: ابهام‌زدایی معنایی کلمات

پیش از ارائه نتایج نکات کلی در مورد پیاده‌سازی و مجموعه‌داده را بیان می‌کنم:

* مجموعه آموزش اولیه را به دو قسمت اعتبارسنجی (۲۰٪) و آموزش (۸۰٪) شکاندم تا از مجموعه اعتبارسنجی برای تعیین ابرپارامتر‌های مدل دسته‌بند استفاده کنم.
* پیاده‌سازی BERT استفاده‌شده تنها می‌تواند برای کلمات یک پنجره کوچک تعبیه ارائه دهد و مابقی کلمات را دور می‌ریزد؛ لذا متن کامل به مدل داده نمی‌شود و تنها چند کلمه قبل و چند کلمه به مدل داده می‌شود و بردار کلمه هدف به دسته‌بند داده می‌شود.
* برخی از داده‌ها دارای خطا است؛ یعنی ممکن است مفهوم نامرتبطی برای یک کلمه مبهم ارائه شده باشد. این متون چه در آموزش و چه در ارزیابی کنار کذاشته می‌شود. در جدول زیر مشخص شده است که برای شکل‌های مختلف یک کلمه مختلف چه مفاهیم در نظر گرفته شده است:

|  |  |  |
| --- | --- | --- |
| کلمه | اشکال مختلف کلمه | مفاهیم کلمه |
| Hard | hard, harder, hardest | HARD1, HARD2, HARD3 |
| Interest | interest, interested, interesting, interests | Interest1, interest2, interest3, interest4, interest5, interest5 |
| Line | line, lines | cord, division, formation, phone, product, text |
| Serve | serve, served, serves | SERVE10, SERVE12 |

* برای هر کلمه مبهم یک SVM جداگانه آموزش داده شده است که کارش تشخیص مفهوم آن کلمه مبهم است.
* برای هر SVM‌ از کرنل‌های مختلف خطی، چندجمله‌ای و RBF استفاده شده است و بهترین دسته‌بند بر اساس صحت روی مجموعه اعتبارسنجی انتخاب شده است.
* برای کلمه Hard تنها مفهوم HARD1 در مجموعه آموزش آمده است؛ لذا دسته‌بند SVM‌ نمی‌توان برای آن آموزش داد. در این شرایط به ازای تمام داده‌ها همان مفهوم HARD1 پیش‌بینی می‌شود.
* معیارهای دقت به ازای هر کلمه مبهم ارائه شده است و در نهایت با میانگین وزن‌دار دقت‌های نهایی نیز ارائه شده است.

آمار مربوط به هر کلمه در هر مجموعه‌داده در جدول زیر آورده شده است:

|  |  |  |  |
| --- | --- | --- | --- |
| کلمه | تعداد داده آموزش | تعداد داده اعتبارسنجی | تعداد داده آزمون |
| Hard | ۱۳۳۴ | ۳۲۷ | ۳۰۰ |
| Interest | ۱۳۴۳ | ۳۳۷ | ۲۸۵ |
| Line | ۱۱۸۵ | ۲۷۰ | ۲۳۶ |
| Serve | ۱۱۱۲ | ۳۰۴ | ۲۷۶ |
| کلی | ۴۹۷۴ | ۱۲۳۸ | ۱۰۹۷ |

بهترین تنظیم برای هر کلمه در جدول زیر آورده شده است:

|  |  |
| --- | --- |
| کلمه | بهترین تنظیم |
| Hard | - |
| Interest | کرنل خطی |
| Line | کرنل RBF |
| Serve | کرنل خطی |

صحت هر دسته‌بند و صحت کلی در جدول زیر آورده شده است:

|  |  |  |  |
| --- | --- | --- | --- |
| کلمه | مجموعه آموزش | مجموعه اعتبارسنجی | مجموعه آزمون |
| Hard | 100.00٪ | 100.00٪ | 100.00٪ |
| Interest | 100.00٪ | 91.69٪ | 89.92٪ |
| Line | 97.38٪ | 92.59٪ | 94.49٪ |
| Serve | 100.00٪ | 100.00٪ | 100.00٪ |
| کلی | 99.38٪ | 96.12٪ | 96.17٪ |

دقت F1 هر دسته‌بند و دقت F1 کلی در جدول زیر آورده شده است:

|  |  |  |  |
| --- | --- | --- | --- |
| کلمه | مجموعه آموزش | مجموعه اعتبارسنجی | مجموعه آزمون |
| Hard | 100.00٪ | 100.00٪ | 100.00٪ |
| Interest | 100.00٪ | 62.96٪ | 74.61٪ |
| Line | 97.20٪ | 92.74٪ | 94.09٪ |
| Serve | 100.00٪ | 100.00٪ | 100.00٪ |
| کلی | 99.33٪ | 88.33٪ | 92.13٪ |