پروژه ی سوم درسNLP ترجمه ماشینی

۱) جمع آوری داده:

برای ترجمه رمان « صد سال تنهایی» را به دو زبان انگلیسی و اسپانیایی انتخاب کردم. پی دی اف هر کدام از زبان های مورد نظر را از اینترنت دانلود کردم و تبدیل آن به text به خاطر تایپی بودن متن کتاب کار سختی نبود. فایل کتاب ها را با نام های one hundred years of solitude_ English.txt و Gabriel García Márquez.txt مشاهده می کنید.

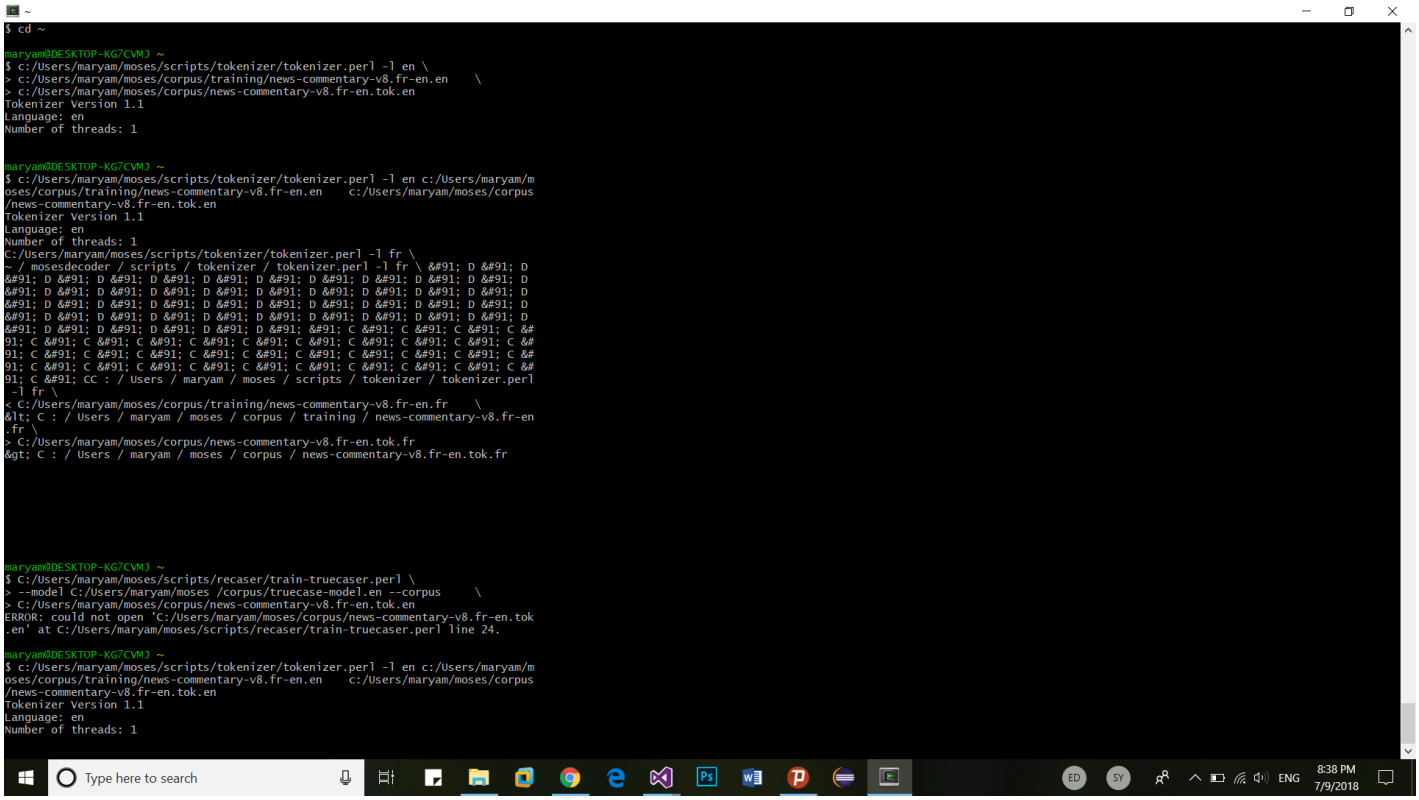
۲) نصب برنامه ی moses:

معمولا برای نصب moses طبق گفته ی سایت بهتر است از سیستم عامل لینوکس استفاده شود. من هم این کار را کردم ولی در نصب boost به مشکل بر خوردم. بنابراین بر روی ویندوز نصب کردم. که از دو لینک زیر برای نصب آن استفاده کردم.

https://jon.dehdari.org/teaching/uds/moses/
Oxkʌelɸhttps://www.youtube.com/watch?v=hh-V

۳) کار با moses و اجرای آن:

http://www.statmt.org/moses/?n=Moses.Baseline مطابق با این لینک مراحل ترجمه ی ماشینی را برای داده ی نمونه ای که در این سایت معرفی شده بود را جلو رفتم. پس از نصب giza++ برای align کردن corpus ها و همچنین دانلود داده نمونه ، متاسفانه در مرحله ی tokenization با مشکل زیر مواجه شدم که process با اینکه مدت زیادی را در حال ران گذاشتم، به پایان نمیرسید و نمیتوانستم ادامه ی مراحل را انجام بدم. بعد از search و پرسیدن از بچه های دیگر هم نتوانستم مشکل را برطرف کنم.


~
$ cd ~
maryam@DESKTOP-KG7CVMJ ~
$ c:/Users/maryam/moses/scripts/tokenizer/tokenizer.perl -l en \
> c:/Users/maryam/moses/corpus/training/news-commentary-v8.fr-en.en \
> c:/Users/maryam/moses/corpus/news-commentary-v8.fr-en.tok.en
Tokenizer Version 1.1
Language: en
Number of threads: 1
maryam@DESKTOP-KG7CVMJ ~
$ c:/Users/maryam/moses/scripts/tokenizer/tokenizer.perl -l en c:/Users/maryam/moses/corpus/training/news-commentary-v8.fr-en.en c:/Users/maryam/moses/corpus/news-commentary-v8.fr-en.tok.en
Tokenizer Version 1.1
Language: en
Number of threads: 1
C:/Users/maryam/moses/scripts/tokenizer/tokenizer.perl -l fr \
~ / mosesdecoder / scripts / tokenizer / tokenizer.perl -l fr \ [D [D
[D [D [D [D [D [D [D [D [D [D
[D [D [D [D [D [D [D [D [D [D
[D [D [D [D [D [D [D [D [D [D
[D [D [D [D [D [D [D [D [D [D
[D [D [D [D [D [[C [C [C &#
91; C [C [C [C [C [C [C [C [C &#
91; C [C [C [C [C [C [C [C [C &#
91; C [C [C [C [C [C [C [C [C &#
91; C [CC : / Users / maryam / moses / scripts / tokenizer / tokenizer.perl
-l fr \
< C:/Users/maryam/moses/corpus/training/news-commentary-v8.fr-en.fr \
< C : / Users / maryam / moses / corpus / training / news-commentary-v8.fr-en
.fr \
> C:/Users/maryam/moses/corpus/news-commentary-v8.fr-en.tok.fr
> C : / Users / maryam / moses / corpus / news-commentary-v8.fr-en.tok.fr
maryam@DESKTOP-KG7CVMJ ~
$ C:/Users/maryam/moses/scripts/recaser/train-truecaser.perl \
> --model C:/Users/maryam/moses /corpus/truecase-model.en --corpus \
> C:/Users/maryam/moses/corpus/news-commentary-v8.fr-en.tok.en
ERROR: could not open 'C:/Users/maryam/moses/corpus/news-commentary-v8.fr-en.tok
.en' at C:/Users/maryam/moses/scripts/recaser/train-truecaser.perl line 24.
maryam@DESKTOP-KG7CVMJ ~
$ c:/Users/maryam/moses/scripts/tokenizer/tokenizer.perl -l en c:/Users/maryam/moses/corpus/training/news-commentary-v8.fr-en.en c:/Users/maryam/moses/corpus/news-commentary-v8.fr-en.tok.en
Tokenizer Version 1.1
Language: en
Number of threads: 1
Type here to search
ENG
8:38 PM
7/9/2018