دانشکده مهندسی کامپیوتر

گزارش فاز اول

بیان دیوانی آذر ۹۸۵۲۲۲۲۹

نیم‌سال دوم

سال تحصیلی ۰۲-۱۴۰۱

از وبسایت وبتون چپتر هارا دانلود میکنیم و در هر صفحه OCR را اجرا میکنیم و متن هر چپتر را استخراج میکنیم و در فرمت [chapter_title].txt ذخیره میکنیم.
از آنجایی که در NER نقش واژه ها اهمیت دارد. جمله های تک واژه ای را حذف میکنیم و سپس به کمک ابزار ntlk و تابع sent_tokenize جملات را تشخیص میدهیم.
حال جملات را با کمک این ابزار برچسب گذاری میکنیم.

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

در این ابزار از جملاتی که تگ ندارند، skip میکنیم بنابراین در داده تمیز شده حذف میشوند.
و سپس تمام متن های تمیز شده چپتر ها را با برچسب هایشان به صورت فایل csv ذخیره میکنیم.

أ. تعداد «واحد» داده ، ب. تعداد جملات

Total	location	character
23990	26	121

ج. تعداد کلمات

Total	location	character
106362	26	145

د. تعداد کلمات منحصر به فرد

Total	location	character
12219	15	24

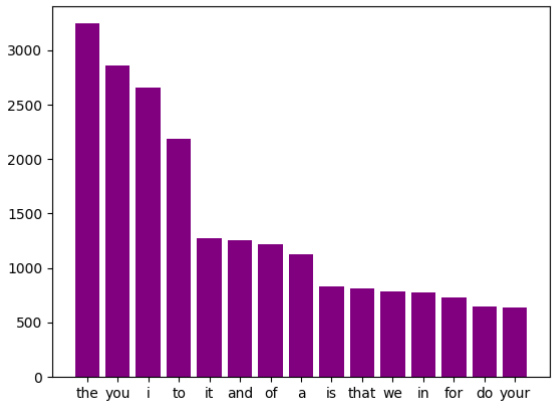
و. ۱۰کلمه پرتکرار غیر مشترک هر برچسب

category	word1	word2	word3	word4	word5	word6	word7	word8	word9	word10
location	nightingale park	allendale park	redcliff's mansion	allendale train station	golden clover	grim goblin	greychapel	tower of ardhalis	mirage opera house	ganbury street
character	lauren	kym	hawkes	tim	sake	sinclair	kieran	davenport	ryan	dakan

ح. ده کلمه برتر بر اساس TF_IDF(W)

category	word1	word2	word3	word4	word5	word6	word7	word8	word9	word10
location	nightingale park	allendale park	redcliff's mansion	allendale train station	golden clover	grim goblin	greychapel	tower of ardhalis	mirage opera house	ganbury street
character	lauren	kym	hawkes	tim	sake	sinclair	kieran	davenport	ryan	dakan

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