Work Shop (S1)

: One-hot

مزایا:

- پیاده سازی آسانی دارد.

معایب:

- دایمنشن هر کلمه بسیار زیاد است (برابر با تعداد کلمات) که فقط یکی از آنها برابر یک و بقیه برابر صفر هستند. یعنی فضای زیادی برای ذخیره اطلاعات هدر می شود.
- دو کلمه کنار هم encoding مشابهی دارند ولی لزوما معنای مشابهی ندارند. پس one-hot معنای کلمات را ذخیره نمی کند.

: Tf_idf

مزایا:

- پیاده سازی آسان
- به راحتی می توان میزان تشابه دو جمله را محاسبه کرد.

معایب:

- این مدل از embedding از روش bag of words استفاده می کند، یعنی جایگاه کلمه در یک جمله برایش مهم نیست و فقط وجود کلمه مهم است.
- معنای کلمات را نمی تواند یاد بگیرد.

: Word2vec

مزایا:

- می تواند معنی کلمات را یاد بگیرد.

معایب:

- برای هر کلمه تنها یک embedding دارد.
- پیاده سازی دشواری دارد.

: Bert

مزایا:

- علاوه بر آن که می تواند معنای کلمات را یاد بگیرد، embedding هر کلمه نسبت به متنش تغییر می کند. یعنی هر کلمه بیش از یک embedding دارد.
- با دیتای زیادی آموزش دیده.
- آموزش آن bidirectional بوده.

معایب:

- پیاده سازی دشواری دارد.