بسمه تعالی

امیرمحمد کاظمینی زاده

۹۴۵۲۱۱۷۱

گزارش قسمت دوم کلاسه‌بندی جملات

قسمت اول:

به طور خلاصه ما سه مرحله اساسی داریم :

جدا سازی کلمات و آماده‌سازی متون. تمامی مراحل قسمت الف در اینجا با بهبود عملکرد، اعمال می‌شوند(رجوع شود به گزارش قبل)

در مرحله دوم ریشه کلمات یافت می‌شود و قسمت‌های اضافی توکن ها حذف می‌شوند.

در مرحله‌ی سوم، stop_words های نهایی حذف می‌شوند.

منابع:

https://github.com/jonsafari/perstem

https://github.com/mhbashari/awesome-persian-nlp-ir

پس از آن مراحل train و test را انجام می‌دهیم. نکته‌ی لازم به ذکر آن است که آموزش به طور semi lazy انجام می‌گیرد.

به علت کم بودن حجم دیتا ۱۰ مرتبه مراحل آموزش و تست را انجام می‌دهیم (تقریبا مانند Cross-validation) و سپس میانگین امتیازات بدست آمده را گزارش می‌کنیم. نکته‌ی مهم آن است که به ازای هرسه مرحله‌ی pipe امان، یک بار سلسله مراحل تست و آموزش انجام می‌شود و نتایج به تفکیک مراحل است:

Clean: پس از مرحله ۳

After1: پس از مرحله ۱

After2: پس از مرحله ۲

نتایج تست ها بدون cross-validation در testing و همراه cross-validition در testing2 موجود است. به‌علت تنوع و زیادی تست‌ها و کلاس ها از آوردن آن‌ها در گزارش خودداری می‌کنیم.

نکته‌ی مهم دیگر آن‌که مراحل تمیزکردن متون وقتی کمتر می‌شود، دقت بالاتر می‌رود. به‌علت وجود داده‌ی بسیارکم، احتمالا تعداد زیادی از کلما جزو stop-word محسوب شده و همچنان ریشه‌یابی باعث از بین‌رفتن اصطلاحات یکتا اما مشابه نامزدها می‌شود.

در خصوص کلمات موثرتر، همانطور که در ارائه آورده شد،با تکرار بیشتر انحصاری رابطه مستقیم دارد. که برای آقای روحانی: ایران، آینده، جامعه، الان، ایجاد، نفت و ... می‌باشد. برای آقای رئیسی: باید، روحانی، کشور، حتما، اسلامی، ورزش، امنیت و...

قسمت ب:

دستور train :

vw -d input -c --passes 10 -f model_output --ngram number --loss_function function

دستور تست:

vw -d test_data -t -i model_input -p report_addr

نکته: آموزش و تست داده را روی داده‌ی after2 انجام داده‌ام. اطلاعات جزئی تر در vowpal res موجود است.

Loss_function	ngram	رئیسی		روحانی	
		precision	recall	precision	recall
HINGE	1	۰/۶۵۳۸	۰/۷۰۸۳	۰/۶۹۵۶۵	۰/۶۴
	2	۰/۶۶۶۶	۰/۶۶۶۶	۰/۶۸	۰/۶۸
	3	۰/۶۰۸۶	۰/۵۸۳۳	۰/۶۱۵۳	۰/۶۴
LOGISTIC	1	۰/۶۹۵	۰/۶۴	۰/۶۹۵۶	۰/۶۴
	2	۰/۶۶۶۶	۰/۷۲	۰/۶۶۶	۰/۷۲
	3	۰/۶۴۲	۰/۷۲	۰/۶۴۲	۰/۷۲

دقیق‌ترین گزینه : 2gram + LOGISTIC loss

برای حالت after1 هم با 2gram + LOGISTIC loss مراحل را تکرار کردم. نتایج ضعیف‌تری حاصل شد.