دانشگاه کردستان
University of Kurdistan
زانکۆی کوردستان

تمرین چهارم جستجو و بازیابی اطلاعات

موضوع: CBOW

استاد

دکتر دانشفر

حل تمرین

پارسا زاهدی

دستورالعمل:

دیتاست: از دیتاست "wiki_00.txt" استفاده کنید.

کدنویسی: باید از زبان پایتون استفاده کنید. فایل تحویلی ترجیحا ipynb باشد. می‌توانید از https://colab.research.google.com/drive/ هم استفاده کنید.

پیش‌پردازش: مانند تمرین‌های قبل انجام شود.

- دیتاست را از فایل بخوانید و آن‌ها را با استفاده از NLTK پیش‌پردازش کنید.
- محتوای فایل را به توکن تبدیل کنید.
- علائم نگارشی و stop words را حذف کنید.

فایل‌های خروجی:

بردار کلمات را با روش CBOW برای همه کلمات را با window size = 2 یا context size و embed_dimension = 10 به‌دست بیاورید.شبکهٔ عصبی یک لایه نهان دارد و خروجی با تابع SoftMax نرمالایز می‌شود. فایل py. یا ipynb. به همراه فایل گزارش (ویس یا pdf) zip کرده و آپلود کنید.

راهنمایی: سرچ کنید CBOW embedding python. از PyTorch یا tensorFlow استفاده کنید.

نمره مثبت: خروجی را به صورت یک نمودار دو بعدی نشان دهید. مانند شکل زیر: