

خلاصه‌سازی اخبار

بهار ۱۴۰۱

فهرست مطالب

-
1. معرفی مجموعه دادگان
 - ✓ مجموعه دادگان تسنیم
 - ✓ مجموعه دادگان عصر ایران
 - ✓ پیش پردازش
 2. مدل‌های پیاده‌سازی شده
 - ✓ مدل MT5
 - ✓ مدل Bert2Bert
 - ✓ مدل WikiBert2WikiBert
 3. برنامک تحت وب پیاده‌سازی شده
 4. نمایش خروجی‌ها
 - ✓ ارائه برنامک تحت وب پروژه
 - ✓ ارائه گیت‌هاب پروژه
 - ✓ ارائه داک پروژه

مجموعه دادگان تسنیم



- حاصل خزش بر روی سایت خبری تسنیم.
- موارد خزش شده، شامل متن، تیتر، خلاصه، حوزه و زمان انتشار خبر می‌شود.
- حوزه‌های خبری‌ای چون؛ سیاسی، اجتماعی، ورزشی، بین‌الملل و ... را پوشش می‌دهد.
- استفاده از زبان برنامه‌نویسی Python و کتابخانه‌ی Scrapy برای خزش.
- شامل بیش از ۶۷ هزار خبر.
- برای یادگیری مدل‌ها صرفاً از ستون متن و خلاصه‌ی خبر استفاده شده است.

مجموعه دادگان عصر ایران



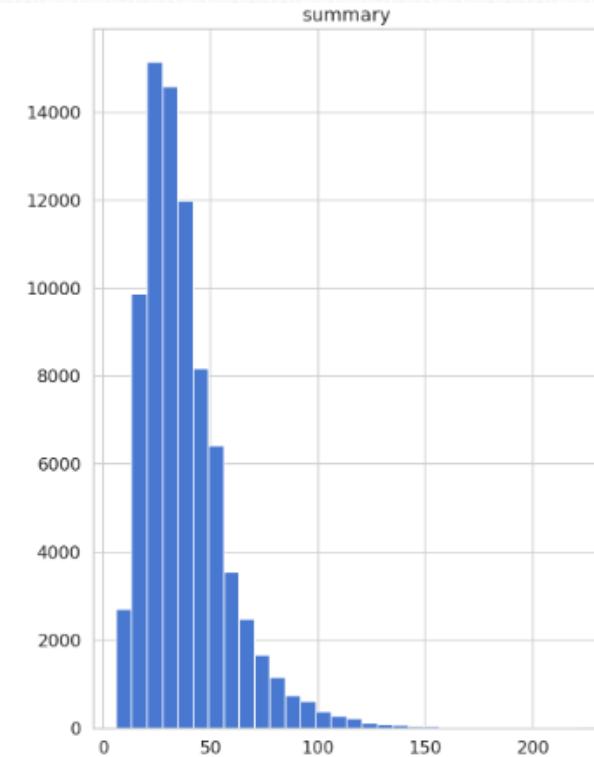
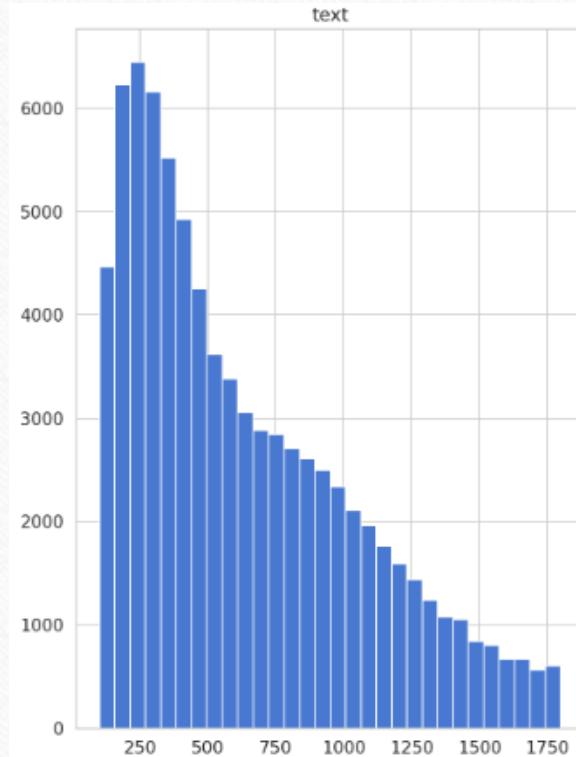
- استفاده از کتابخانه Scrappy جهت انجام عملیات خزش به صورت چند نخی
- استفاده از صفحه آرشیو سایت خبرگزاری عصر ایران جهت خزش اخبار شامل انواع مختلف خبر از جمله اخبار اجتماعی، اقتصادی و فرهنگی
- ذخیره ۱۵۰,۰۰۰ رکورد خبر و خلاصه و انتشار کد خزش جهت استخراج اخبار به روزتر و یا ایجاد مجموعه دادگان بزرگتر
- انجام پیش‌پردازش (در ادامه توضیح داده می‌شود) برروی مجموعه داده و خلاصه‌سازی خبر با استفاده از قسمت خبر و خلاصه ایجاد شده.

پیش‌پردازش دادگان



- استفاده از دادگان BBC، تبیان و pn-summary علاوه بر تسنیم و عصر ایران
- یکسان‌سازی حروف عربی و فارسی و جایگزینی حروف عربی با فارسی مانند: "ی" و "ي"
- حذف کاراکترهای خاص که تاثیری در پردازش ندارند. مانند: \xa0
- حذف space اضافه در متن
- نرمالسازی داده با استفاده از کتابخانه هضم
- محاسبه ۹۰ درصد طول اخبار و خلاصه آن‌ها و حذف اخبار خیلی بلند و خلاصه‌های خیلی کوتاه
- بررسی کوتاه‌تر بودن خلاصه نسبت به خبر
- محاسبه تعداد توکن‌های خبر و خلاصه و در نظر گرفتن دو ستون در دادگان برای آن‌ها

نمونه‌ای از نمودار رسم شده برای دادگان



مدل mt5

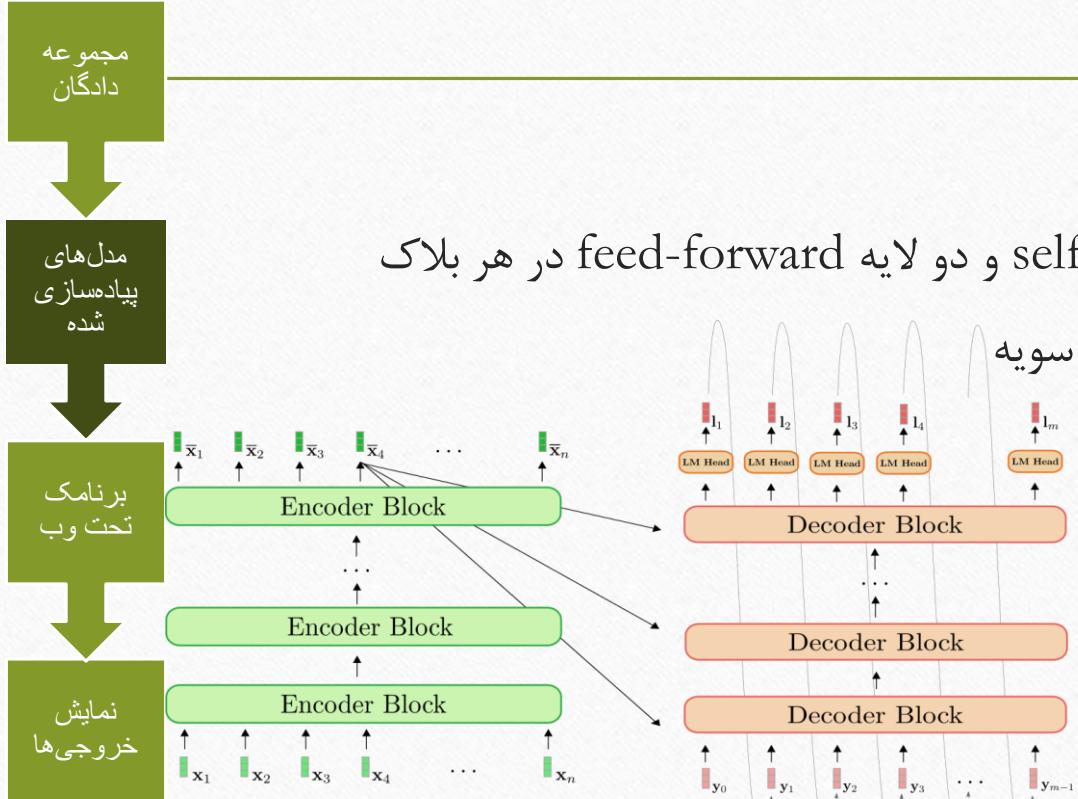


	Rough-1	Rough-2	Rough-1
Pn-summary	31.61	14.64	26.27
BBC	22.98	6.17	18.48
merged	25.53	8.66	20.38

- استفاده از وزن‌های مدل mt5-small •
- مبتنی بر رمزگذار- رمز گشا •
- استفاده از دادگان تسنیم، BBC و pn-summary •

پارامتر	مقدار
#Beams	2
Length Penalty	2.0
Early Stopping status	ACTIVE
Repetition Penalty	1.0

BERT2BERT مدل



اضافه کردن لایه cross-attention بین لایه self-attention و دو لایه feed-forward در هر بلاک

- یک مدل رمزنگار-رمزگشا Warm-starting

- تبدیل لایه‌های دوسویه self-attention به لایه‌های یک سویه

- اضافه کردن لایه مدل زبانی به بالای رمزگشا

برای ساختن توزیع احتمال شرطی

Model	R1	R2	RL
ParsBERT-sum	85/32	45/17	5/28

WikiBert2WikiBert مدل

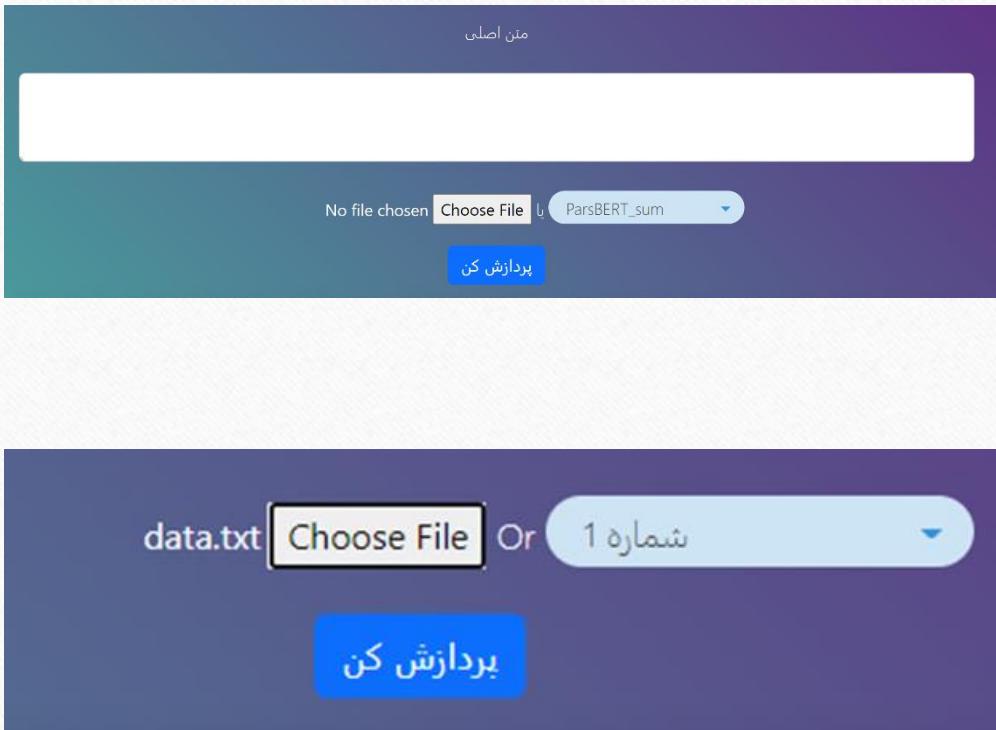


- استفاده از وزن‌های مدل زبانی برتر در حالت Fine-tune شده بر روی مجموعه دادگان ویکی‌پدیا با ساختار رمزنگار و رمزگشا
- نتایج آموزش مدل بر روی دادگان BBC و PN-Summary

	Rough-1	Rough-2	Rough-1
Pn-summary	35.51	15.65	30.91
BBC	32.75	9.72	27.80

پارامتر	مقدار
Num_train_epochs	5
Warm_up_steps	500
Per_device_eval_batch_size	1
Per_device_train_batch_size	1
Eval_steps	500
Gradient_accumulation_steps	16

پیادهسازی برنامک تحت وب



- ✓ امکان ایجاد حساب کاربری و ورود و خروج
- ✓ امکان انتخاب میان مدل‌های متفاوت آموزش داده شده
- ✓ امکان آپلود فایل خبر
- ✓ ذخیره خروجی‌های مدل‌ها با ثبت متن اصلی، خلاصه، شماره مدل مورد استفاده و مدت زمان پردازش

نمایش خروجی‌ها(پایان ارائه)

متن اصلی

به گزارش شانا، علی کاردر امروز (۷۶ دی ماه) در مراسم تودیع محسن قمصری، مدیر سابق امور بین الملل شرکت ملی نفت ایران و معارفه سعید خوشرو، مدیر جدید امور بین الملل این شرکت، گفت: مدیریت امور بینالملل به عنوان یکی از تاثیرگذارترین مدیریت‌های شرکت ملی نفت ایران در دوران

No file chosen Choose File Or شماره ۱

برداش کن

Processing time	خلاصه	شماره مدل nlp	متن اصلی
2.05793	مدیر سابق امور بین الملل و بین الملل شرکت ملی نفت ایران در دوران تحریم به دلیل پیرامون تحریم وظیفه اصلی این شرکت ملی نفت ایران و معا...	1	به گزارش شانا، علی کاردر امروز (۷۶ دی ماه) در مراسم تودیع محسن قمصری، مدیر جدید امور بین الملل شرکت ملی نفت ایران در دوران تحریم به دلیل پیرامون تحریم وظیفه اصلی این شرکت ملی در دوران تحریمهای طالمانه غرب ایران بود.

نتیجه

مدیر سابق امور بین الملل و بین الملل شرکت ملی نفت ایران در دوران تحریم به دلیل پیرامون تحریم وظیفه اصلی این شرکت ملی در دوران تحریمهای طالمانه غرب ایران بود.

20.05793 -

بازگشت به صفحه ای قبل