محمد مهدی یادگار

در این فاز چندین بخش را می خواهیم اجرا کنیم. درباره بخش هایی که تا کنون اجرا شده و در گیت موجود است به طور مختصر در اینجا توضیحاتی داده می شود.

اولین بخش آن ایجاد word2vec برروی دادگان خودمان بود. اینکار را با استفاده از کد تمرین ۲ انجام دادم. برای انجام اینکار یک کد داریم که کل متنی که داریم را به یک فایل تکست تبدیل کند سپس این فایل را به مدل word2vec می دهیم، البته در سه بخش. بخش اول دادگان در تگ حقوقی، بخش دوم دادگان در تگ کیفری، بخش سوم دادگان همه ی متون را شامل میشود. این مدل ها بطور جداگانه سیو و بررسی می شوند.

در دومین بخش کاری که می کنیم این است که توکنایزشن را در ۴ اندازه مختلف انجام دهیم. اندازه ها را از ۱۵۰۰ تا ۶۲۰۰ گذاشته ام. در هر کدام از این اندازه ها متن را ۵ بار به ۲۰ درصد و ۸۰ درصد تقسیم میکنیم و آموزش می دهیم و بهترین حالت را انتخاب می کنیم. در نهایت همانطور که انتظار می رفت با بیشترین تعداد subword ها یعنی ۶۲۰۰ بهترین جواب گرفته شد.

در بخش سوم هم مدل parsing را روی دادگان فارسی باید آموزش میدادم که برای اینکار دیتاست فارسی را از سایت مربوطه دانلود کردم. مدل را برروی آن اجرا کردم و نتیجه UAS=87 شد. که نسبتا قابل قبول است.