دانشگاه اصفهان
دانشکده مهندسی کامپیوتر

تمرین اول: فاز دوم پروژه سخنرانی‌های TED	**استاد:** دکتر احمد براآنی
درس: بازیابی پیشرفته اطلاعات	**دستیار:** الهام اسماعیلی
نام و نام خانوادگی: سید عمید اسدالهی مجد	**شماره دانشجویی:** ۴۰۰۳۶۱۴۰۰۴

آدرس گیت: https://github.com/amidmajd/ted-talk-classification

تبدیل متن به بردار با روش Skip-Gram

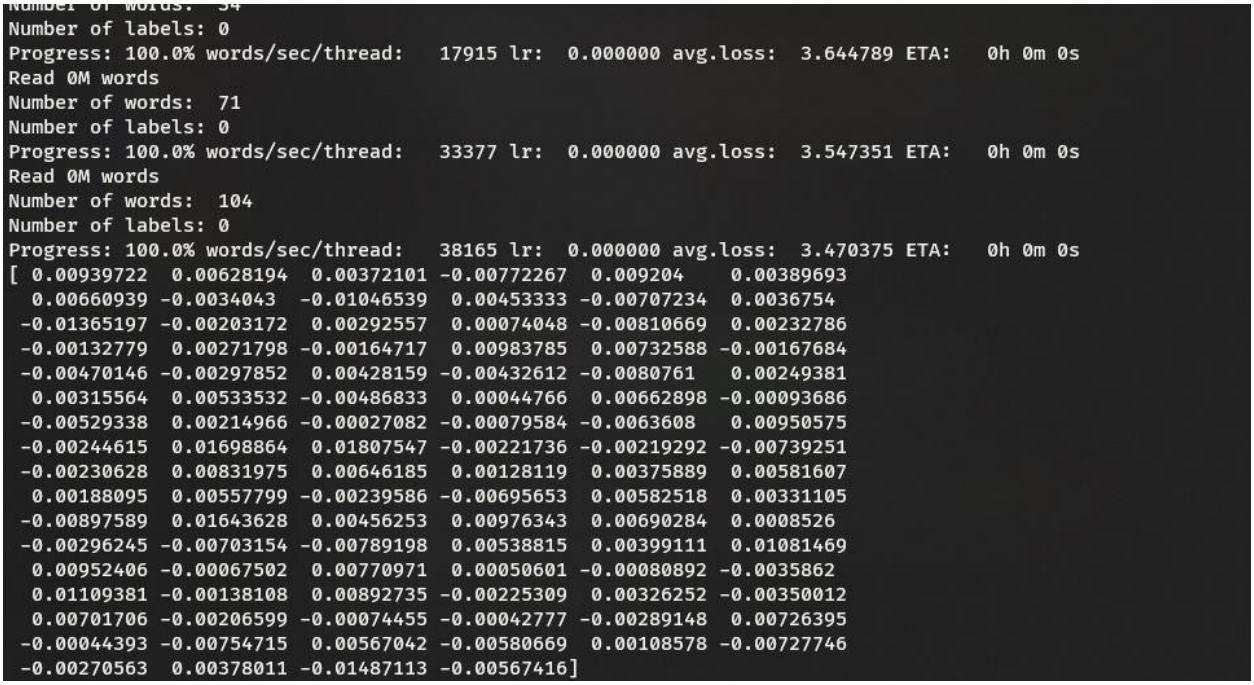
یکی از الگوریتم‌های تبدیل متن به بردار غیرنظارتی، الگوریم Skip-Gram است. این الگوریتم بدون نیاز به برچسب و ناظر یاد می‌گیرد تا کلمه هدف را با استفاد از کلمات همسایه‌ی کلمه هدف در جمله محاسبه کند.

در این روش با استفاده از یک پنجره کشویی که برروی جمله و کلمات حرکت می‌کند سعی می‌کند تا کلمه بعدی را حدس بزند. در این روش ابتدا بردار کلمات نیز تشکیل می‌شود که در این تمرین مورد استفاده قرار می‌گیرد.

توضیح کد

در این کد ابتدا از ایندکس الستیک‌سرچی که در فاز اول پروژه ساخته شده‌بود خوانده می‌شود و بخش زیرنویس استخراج می‌شود. سپس به‌دلیل نیاز به آدرس فایل به‌عنوان ورودی کتابخوانه‌ی FastText، نیاز است تا تک‌تک زیرنویس‌ها به‌صورت فایل ذخیره شوند. پس از ذخیره آن‌ها در پوشه temp ساخته شده در مسیر کد، نام هر فایل را به‌عنوان ورودی تابع train_unsupervised از کتاب‌خوانه‌ی FastText استفاده می‌کنیم. پس از ساخت مدل برای هر زیرنویس، آن‌ها را در یک آرایه ذخیره می‌نمایی. در انتها یک نمونه بردار از کلمه the از یک زیرنویس نمایش داده می‌شود.

تصاویری از نتایج اجرای برنامه



بردار کلمه **the** از یک زیرنویس