دانشکده مهندسی کامپیوتر

گزارش فاز اول

بیان دیوانی آذر ۹۸۵۲۲۲۲۹

نیم‌سال دوم

سال تحصیلی ۰۲-۱۴۰۱

آماده سازی داده آموزشی

از وبسایت وبتون[۱] چپتر هارا دانلود میکنیم و در هر صفحه OCR را اجرا میکنیم و متن هر چپتر را استخراج میکنیم و در فرمت [chapter_number].txt ذخیره میکنیم.
از آنجایی که طبق سیاست وبسایت وبتون، تصاویر وبتون باید سایز canvas مشخصی داشته باشد و این سایز کوچک است. آرتیست ها روی سایز canvas بزرگتری کار میکنند و سپس آن را برش میدهند. به همین دلیل بعضی از متن ها بین دو تصویر نصف میشوند. برای وصل کردن تصاویر باید وجود نوشته در پایین یا بالای تصویر را تشخیص میدادیم. برای اینکار چک کردیم در پیکسل های انتهایی، پیکسل های مشکی درون پیکسل های سفید رنگ محاصره شده اند یا نه. با این روش توانستیم تقریبا تمام متون دو نیمه شده را به هم وصل کنیم.

در قسمت تمیز کردن داده اشتباهات متداول ocr را اصلاح کردیم. برای مثال تعویض $ با S و همچنین خطوط تک واژه که اهمیتی در مدل NER ما ندارند را حذف کردیم.

سپس به کمک ابزار ntlk و تابع word_tokenize متن را به واژه ها میشکانیم.
و سپس به کمک ابزار spacy جملات را تشخیص میدهیم.

برچسب گذاری

برای برچسب گذاری ابتدا از این وب ابزار مخصوص spacy استفاده کردیم که متاسفانه باید کل کار را دستی انجام میدادیم و به همین دلیل خیلی زمانبر و انرژی بر بود.

https://tecoholic.github.io/ner-annotator

و دنبال روش جدید گشتیم. تگ های NER را از سایت فندوم وبتون [۲] crawl کردیم.
و سپس این کلمه هارا در تمام متون سرچ کردیم و index شروع و پایان با تگ مورد نظر را در فایلی به نام train_data.json ذخیره کردیم.

---

[۱] webtoon.com

[۲] purple-hyacinth.fandom.com

تعداد جملات قبل تمیز شدن و برچسب گذاری: ۹۵۴۰

تعداد جملات بعد تمیز شدن و برچسب گذاری: ۱۱۷۷

آمار داده ها

أ. تعداد «واحد» داده ، ب. تعداد جملات

Total	PERSON	LOC
9539	1124	78

ج. تعداد کلمات

Total	PERSON	LOC
78421	2020	94

د. تعداد کلمات منحصر به فرد

Total	PERSON	LOC
8975	131	13

و. ۱۰کلمه پرتکرار غیر مشترک هر برچسب

Tag	word1	word2	word3	word4	word5	word6	word7	word8	word9	word10
PERSON	sinclair	hawkes	phantom	scythe	leader	ladell	sake	lauren	anslow	hermann
LOC	allendale	allendale train station	golden clover	the carmine camelia	greychapel	grim goblin	l'arlequin	docks	tower of ardhalis	Allendale

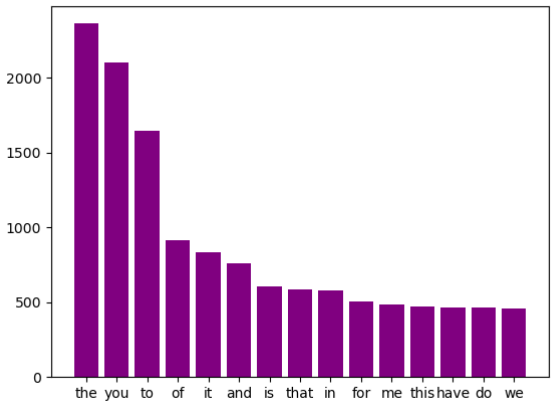
ز. ده کلمه مشترک برتر براساس frequency normalized relative

	word1	word2	word3	word4	word5	word6	word7	word8	word9	word10
PERSON/LOC	leader	have	your	this	not	lauren	sinclair	hermann	kieran	police
LOC/PERSON	greychapel	tragedy	l'arlequin	allendale	grim	goblin	camelia	band	foot	entrance

ح. ده کلمه برتر بر اساس TF_IDF(W)

	Tag	word1	word2	word3	word4	word5	word6	word7	word8	word9	word10
0	PERSON	(lauren, -4.11)	(Lauren, -2.89)	(abel, -1.79)	(Ladell, -1.79)	(William, -1.39)	(Laureni, -1.39)	(., -0.93)	(have, -0.91)	(to, -0.87)	(sake, -0.86)
1	LOC	(l'arlequin, -1.1)	(allendale, -0.36)	(i, -0.33)	(on, -0.33)	(grim, -0.32)	(goblin, -0.32)	(tragedy, -0.3)	('s, -0.27)	(., -0.25)	(you, -0.25)

ط. هیستوگرام تعداد تکرار هر کلمه منحصر به فرد به ترتیب از فرکانس بالا به پایین



لینک github :

https://github.com/Bayany/NLP_NER

لینک huggingface:

https://huggingface.co/datasets/Bayany/NER