RNN

فصل پنجم

شرکت هوش مصنوعی رسا

مدرس : فرید هاشمی نژاد

Word2Vec

جلسه سی ام

شرکت هوش مصنوعی رسا

مدرس : فرید هاشمی نژاد

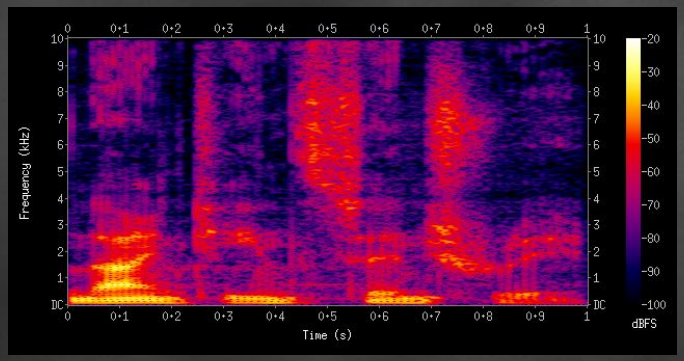
- Word2Vec
- Gensim
- NLP
- From String to vectors
- [‘Hi’ , ’How’ , ’are’ , ’you’]
- NLP کلاسیک : اعدادی که بیانگر رابطه تکرار کلمات با متن است

- Count Based : تکرار کلمات در متن
- Predictive Based : همسایگی کلمات بر پایه یک فضای برداری تخمین زده می شود
- یکی از شبکه های عصبی معروف NLP
- Word2Vec : Mikolov & el
- جایگذاری کلمات را یاد می گیرد: با
- مدل کردن هر کلمه به عنوان برداری n بعدی

- توصیف داده ها
- Representation of Data

تصویر (dense)

صدا (dense)



متن (sparse)

0	0	0	0.7	0.2	...

- Word2Vec
- شروع آموزش با جایگذاری رندم اتفاق می افتد
- و با backpropagation مدل به سمت بدست آوردن مقدار دقیق پیش میرود
- ابعاد بیشتر، نشان دهنده زمان آموزش بیشتر و البته اطلاعات بیشتر نسبت به هر کلمه است

- کلمات مشابه با دنباله های نزدیکتر پیدا می شوند
- فعل ها جنسیت و . . .

- Prediction Target
- Skip-Gram model
- سک استخوان را می جود
- مناسب برای دیتاست های بزرگتر
- CBOW(Continuous Bag of Words)
- سک استخوان را می جود
- برای دیتاست های کوچکتر مناسبه

- سگ wt=? را می جود
- استخوان سیب ماشین کتاب
- کلمه هدف کلمات نویز

- طبقه بند های نویز
- لایه مخفی
- لایه projection

- Noise-contrastive Training
- **کلمه هدف بر اساس ماکزیمم یابی پیدا می شود .**

- $J_{NEG} = \log Q_\theta(D=1|w_t,h) + k_{n\sim pnoise} E[\log Q_\theta(D=0|w_n,h)]$
- $\log Q_\theta(D=1|w_t,h)$ **رگرسیون لجستیکی باینری**
- H **متن و** θ **پارامتر دیتاست** D ،W_t **کلمه درست یا هدف**
- Wn **کلمه ای که با توزیع نویز ساخته شده**
- **هدف : احتمال بالا برای کلمه درست و احتمال پایین برای کلمات نویزی**
- **سپس کاهش ابعاد به ازای هر کلمه از بردار ۱۰۰ تایی برای مثال به ۲ تایی**