

پروژه نهایی درس هوش محاسباتی:

به همراه این صورت پروژه و دیتا، یک مقاله آپلود شده است که روشی برای کلاستر بندی را ارائه میدهد. مقاله را به دقت مطالعه کنید و روش کلاسترینگ بیان شده در مقاله را پیاده سازی کنید. ۱ الی ۲ نمره از بارم پروژه بصورت سوالات کتبی در امتحان از این مقاله خواهد بود. (خود مقاله و توانایی استفاده از روش مطرح شده در آن برای حل مسائل مختلف).

در کلاس دو الگوریتم word2vec و Bag of Words برای پردازش زبان طبیعی (NLP) با جزئیات بررسی شدند. یکی از مشکلات روش Bag of Words این است که هر کلمه را بطور مجزا در نظر گرفته و ارتباط بین کلماتی که معنی یکسانی دارند را در نظر نمیگیرد. روشی ارائه دهید که با استفاده از الگوریتم word2vec بتوانیم این مشکل را برطرف کرده و الگوریتم Bag of Words را بهبود دهیم.

دیتاستی که در اختیار شما قرار داده شده است کامنت های کاربران سایت Reddit برای سه موضوع فیزیک، شیمی و زیست می باشد. روش پیشنهادی خود برای بهبود الگوریتم Bag of Words را پیاده سازی کرده و از آن برای بازنمایی این کامنت ها استفاده کنید. اگر در بدست آوردن بازنمایی ها نیازی به استفاده از کلاسترینگ هست، از کدی که برای قسمت اول پیاده سازی کردید استفاده کنید. (مجاز به استفاده از سایر روش های کلاسترینگ نیستید). حال با استفاده از داده های موجود در فایل train، مدلی را آموزش دهید موضوع یک کامنت را تشخیص دهد. دقت مدل را روی داده های فایل test گزارش کنید.

میتوانید پروژه را در قالب گروه های ۲ نفره انجام دهید. مشابه تمرین های کلاسی، برای پیاده سازی تمرین میتوانید از متلب یا پایتون استفاده کنید. استفاده از کتابخانه ها و منابع آنلاین برای خواندن و پردازش داده و الگوریتم های Bag of Words و word2vec مشکلی ندارد. برای word2vec میتوانید از مدل های از قبل تعلیم دیده استفاده کنید. در صورت مشاهده هر گونه شباهت نامتعارف بین کد دانشجویان و یا مشابهت با کد های موجود در اینترنت با افراد متخلف برخورد خواهد شد. مهلت تحویل پروژه تا پایان روز دوشنبه ۱۳ تیر می باشد. لطفا تا این تاریخ کد و گزارش خود را در سایت VU بارگذاری کنید. موفق باشید!