



تمرین اول: فاز دوم پروژه سخنرانی های TED

درس: بازیابی پیشرفته اطلاعات

نام و نام خانوادگی: سید عمید اسدالهی مجد

استاد: دکتر احمد برآنی

دستیار: الهام اسماعیلی

شماره دانشجویی: ۴۰۰۳۶۱۴۰۰۴

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

تبدیل متن به بردار با روش Skip-Gram

یکی از الگوریتم های تبدیل متن به بردار غیرنظارتی، الگوریتم Skip-Gram است. این الگوریتم بدون نیاز به برچسب و ناظر یاد می گیرد تا کلمه هدف را با استفاده از کلمات همسایه ی کلمه هدف در جمله محاسبه کند.

در این روش با استفاده از یک پنجره کشویی که بر روی جمله و کلمات حرکت می کند سعی می کند تا کلمه بعدی را حدس بزند. در این روش ابتدا بردار کلمات نیز تشکیل می شود که در این تمرین مورد استفاده قرار می گیرد.

توضیح کد

در این کد ابتدا از ایندکس الستیک سرچی که در فاز اول پروژه ساخته شده بود خوانده می شود و بخش زیرنویس استخراج می شود. سپس به دلیل نیاز به آدرس فایل به عنوان ورودی کتابخانه ی FastText، نیاز است تا تک تک زیرنویس ها به صورت فایل ذخیره شوند. پس از ذخیره آن ها در پوشه temp ساخته شده در مسیر کد، نام هر فایل را به عنوان ورودی تابع train_unsupervised از کتابخانه ی FastText استفاده می کنیم. پس از ساخت مدل برای هر زیرنویس، آن ها را در یک آرایه ذخیره می نماییم. در انتها یک نمونه بردار از کلمه the از یک زیرنویس نمایش داده می شود.

تصاویری از نتایج اجرای برنامه

```
Number of words: 34
Number of labels: 0
Progress: 100.0% words/sec/thread: 17915 lr: 0.000000 avg.loss: 3.644789 ETA: 0h 0m 0s
Read 0M words
Number of words: 71
Number of labels: 0
Progress: 100.0% words/sec/thread: 33377 lr: 0.000000 avg.loss: 3.547351 ETA: 0h 0m 0s
Read 0M words
Number of words: 104
Number of labels: 0
Progress: 100.0% words/sec/thread: 38165 lr: 0.000000 avg.loss: 3.470375 ETA: 0h 0m 0s
[ 0.00939722 0.00628194 0.00372101 -0.00772267 0.009204 0.00389693
 0.00660939 -0.0034043 -0.01046539 0.00453333 -0.00707234 0.0036754
-0.01365197 -0.00203172 0.00292557 0.00074048 -0.00810669 0.00232786
-0.00132779 0.00271798 -0.00164717 0.00983785 0.00732588 -0.00167684
-0.00470146 -0.00297852 0.00428159 -0.00432612 -0.0080761 0.00249381
 0.00315564 0.00533532 -0.00486833 0.00044766 0.00662898 -0.00093686
-0.00529338 0.00214966 -0.00027082 -0.00079584 -0.0063608 0.00950575
-0.00244615 0.01698864 0.01807547 -0.00221736 -0.00219292 -0.00739251
-0.00230628 0.00831975 0.00646185 0.00128119 0.00375889 0.00581607
 0.00188095 0.00557799 -0.00239586 -0.00695653 0.00582518 0.00331105
-0.00897589 0.01643628 0.00456253 0.00976343 0.00690284 0.0008526
-0.00296245 -0.00703154 -0.00789198 0.00538815 0.00399111 0.01081469
 0.00952406 -0.00067502 0.00770971 0.00050601 -0.00080892 -0.0035862
 0.01109381 -0.00138108 0.00892735 -0.00225309 0.00326252 -0.00350012
 0.00701706 -0.00206599 -0.00074455 -0.00042777 -0.00289148 0.00726395
-0.00044393 -0.00754715 0.00567042 -0.00580669 0.00108578 -0.00727746
-0.00270563 0.00378011 -0.01487113 -0.00567416]
```

بردار کلمه the از یک زیرنویس