بررسی مدل‌های جادهی کلمات یا word embedding

پایان‌نامه‌ی کارشناسی ارشد رشته‌ی علوم کامپیوتر

محمد فاضلی
استاد راهنما: دکتر سید رضا مقدسی

دانشگاه صنعتی شریف

۲۸ نوامبر ۲۰۱۹

به نام خدا

محتوی ارایه

معرفی تحقیق

- بررسی و ارائه‌ی روش هایی برای نمایش متن
- بررسی مدل‌هایی زبانی که رابطه‌ی نزدیکی با راه حل ما دارند

معرفی تحقیق

- بررسی و ارائه‌ی روش هایی برای نمایش متن
- بررسی مسئله‌ی تخمین احتمال وقوع یک متن

- بررسی مدل‌هایی زبانی که رابطه‌ی نزدیکی با راه حل ما دارند

معرفی تحقیق

- بررسی و ارائه‌ی روش هایی برای نمایش متن
- بررسی مسئله‌ی تخمین احتمال وقوع یک متن
- چون رابطه‌ی بسیار نزدیک با روش های جدیدنمایش متن دارد

- بررسی مدل‌هایی زبانی که رابطه‌ی نزدیکی با راه حل ما دارند

معرفی تحقیق

- بررسی و ارائه‌ی روش هایی برای نمایش متن
- بررسی مسئله‌ی تخمین احتمال وقوع یک متن
- چون رابطه‌ی بسیار نزدیک با روش های جدیدنمایش متن دارد
- حل مشکلات و افزایش کارایی مدل‌های کلاسیک
- بررسی مدل‌هایی زبانی که رابطه‌ی نزدیکی با راه حل ما دارند

معرفی تحقیق

- بررسی و ارائه‌ی روش هایی برای نمایش متن
- بررسی مسئله‌ی تخمین احتمال وقوع یک متن
- چون رابطه‌ی بسیار نزدیک با روش های جدیدنمایش متن دارد
- حل مشکلات و افزایش کارایی مدل‌های کلاسیک
- بررسی مدل‌هایی زبانی که رابطه‌ی نزدیکی با راه حل ما دارند

معرفی تحقیق

ادامه

- معرفی مدل‌های زبانی موفق اولیه با شبکه‌ی عصبی

معرفی تحقیق

ادامه

- معرفی مدل‌های زبانی موفق اولیه با شبکه‌ی عصبی
- معرفی جادهی کلمات (word embedding) به خصوص مدل word2vec

معرفی تحقیق

ادامه

- معرفی مدل‌های زبانی موفق اولیه با شبکه‌ی عصبی
- معرفی جادهی کلمات (word embedding) به خصوص مدل word2vec
- بررسی روش‌های تعمیم جادهی کلمات به سندهای متنی

معرفی تحقیق

ادامه

- بررسی کارایی مدل‌های معرفی شده روی زبان فارسی

پردازش متن

- به خاطر رشد اینترنت درس رسی به منابع دیجیتال متنی زیاد است.

- این مدل‌ها برای داده‌های غیر متن نیز قابل استفاده است

پردازش متن

- به خاطر رشد اینترنت درس رسی به منابع دیجیتال متنی زیاد است.
- محبوبیت محصولاتی مانند گوگل، نشان می‌دهد پیدا کردن و منظم کردن این داده‌ها جای کار بسیاری دارد.

- این مدل‌ها برای داده‌های غیر متن نیز قابل استفاده است

پردازش متن

- به خاطر رشد اینترنت درس رسی به منابع دیجیتال متنی زیاد است.
- محبوبیت محصولاتی مانند گوگل، نشان می‌دهد پیدا کردن و منظم کردن این داده‌ها جای کار بسیاری دارد.
- در هر ثانیه۶هزار توییت تولید می‌شود،۵۰۰میلیون در روز،سالی ۲۰۰ میلیارد
- این مدل‌ها برای داده‌های غیر متن نیز قابل استفاده است

پردازش متن

- به خاطر رشد اینترنت درس رسی به منابع دیجیتال متنی زیاد است.
- محبوبیت محصولاتی مانند گوگل، نشان می‌دهد پیدا کردن و منظم کردن این داده‌ها جای کار بسیاری دارد.
- در هر ثانیه۶هزار توییت تولید می‌شود،۵۰۰میلیون در روز،سالی ۲۰۰ میلیارد
- این مدل‌ها برای داده‌های غیر متن نیز قابل استفاده است

پردازش متن

ادامه

- متن به شکلی که رشته‌ی کاراکترها برای بررسی مدل‌های معمول یادگیری ماشین مناسب نیست
- هدف ما این است که زبان انسان را به زبان برداری بیان کنیم

پردازش متن

ادامه

- متن به شکلی که رشته‌ی کاراکترها برای بررسی مدل‌های معمول یادگیری ماشین مناسب نیست
- این مدل‌ها به زبان بردار سخن می‌گویند
- هدف ما این است که زبان انسان را به زبان برداری بیان کنیم

پردازش متن

ادامه

- متن به شکلی که رشته‌ی کاراکترها برای بررسی مدل‌های معمول یادگیری ماشین مناسب نیست
- این مدل‌ها به زبان بردار سخن می‌گویند
- هدف ما این است که زبان انسان را به زبان برداری بیان کنیم

نمایش معنی کلمه

- برای این کار از نمایش معنی یک کلمه شروع می‌کنیم.

معنی کلمه به صورت Formal

شکل مرسوم در منطق

All cats chase mice:
$\forall x \forall y((cat'(x) \wedge mouse'(y)) \rightarrow chase'(x, y))$
Each mouse is connected to a computer:
$\forall x(mouse'(x) \rightarrow \exists y(comp'(y) \wedge connect'(x, y)))$

معنی کلمه به صورت محدودیت‌هایی که رابطه‌های تعریف شده روی آن می‌گذارد تعریف می‌شود.

فرض توزیعی

فرضیه

کلمات که متن اطرافشان شبیه به هم است مستعد شباهت معنایی هستند.

فرض توزیعی

فرضیه

کلمات که متن اطرافشان شبیه به هم است مستعد شباهت معنایی هستند.

مثال:

- «دکتر» و «پزشک» در یک محیط در متن ظاهر می‌شوند.
- «پزشک» و «مطب» در یک متن نزدیک هم ظاهر می‌شوند.

فرض توزیعی

- این فرض از نظر فلسفی Wittgenstein ریشه گرفته.

فرض توزیعی

- این فرض از نظر فلسفی Wittgenstein ریشه گرفته.
- در زبان شناسی به شکل مطرح شده توسط زبان شناسان در دهه‌ی ۱۹۶۰ وارد شده

فرض توزیعی

- این فرض از نظر فلسفی Wittgenstein ریشه گرفته.
- در زبان شناسی به شکل مطرح شده توسط زبان شناسان در دهه‌ی ۱۹۶۰ وارد شده
- «تفاوت معنی دو کلمه با تفاوت بین محیط آن‌ها رابطه‌ی تقریبی دارد» Harris

فرض توزیعی

- این فرض از نظر فلسفی Wittgenstein ریشه گرفته.
- در زبان شناسی به شکل مطرح شده توسط زبان شناسان در دهه‌ی ۱۹۶۰ وارد شده
- «تفاوت معنی دو کلمه با تفاوت بین محیط آن‌ها رابطه‌ی تقریبی دارد» Harris
-

نمایش کلمه

- نمایش کلمه به صورت بردار اولین بار در [۵۷CT] مطرح شد.

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24
life	6.68	5.59	5.89

مدل فضای برداری

- مدل کلمه را به شکل یک بردار در فضای $R^{|D|}$ نمایش داده می‌شود که در ازای هر سندی که کلمه در آن ظاهر شده، مقدار بعد آن سند برابر عدد متناسب با تکرار آن کلمه در آن بعد است.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

کاهش بعد مدل فضای برداری

- جادهی کلمه‌های در فضای با بعد پایین تر مشکل تنکی و بعد بالا را حل می‌کند

کاهش بعد مدل فضای برداری

- جادهی کلمه‌های در فضای با بعد پایین تر مشکل تنکی و بعد بالا را حل می‌کند
- سند مثال مناسبی برای نمایش معنی کلمه نیست

کاهش بعد مدل فضای برداری

- جادهی کلمه‌های در فضای با بعد پایین تر مشکل تنکی و بعد بالا را حل می‌کند
- سند مثال مناسبی برای نمایش معنی کلمه نیست
- چون برای تفاوت کافی برای همه‌ی کلمه‌ها به سندهای زیادی نیاز داریم.

کاهش بعد مدل فضای برداری

- جادهی کلمه‌های در فضای با بعد پایین تر مشکل تنکی و بعد بالا را حل می‌کند
- سند مثال مناسبی برای نمایش معنی کلمه نیست
- چون برای تفاوت کافی برای همه‌ی کلمه‌ها به سندهای زیادی نیاز داریم.
- هدف ما همین جادهی در فضایی با بعد پایین تر است

کاهش بعد مدل فضای برداری

- جادهی کلمه‌های در فضای با بعد پایین تر مشکل تنکی و بعد بالا را حل می‌کند
- سند مثال مناسبی برای نمایش معنی کلمه نیست
- چون برای تفاوت کافی برای همه‌ی کلمه‌ها به سندهای زیادی نیاز داریم.
- هدف ما همین جادهی در فضایی با بعد پایین تر است
- به صورتی که ویژگی‌های معنایی کلمات بهتر مشخص شوند

مدل زبان

- مدلی که رابطه‌ی نزدیکی با روش‌های جادهی کلمه در بعد پایین تر دارد مدل زبان است
- کاربردهای خودش را نیز مستقلا دارد

مدل زبان

- مدلی که رابطه‌ی نزدیکی با روش‌های جادهی کلمه در بعد پایین تر دارد مدل زبان است
- مدل زبان نشان دهنده‌ی احتمال وقوع یک دنباله از کلمات در زبان است
- کاربردهای خودش را نیز مستقلا دارد

مدل زبان

- مدلی که رابطه‌ی نزدیکی با روش‌های جادهی کلمه در بعد پایین تر دارد مدل زبان است
- مدل زبان نشان دهنده‌ی احتمال وقوع یک دنباله از کلمات در زبان است
- کاربردهای خودش را نیز مستقلا دارد

مدل زبان

- مدلی که رابطه‌ی نزدیکی با روش‌های جادهی کلمه در بعد پایین تر دارد مدل زبان است
- مدل زبان نشان دهنده‌ی احتمال وقوع یک دنباله از کلمات در زبان است
- کاربردهای خودش را نیز مستقلا دارد

مدل زبان

- مدلی که رابطه‌ی نزدیکی با روش‌های جادهی کلمه در بعد پایین تر دارد مدل زبان است
- مدل زبان نشان دهنده‌ی احتمال وقوع یک دنباله از کلمات در زبان است
- کاربردهای خودش را نیز مستقلا دارد

تعریف n-gram:
به n کلمه‌ی پشت هم n-gram می‌گویند.
مثال:
«علی از محل کار به خانه آمد»
unigram: { «علی»، «از»، «محل» و...}
bigram: { «علی از»، «از محل»، «محل کار» و... }
trigram: { «علی از محل»، «از محل کار» و ...}

مدل زبان

احتمال وقوع یک دنباله از پیش آمد های تصادفی

$$P(x_1, x_2, \ldots, x_n) = P(x_1) \times P(x_2|x_1) \times P(x_3|x_1, x_2) \times \ldots \times P(x_n|x_1, x_2, \ldots, x_{n-1})$$

برای ساده سازی به شمارش bigram ها داریم :

مدل زبان

احتمال وقوع یک دنباله از پیش آمد های تصادفی

$$P(x_1, x_2, \ldots, x_n) = P(x_1) \times P(x_2|x_1) \times P(x_3|x_1, x_2) \times \ldots \times P(x_n|x_1, x_2, \ldots, x_{n-1})$$

برای ساده سازی به شمارش bigram ها داریم :

$$P(x_1|x_2, x_3, \ldots, x_n) \approx P(x_1|x_2)$$

مدل زبان

مثال

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

شکل: جدول نشان دهنده‌ی بعضی از تعداد تکرار bigram ها، داخل پیکره‌ای از ۹۲۲۲ جمله در مورد رستوران‌های منطقه‌ی برکلی کالیفرنیا است [Jur+94]

مدل زبان

مثال

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

شکل: تعداد unigram ها در پیکره. [Jur+94]

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

شکل: با تقسیم تعداد bigram به تخمین درست نمایی بیشینه‌ی احتمال مدل می‌رسیم. [Jur+94]

مدل زبان

مثال

$$P(i|<s>) = ۰/۲۵$$
$$P(english|want) = ۰/۰۰۱۱$$
$$P(food|english) = ۰/۵$$
$$P(</s>|food) = ۰/۶۸$$

$$P(<s> \ i \ want \ english \ food \ </s>)$$
$$= P(i|<s>)P(want|i)P(english|want)$$
$$P(food|english)P(</s>|food)$$
$$= ۰/۲۵ \times ۰/۳۳ \times \times ۰/۰۰۱۱ \times ۰/۵ \times ۰/۶۸$$
$$= ۰/۰۰۰۰۳۱$$

اندازه‌گیری کارایی مدل زبان

سرگشتگی

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \dots w_{i-1})}}$$

سرگشتگی

شهود

اگر داده‌های رشته‌های ده دهی باشند، که رقم به صورت تصادفی یکنواخت از بین ۰ تا ۹ انتخاب شده داریم:

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}}$$

$$= \frac{1}{10}^{-1} = 10$$

تست روی داده‌گان واقعی با مدل n-gram

سرگشتگی

شهود

اگر داده‌های رشته‌های ده دهی باشند، که رقم به صورت تصادفی یکنواخت از بین ۰ تا ۹ انتخاب شده داریم:

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}}$$

$$= \frac{1}{10}^{-1} = 10$$

تست روی داده‌گان واقعی با مدل n-gram

	Unigram	Bigram	Trigram
Perplexity	962	170	109

شکل: سرگشتگی مدل unigram ، bigram و trigram آموزش داده شده روی داده‌های Wall Street Journal

مثال کارایی مدل

شهود

اگر داده‌های رشته‌های ده دهی باشند، که رقم به صورت تصادفی یکنواخت از بین ۰ تا ۹ انتخاب شده داریم:

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}}$$

$$= \frac{1}{10}^{-1} = 10$$

تست روی داده‌گان واقعی با مدل n-gram

مثال کارایی مدل

شهود

اگر داده‌های رشته‌های ده دهی باشند، که رقم به صورت تصادفی یکنواخت از بین ۰ تا ۹ انتخاب شده داریم:

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}}$$

$$= \frac{1}{10}^{-1} = 10$$

تست روی داده‌گان واقعی با مدل n-gram

	Unigram	Bigram	Trigram
Perplexity	962	170	109

شکل: سرگشتگی مدل unigram ، bigram و trigram آموزش داده شده روی داده‌های Wall Street Journal

مثال کارایی مدل

مصورسازی

1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
3 gram	They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

شکل: مصورسازی مدل‌های زبان با داده‌های [MJ, Language Wall Street Journal Modeling section]

مشکل مدل

- اگر کل کلمه‌ها $V = ۲۹۰۵۱$ باشد(کل متن‌های نوشته شده توسط شکسپیر).

همان تعداد است.

مشکل مدل

- اگر کل کلمه‌ها $V = ۲۹۰۵۱$ باشد(کل متن‌های نوشته شده توسط شکسپیر).
- کل bigram ها $V \approx ۸۴۴۰۰۰۰۰۰$ می‌شود.

همان تعداد است.

مشکل مدل

- اگر کل کلمه‌ها $V = ۲۹۰۵۱$ باشد(کل متن‌های نوشته شده توسط شکسپیر).
- کل bigram ها $V \approx ۸۴۴۰۰۰۰۰۰$ می‌شود.
- کل طول داده‌گان شکسپیر $N = ۸۸۴۶۵۷$ است.

همان تعداد است.

مشکل مدل

- اگر کل کلمه‌ها $V = ۲۹۰۵۱$ باشد(کل متن‌های نوشته شده توسط شکسپیر).
- کل bigram ها $V \approx ۸۴۴۰۰۰۰۰۰$ می‌شود.
- کل طول داده‌گان شکسپیر $N = ۸۸۴۶۵۷$ است.
- تعداد 4-gram ها $V^۴ \approx ۷ \times ۱۰^{۱۷}$ است در حالی که تعداد کل 4-gram همان تعداد است.

مشکل مدل

- اگر حتی یکی از n-gram ها موجود نباشد احتمال صفر می‌شود که سرگشتگی را غیر قابل محاسبه می‌کند.

کم تکرار تر اصافه کرد.

مشکل مدل

- اگر حتی یکی از n-gram ها موجود نباشد احتمال صفر می‌شود که سرگشتگی را غیر قابل محاسبه می‌کند.
- می‌توان مشکل را با کمتر کردن، احتمال از پر تکرار و دادن به صفرها و یا کم تکرارها حل کرد.

کم تکرار تر اصافه کرد.

مشکل مدل

- اگر حتی یکی از n-gram ها موجود نباشد احتمال صفر می‌شود که سرگشتگی را غیر قابل محاسبه می‌کند.
- می‌توان مشکل را با کمتر کردن، احتمال از پر تکرار و دادن به صفرها و یا کم تکرارها حل کرد.
- می‌توان، از (n-1)-gram ، (n-2)-gram و ... را interpolate کرد.

کم تکرار تر اصافه کرد.

مشکل مدل

- اگر حتی یکی از n-gram ها موجود نباشد احتمال صفر می‌شود که سرگشتگی را غیر قابل محاسبه می‌کند.
- می‌توان مشکل را با کمتر کردن، احتمال از پر تکرار و دادن به صفرها و یا کم تکرارها حل کرد.
- می‌توان، از (n-1)-gram ، (n-2)-gram و ... را interpolate کرد.
-

$$P_{interp}(w_3|w_1,w_2) = \lambda_1 P(w_3|w_1,w_2) + \lambda_2 P(w_3|w_2) + \lambda_3 P(w_3)$$
$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

کم تکرار تر اصافه کرد.

مشکل مدل

- اگر حتی یکی از n-gram ها موجود نباشد احتمال صفر می‌شود که سرگشتگی را غیر قابل محاسبه می‌کند.
- می‌توان مشکل را با کمتر کردن، احتمال از پر تکرار و دادن به صفرها و یا کم تکرارها حل کرد.
- می‌توان، از (n-1)-gram ، (n-2)-gram و ... را interpolate کرد.
-

$$P_{interp}(w_3|w_1,w_2) = \lambda_1 P(w_3|w_1,w_2) + \lambda_2 P(w_3|w_2) + \lambda_3 P(w_3)$$
$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

- می‌توان λ ها را بر اساس کلمه‌ها یادگرفت

کم تکرار تر اصافه کرد.

مشکل مدل

- اگر حتی یکی از n-gram ها موجود نباشد احتمال صفر می‌شود که سرگشتگی را غیر قابل محاسبه می‌کند.
- می‌توان مشکل را با کمتر کردن، احتمال از پر تکرار و دادن به صفرها و یا کم تکرارها حل کرد.
- می‌توان، از (n-1)-gram ، (n-2)-gram و ... را interpolate کرد.
-

$$P_{interp}(w_3|w_1,w_2) = \lambda_1 P(w_3|w_1,w_2) + \lambda_2 P(w_3|w_2) + \lambda_3 P(w_3)$$
$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

- می‌توان λ ها را بر اساس کلمه‌ها یادگرفت
- می‌توان کمتر کردن را از کلمه‌های با احتمال بالاتر انجام داد و از کلمه‌های کم تکرار تر اصافه کرد.

مدل زبان شبکه‌ی عصبی

- مدل‌های قبلی، کافی نبودند.
- Bengio **andothers** این کار را انجام دادند.

مدل زبان شبکه‌ی عصبی

- مدل‌های قبلی، کافی نبودند.
- روش های smoothing را می‌شود یاد گرفت.

- **andothers** Bengio این کار را انجام دادند.

مدل زبان شبکه‌ی عصبی

- مدل‌های قبلی، کافی نبودند.
- روش های smoothing را می‌شود یاد گرفت.
- از داده‌ها این تغییر مقدارها را یادگرفت.
- Bengio **andothers** این کار را انجام دادند.

مدل زبان شبکه‌ی عصبی

- مدل‌های قبلی، کافی نبودند.
- روش های smoothing را می‌شود یاد گرفت.
- از داده‌ها این تغییر مقدارها را یادگرفت.
- Bengio **andothers** این کار را انجام دادند.

مدل زبان شبکه‌ی عصبی

- رابطه‌ی زیر را می‌توان مستقیم با شبکه‌ی عصبی مدل کرد:

$$P(w_k|w_{k-۱}w_{k-۲}\dots,w_{k-(n-۱)})$$

مدل را ارائه داد.

مدل زبان شبکه‌ی عصبی

- رابطه‌ی زیر را می‌توان مستقیم با شبکه‌ی عصبی مدل کرد:

$$P(w_k | w_{k-1} w_{k-2} \dots, w_{k-(n-1)})$$

- تعداد تلاش ناموفق برای این کار وجود داشت.

مدل را ارائه داد.

مدل زبان شبکه‌ی عصبی

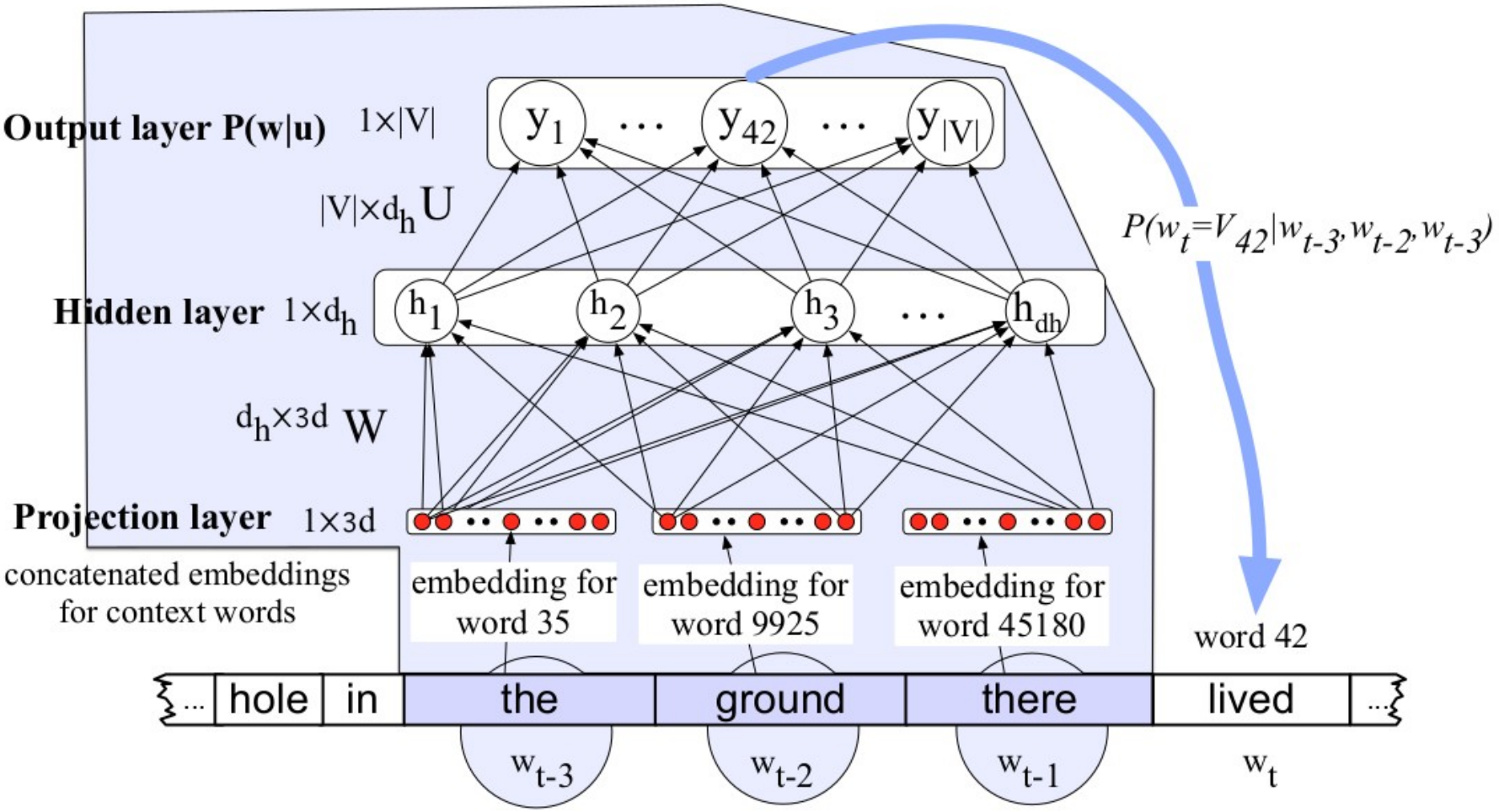
- رابطه‌ی زیر را می‌توان مستقیم با شبکه‌ی عصبی مدل کرد:

$$P(w_k | w_{k-1} w_{k-2} \dots, w_{k-(n-1)})$$

- تعداد تلاش ناموفق برای این کار وجود داشت.
- اگر یک جادهی برای کلمه داشته باشیم می‌شود به شکلی که نشان می‌دهیم مدل را ارائه داد.

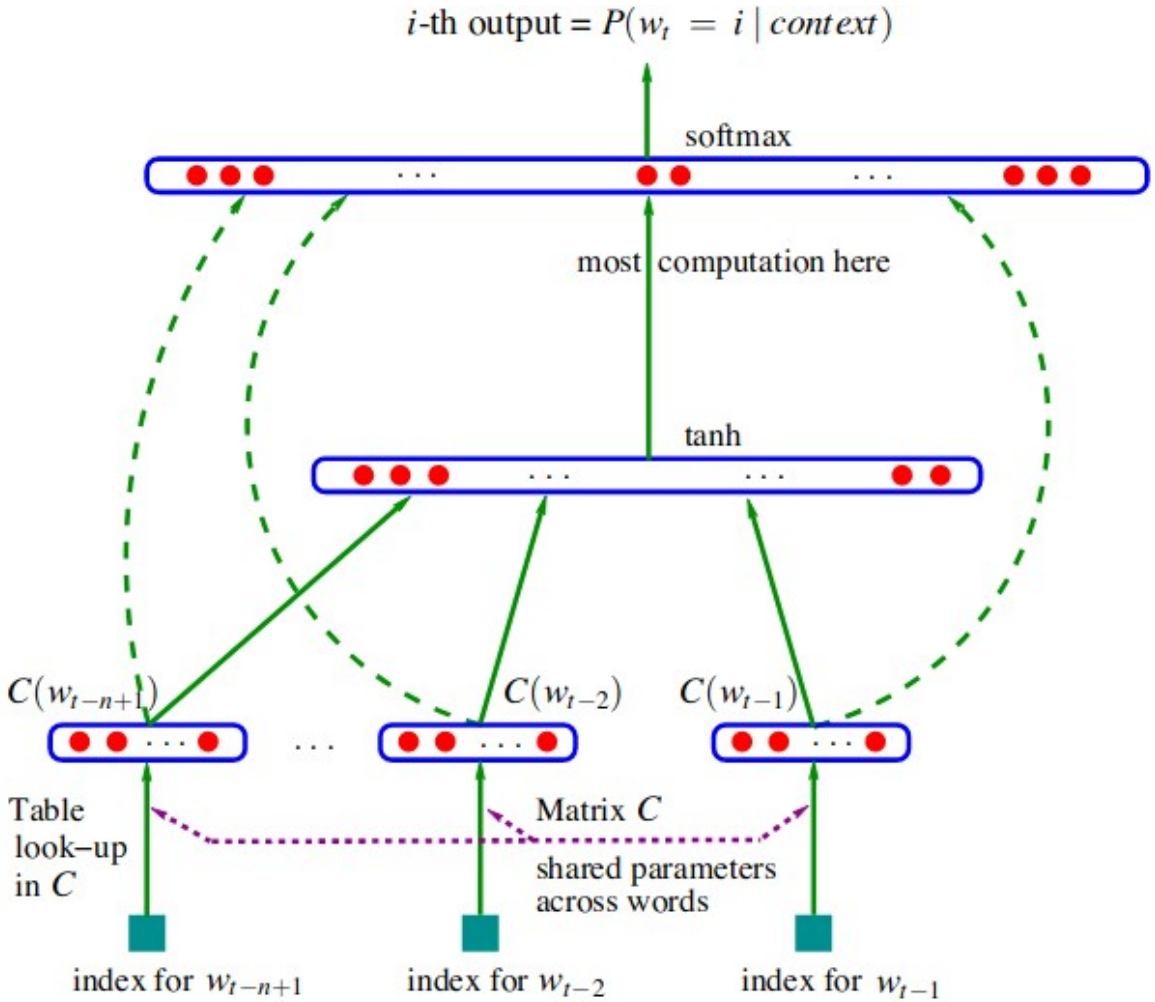
مدل زبان شبکه‌ی عصبی

مدل از روی یک جادهی کلمه



مدل زبان شبکه‌ی عصبی

مدل از روی داده‌ی خام



شکل: (عکس از [Ben+03])

مدل زبان شبکه‌ی عصبی

شهود مدل

- مدل مقدار زیر را تخمین بزند

$$P(eating|the, cat, is)$$

تخمین بزند.

مدل زبان شبکه‌ی عصبی

شهود مدل

- مدل مقدار زیر را تخمین بزند

$$P(eating|the, cat, is)$$

- فرض کنید، مدل بخش زیر را ندیده:

the cat is eating

تخمین بزند.

مدل زبان شبکه‌ی عصبی

شهود مدل

- مدل مقدار زیر را تخمین بزند

$$P(eating|the, cat, is)$$

- فرض کنید، مدل بخش زیر را ندیده:

the cat is eating

- اما مقدار زیر را دیده :

the dog is eating

تخمین بزند.

مدل زبان شبکه‌ی عصبی

شهود مدل

- مدل مقدار زیر را تخمین بزند

$$P(eating|the, cat, is)$$

- فرض کنید، مدل بخش زیر را ندیده:

the cat is eating

- اما مقدار زیر را دیده :

the dog is eating

- و از دیدن جمله‌هایی مانند :

the cat was sleeping و the cat was sleeping

را قبلا دیده، پس نمایش dog و cat نزدیک هم است پس می‌تواند مقدار تخمین بزند.

محاسبه‌ی گرادیان ورودی

$$\nabla_{C(w)} l = \sum_{i=۱}^{n-۱} ۱_{(w_i=w)} W_i^{\top} \nabla_{a(x)} l \qquad (۱)$$

$$۱_{(w_i=w)} = \begin{cases} ۱ \; if \; w_i = w \\ ۰ \; if \; w_i \neq w \end{cases}$$

مدل زبانی شبکه عصبی

	n	c	h	m	direct	mix	train.	valid.	test.
MLP1	5		50	60	yes	no	182	284	268
MLP2	5		50	60	yes	yes		275	257
MLP3	5		0	60	yes	no	201	327	310
MLP4	5		0	60	yes	yes		286	272
MLP5	5		50	30	yes	no	209	296	279
MLP6	5		50	30	yes	yes		273	259
MLP7	3		50	30	yes	no	210	309	293
MLP8	3		50	30	yes	yes		284	270
MLP9	5		100	30	no	no	175	280	276
MLP10	5		100	30	no	yes		265	**252**
Del. Int.	3						31	352	336
Kneser-Ney back-off	3							334	323
Kneser-Ney back-off	4							332	321
Kneser-Ney back-off	5							332	321
class-based back-off	3	150						348	334
class-based back-off	3	200						354	340
class-based back-off	3	500						326	**312**
class-based back-off	3	1000						335	319
class-based back-off	3	2000						343	326
class-based back-off	4	500						327	312

تست روی داده‌گان APNews

	n	h	m	direct	mix	train.	valid.	test.
MLP10	6	60	100	yes	yes		104	**109**
Del. Int.	3						126	132
Back-off KN	3						121	127
Back-off KN	4						113	119
Back-off KN	5						112	**117**

مشکل پیچیدگی مدل

آزمایش روی یک داده‌گان معمولی و در دست رس مدت زمان طولانی طول می‌کشد.

پیچیدگی آموزش مدل $O(ndh + h|V|)$

$$V \approx ۱۸۰۰۰$$
$$d = ۱۰۰$$
$$h = ۶۰$$
$$n = ۶$$

$$ndh + h|V| = ۳۶۰۰۰ + ۱۰۸۰۰۰۰$$

راه‌حل مشکل

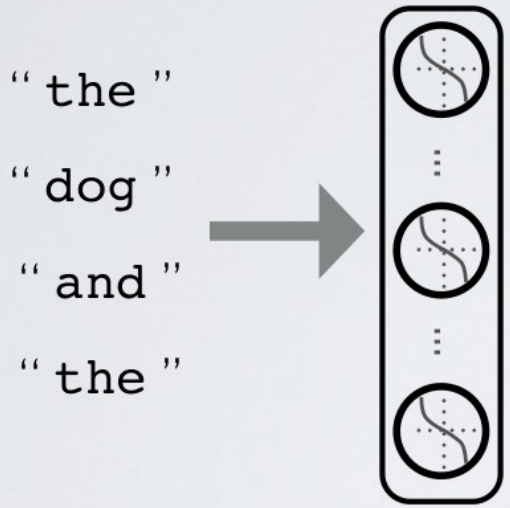
- اصل مشکل از لایه پنهان به لایه آخر می‌آید

hierarchical softmax می‌دهیم.

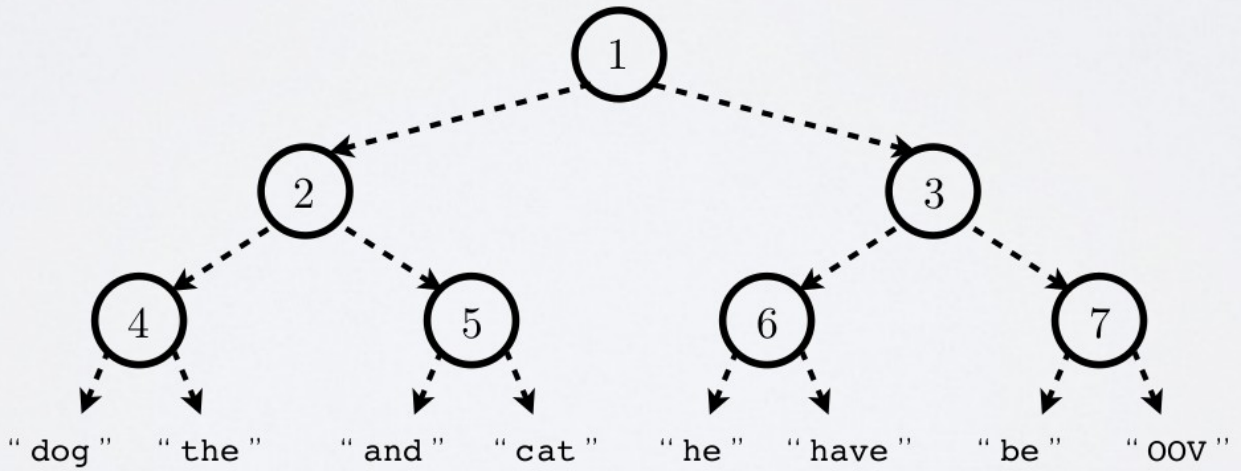
راه‌حل مشکل

- اصل مشکل از لایه پنهان به لایه آخر می‌آید
- برای حل کردن این مشکل، خروجی را به جای softmax به مدل hierarchical softmax می‌دهیم.

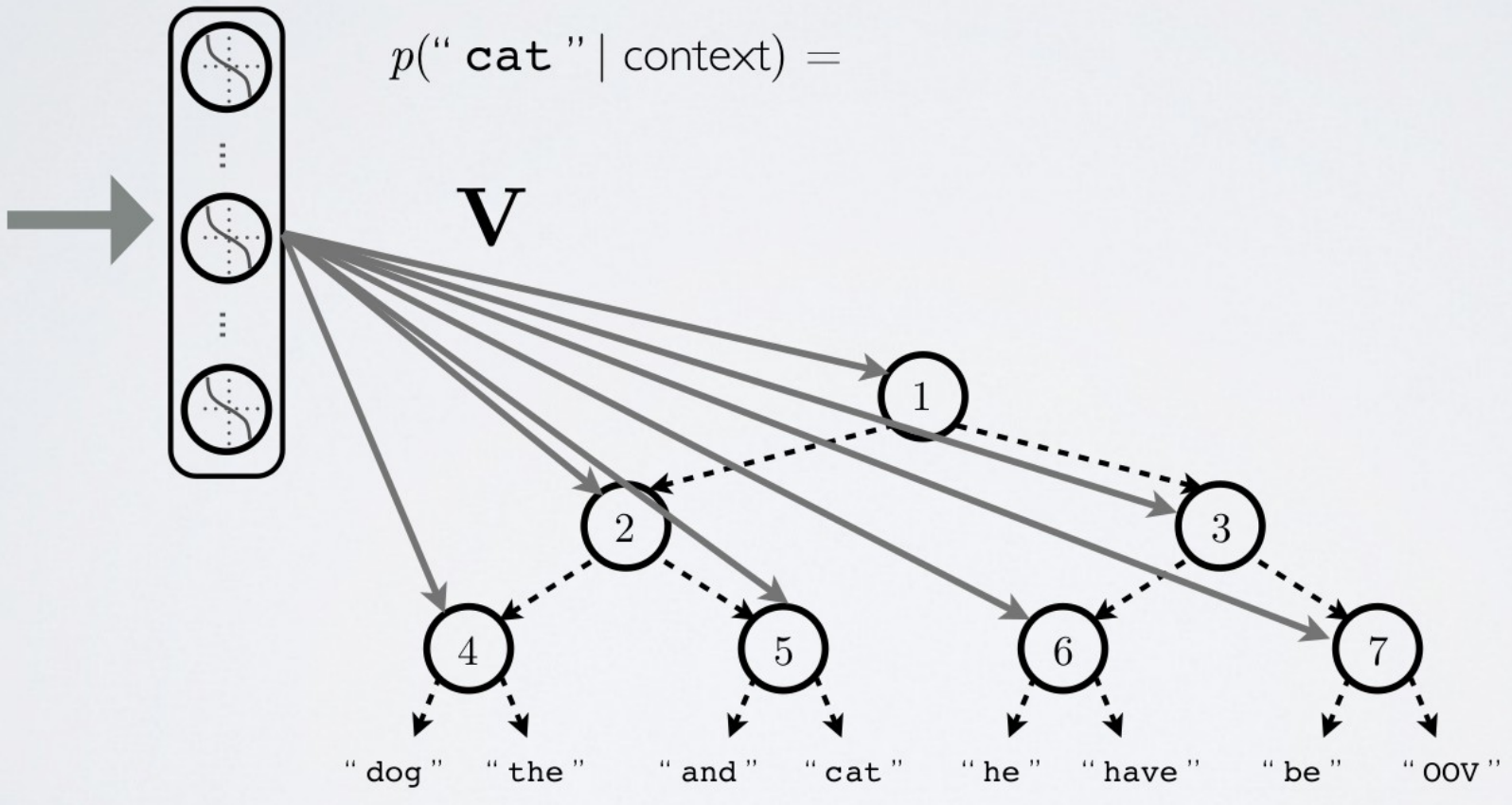
- Example: [“ the ”,“ dog ”,“ and ”, “ the ”, “ cat ”]



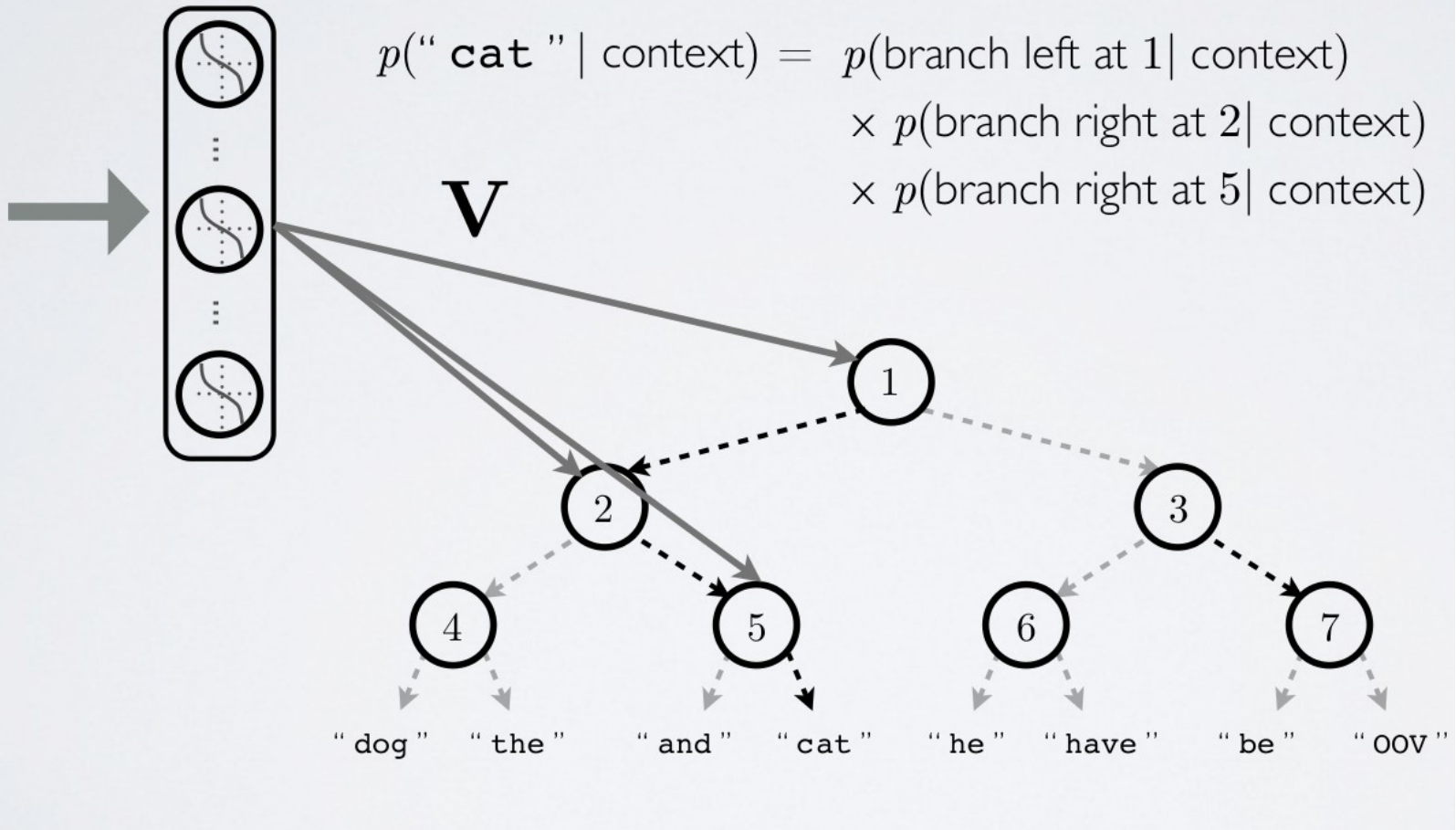
$p(\text{“ cat ”} \mid \text{context}) =$



- Example: [“ the ”,“ dog ”,“ and ”, “ the ”, “ cat ”]



- Example: [“ the ”,“ dog ”,“ and ”, “ the ”, “ cat ”]



مدل شبکه زبان عصبی

- شکل درخت دودویی خیلی تاثیر دارد

که به نام TF-IDF معروف است

مدل شبکه زبان عصبی

- شکل درخت دودویی خیلی تاثیر دارد
- زمان آموزش لایه آخر خیلی کمتر می‌شود

که به نام TF-IDF معروف است

مدل شبکه زبان عصبی

- شکل درخت دودویی خیلی تاثیر دارد
- زمان آموزش لایه آخر خیلی کمتر می‌شود
- زمان آزمون کمتر نمی‌شود

که به نام TF-IDF معروف است

مدل شبکه زبان عصبی

- شکل درخت دودویی خیلی تاثیر دارد
- زمان آموزش لایه آخر خیلی کمتر می‌شود
- زمان آزمون کمتر نمی‌شود
- درخت وردنت را می‌توانیم بسازیم

که به نام TF-IDF معروف است

مدل شبکه زبان عصبی

- شکل درخت دودویی خیلی تاثیر دارد
- زمان آموزش لایه آخر خیلی کمتر می‌شود
- زمان آزمون کمتر نمی‌شود
- درخت وردنت را می‌توانیم بسازیم
- برای تبدیل به درخت دودویی از الگوریتم سلسله مراتبی K-Means استفاده می‌کنیم

که به نام TF-IDF معروف است

مدل شبکه زبان عصبی

- شکل درخت دودویی خیلی تاثیر دارد
- زمان آموزش لایه آخر خیلی کمتر می‌شود
- زمان آزمون کمتر نمی‌شود
- درخت وردنت را می‌توانیم بسازیم
- برای تبدیل به درخت دودویی از الگوریتم سلسله مراتبی K-Means استفاده می‌کنیم
- از مدل فضای برداری استفاده می‌کنیم، با وزن دهی زیر

$$w_i^j = count_{word}(w_i, doc_j) * \log\left(\frac{T_{documents}}{containing_{documents}(w_i)}\right) \quad (۲)$$

که به نام TF-IDF معروف است

مدل شبکه زبان عصبی

architecture	Time per epoch (s)	Time per ex. (ms)	speed-up
original neural net	416 300	462.6	1
importance sampling	6 062	6.73	68.7
hierarchical model	1 609	1.79	258

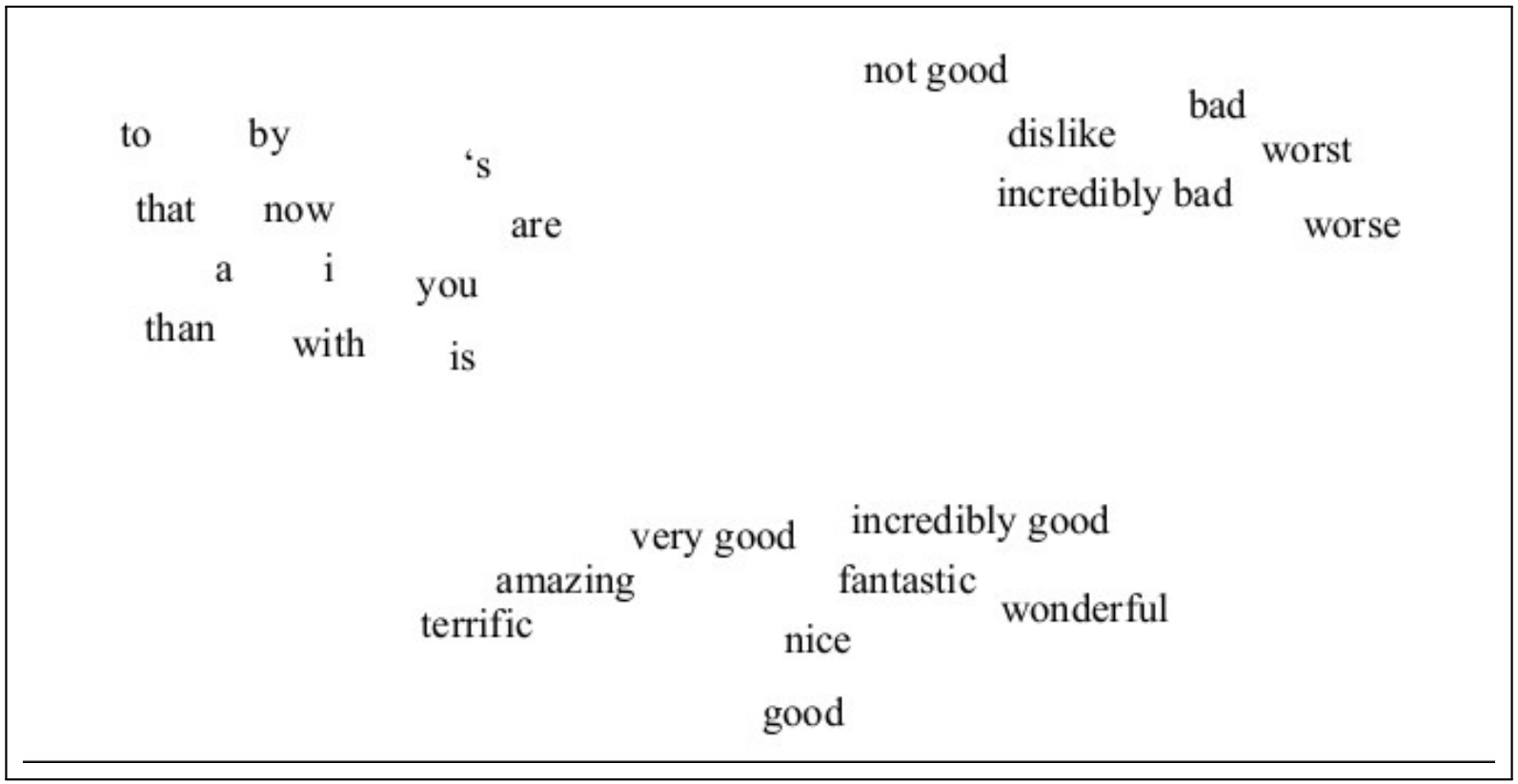
	Validation perplexity	Test perplexity
trigram	299.4	268.7
class-based	276.4	249.1
original neural net	213.2	195.3
importance sampling	209.4	192.6
hierarchical model	241.6	220.7

مدل شبکه زبان عصبی

architecture	Time per epoch (s)	Time per ex. (ms)	speed-up
original neural net	416 300	462.6	1
importance sampling	6 062	6.73	68.7
hierarchical model	1 609	1.79	258

	Validation perplexity	Test perplexity
trigram	299.4	268.7
class-based	276.4	249.1
original neural net	213.2	195.3
importance sampling	209.4	192.6
hierarchical model	241.6	220.7

مدل جادهی کلمات



مدل جادهی کلمات

- هدف این مدل‌ها این است که رابطه‌های بین کلمات را نشان بدهند

softmax مشکل لایه پنهان بیشتر مشخص می‌شود

مدل جادهی کلمات

- هدف این مدل‌ها این است که رابطه‌های بین کلمات را نشان بدهند
- در بخش قبلی حداکثر ۱۰۰ بوده و افزایش آن باعث افزایش پیچیدگی محاسباتی مدل می‌شود

softmax مشکل لایه پنهان بیشتر مشخص می‌شود

مدل جادهی کلمات

- هدف این مدل‌ها این است که رابطه‌های بین کلمات را نشان بدهند
- در بخش قبلی حداکثر ۱۰۰ بوده و افزایش آن باعث افزایش پیچیدگی محاسباتی مدل می‌شود
- [Mik+13b] متوجه شد که با حذف لایه پنهان، کارایی محاسباتی مدل افزایش می‌کند، که می‌شود روی داده‌های بیشتر آموزش داد

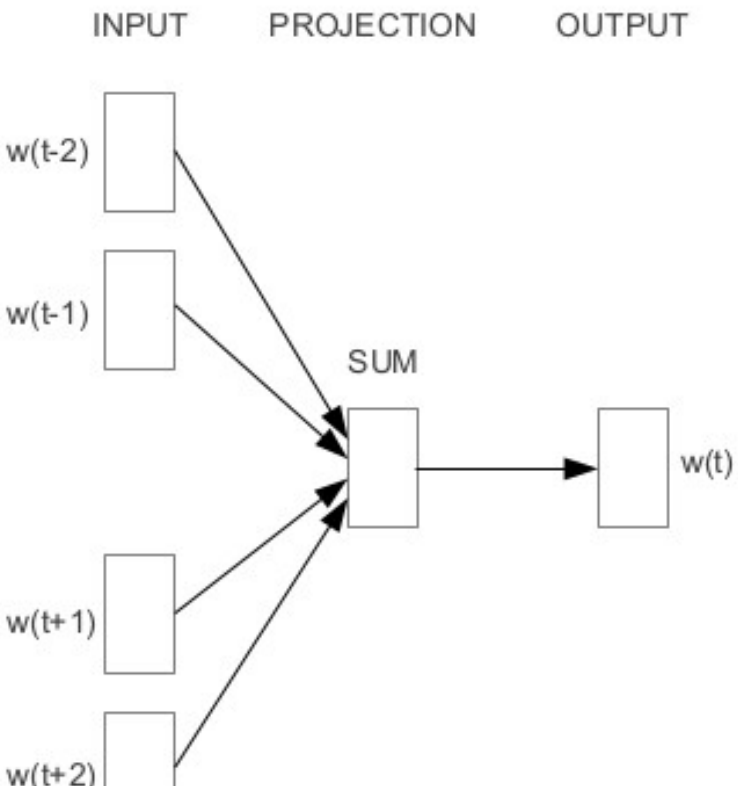
softmax مشکل لایه پنهان بیشتر مشخص می‌شود

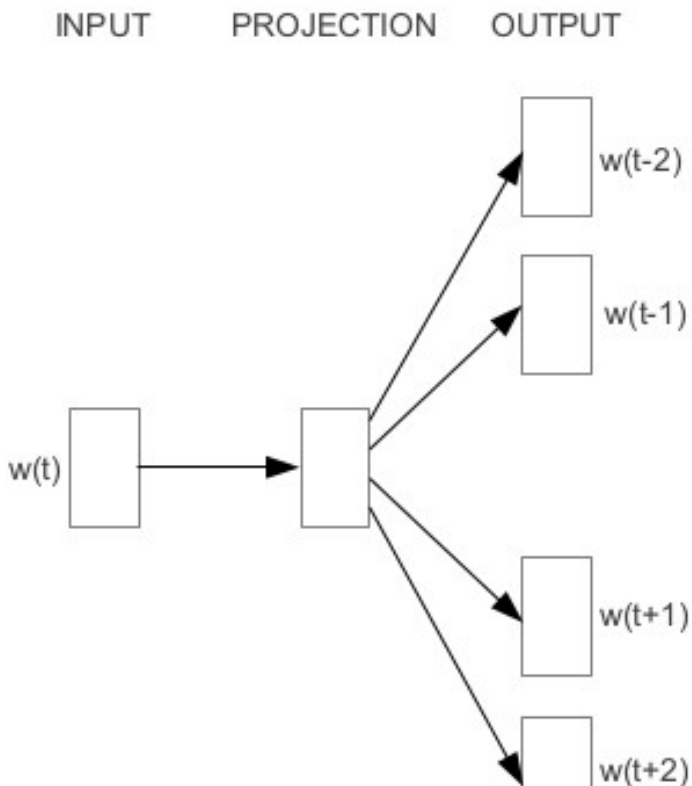
مدل جادهی کلمات

- هدف این مدل‌ها این است که رابطه‌های بین کلمات را نشان بدهند
- در بخش قبلی حداکثر ۱۰۰ بوده و افزایش آن باعث افزایش پیچیدگی محاسباتی مدل می‌شود
- [Mik+13b] متوجه شد که با حذف لایه پنهان، کارایی محاسباتی مدل افزایش می‌کند، که می‌شود روی داده‌های بیشتر آموزش داد
- که کارایی مدل را بسیار افزایش می‌دهد، چون بعد از کاهش مشکل لایه softmax مشکل لایه پنهان بیشتر مشخص می‌شود

مدل جادهی کلمات

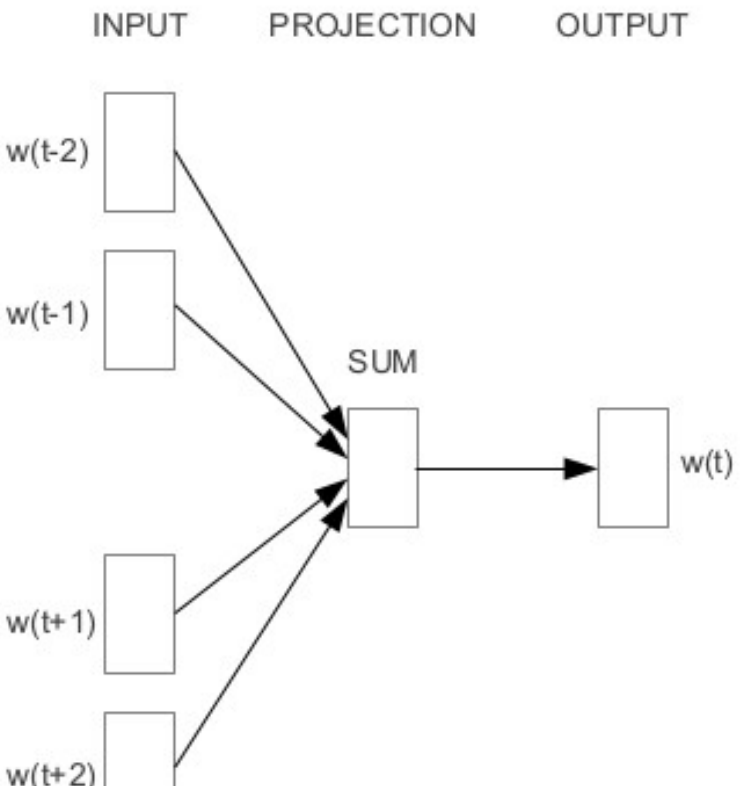
- دو نوع مدل وجود دارد CBOW و Skipgram

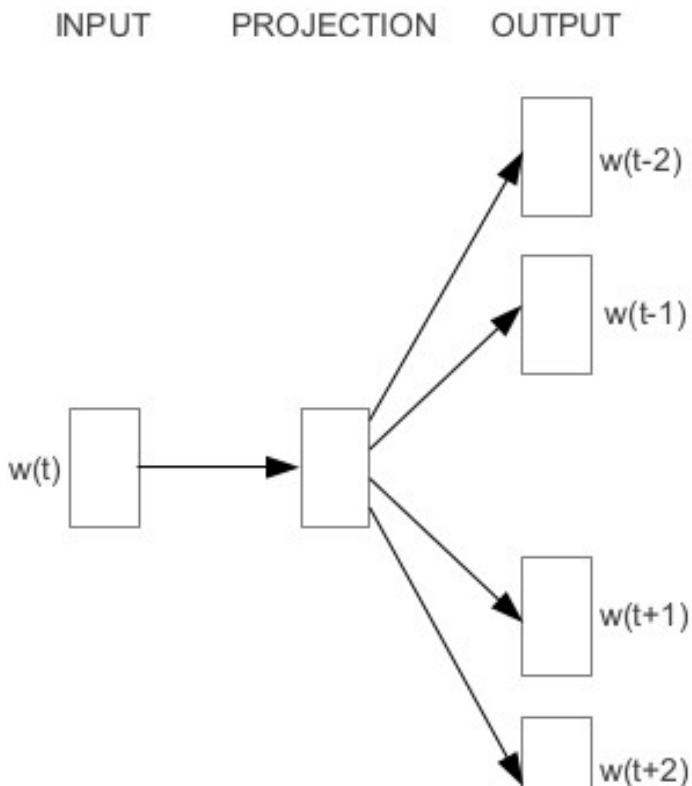




مدل جادهی کلمات

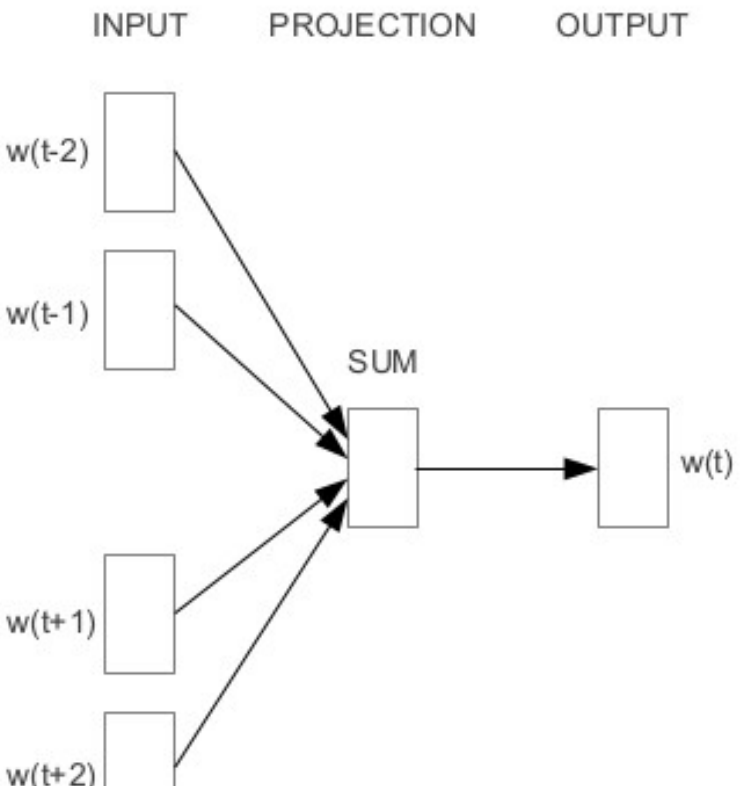
- دو نوع مدل وجود دارد CBOW و Skipgram
- ورودی one-hot encoding است.

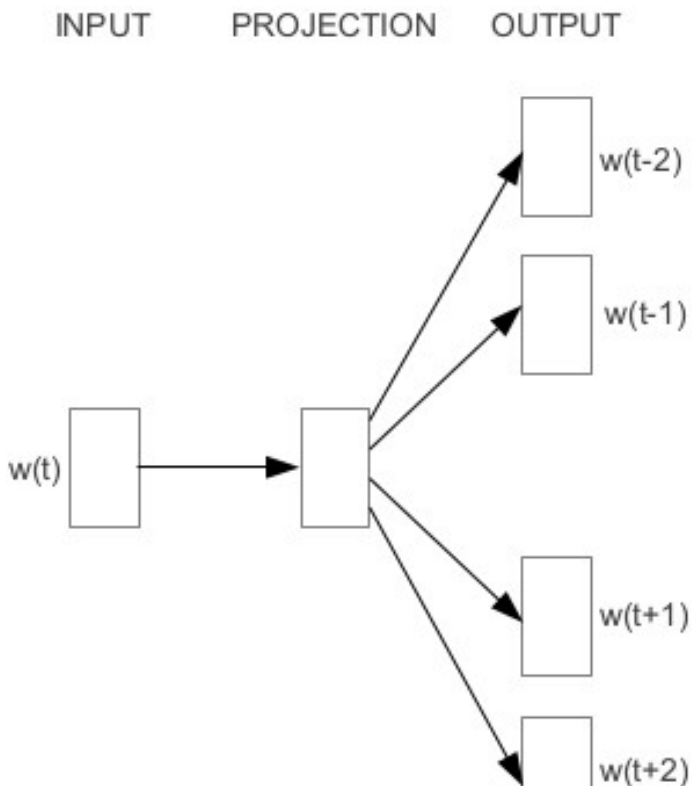




مدل جادهی کلمات

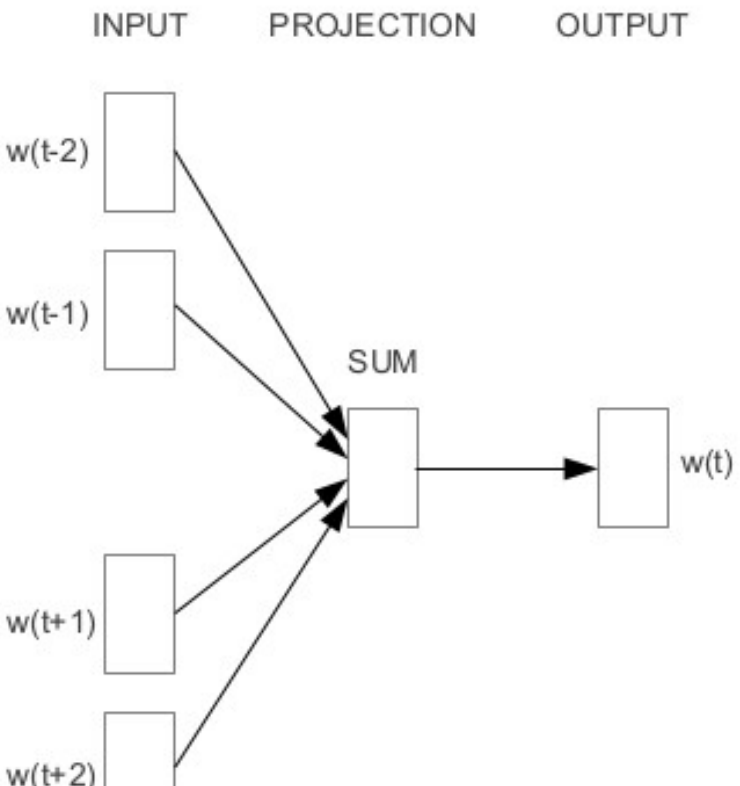
- دو نوع مدل وجود دارد CBOW و Skipgram
- ورودی one-hot encoding است.
- لایه بعدی، یک لایه projection است

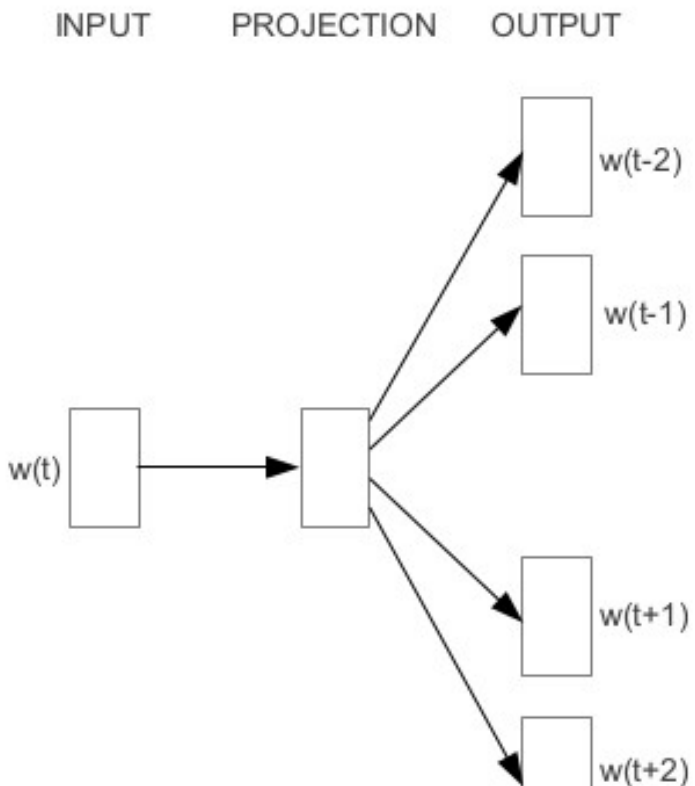




مدل جادهی کلمات

- دو نوع مدل وجود دارد CBOW و Skipgram
- ورودی one-hot encoding است.
- لایه بعدی، یک لایه projection است
- خروجی لایه سلسله مراتبی softmax است، با درختی خاص





مدل CBOW

- برای ورودی مدل، CBOW برای ورودی چند کلمه‌ی اطراف یک $2N + 1$ تایی کلمه را از متن می‌گیرد

اطراف پیش بینی کند

مدل CBOW

- برای ورودی مدل، CBOW برای ورودی چند کلمه‌ی اطراف یک $۲N + ۱$ تایی کلمه را از متن می‌گیرد
- N کلمه‌ی سمت راست و N کلمه‌ی سمت چپ را به عنوان ورودی به مدل می‌دهیم.

اطراف پیش بینی کند

مدل CBOW

- برای ورودی مدل، CBOW برای ورودی چند کلمه‌ی اطراف یک $۲N + ۱$ تایی کلمه را از متن می‌گیرد
- N کلمه‌ی سمت راست و N کلمه‌ی سمت چپ را به عنوان ورودی به مدل می‌دهیم.
- برای خروجی کلمه‌ی وسطی این $۲N + ۱$ کلمه را از مدل می‌خواهیم و با این کلمه به عنوان خروجی آموزش می‌دهیم.

اطراف پیش بینی کند

مدل CBOW

- برای ورودی مدل، CBOW برای ورودی چند کلمه‌ی اطراف یک $۲N + ۱$ تایی کلمه را از متن می‌گیرد
- N کلمه‌ی سمت راست و N کلمه‌ی سمت چپ را به عنوان ورودی به مدل می‌دهیم.
- برای خروجی کلمه‌ی وسطی این $۲N + ۱$ کلمه را از مدل می‌خواهیم و با این کلمه به عنوان خروجی آموزش می‌دهیم.
- وزن بین لایه‌ی ورودی و میانی، جادهی کلمات مورد نظر هستند.

اطراف پیش بینی کند

مدل CBOW

- برای ورودی مدل، CBOW برای ورودی چند کلمه‌ی اطراف یک $۲N + ۱$ تایی کلمه را از متن می‌گیرد
- N کلمه‌ی سمت راست و N کلمه‌ی سمت چپ را به عنوان ورودی به مدل می‌دهیم.
- برای خروجی کلمه‌ی وسطی این $۲N + ۱$ کلمه را از مدل می‌خواهیم و با این کلمه به عنوان خروجی آموزش می‌دهیم.
- وزن بین لایه‌ی ورودی و میانی، جادهی کلمات مورد نظر هستند.
- در لایه‌ی projection وزن‌های کلمه‌های ورودی جمع می‌شود.

اطراف پیش بینی کند

مدل CBOW

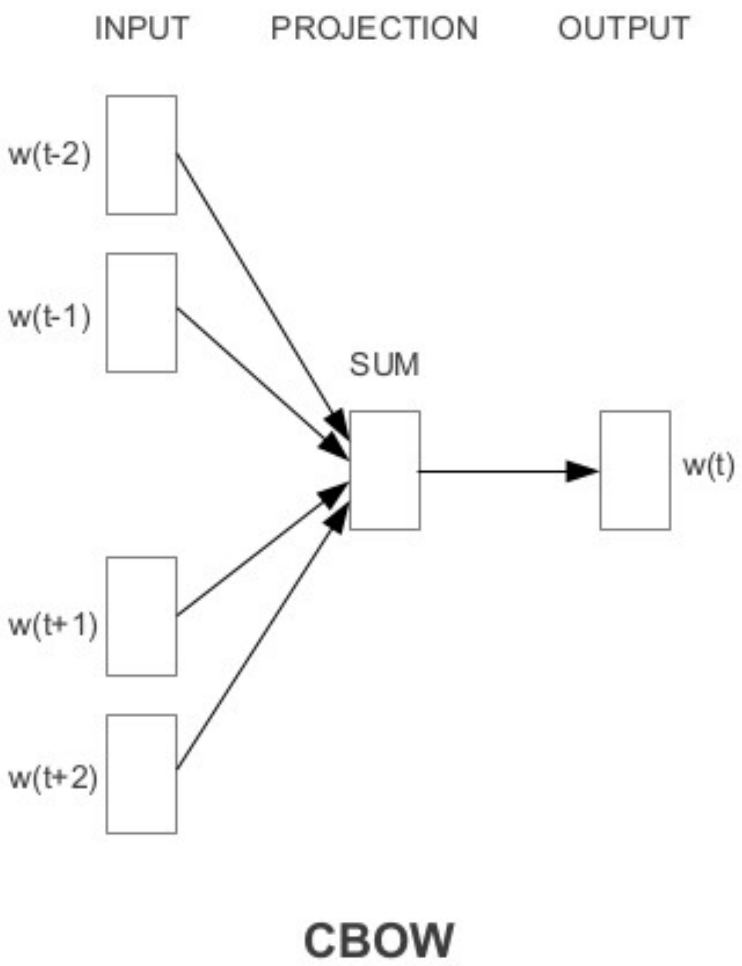
- برای ورودی مدل، CBOW برای ورودی چند کلمه‌ی اطراف یک $۲N + ۱$ تایی کلمه را از متن می‌گیرد
- N کلمه‌ی سمت راست و N کلمه‌ی سمت چپ را به عنوان ورودی به مدل می‌دهیم.
- برای خروجی کلمه‌ی وسطی این $۲N + ۱$ کلمه را از مدل می‌خواهیم و با این کلمه به عنوان خروجی آموزش می‌دهیم.
- وزن بین لایه‌ی ورودی و میانی، جادهی کلمات مورد نظر هستند.
- در لایه‌ی projection وزن‌های کلمه‌های ورودی جمع می‌شود.
- پس ترتیب کلمات مهم نیست، برای همین نام of Bag Continous Words گذاشته شده

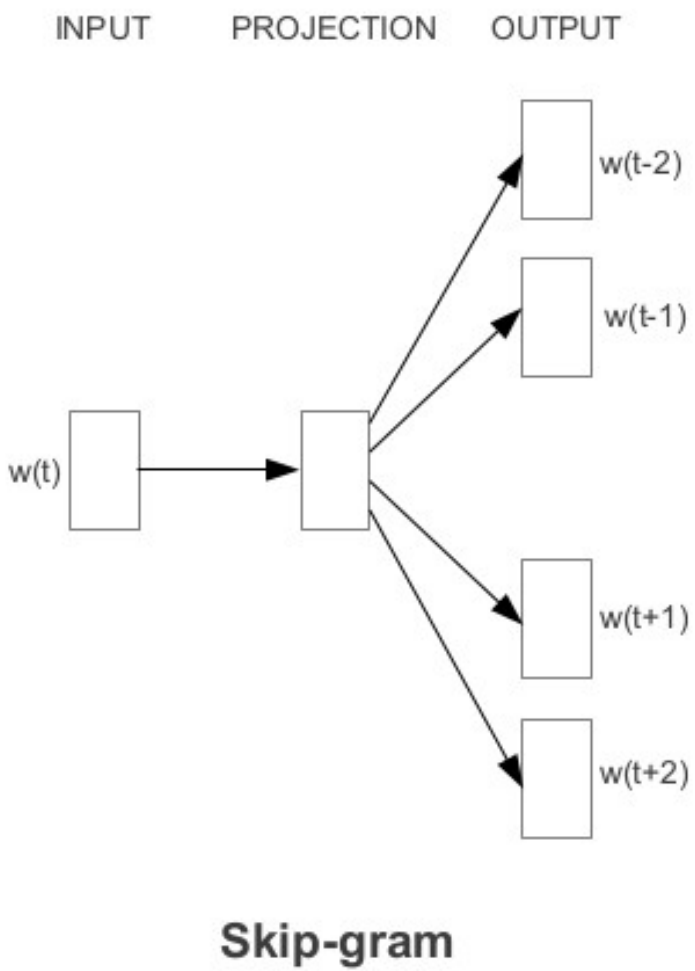
اطراف پیش بینی کند

مدل CBOW

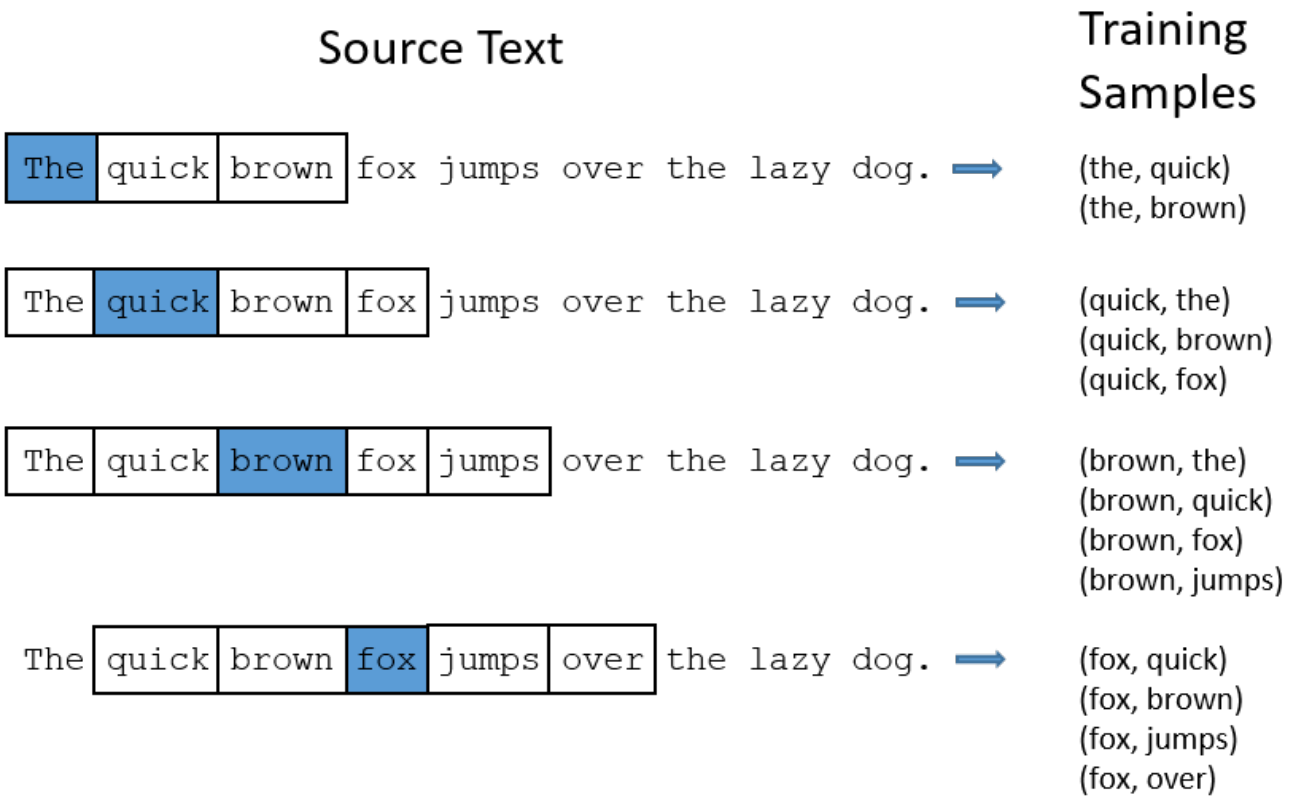
- برای ورودی مدل، CBOW برای ورودی چند کلمه‌ی اطراف یک $۲N + ۱$ تایی کلمه را از متن می‌گیرد
- N کلمه‌ی سمت راست و N کلمه‌ی سمت چپ را به عنوان ورودی به مدل می‌دهیم.
- برای خروجی کلمه‌ی وسطی این $۲N + ۱$ کلمه را از مدل می‌خواهیم و با این کلمه به عنوان خروجی آموزش می‌دهیم.
- وزن بین لایه‌ی ورودی و میانی، جادهی کلمات مورد نظر هستند.
- در لایه‌ی projection وزن‌های کلمه‌های ورودی جمع می‌شود.
- پس ترتیب کلمات مهم نیست، برای همین نام of Bag Continous Words گذاشته شده
- این مدل یاد می‌گیرد، با گرفتن پاره‌ای کلمه، کلمه‌ی میانی را از کلمه‌های اطراف پیش بینی کند

مدل Skipgram





مدل Skipgram



مدل Skipgram

- یک بازه‌ی متن به طول $۲N + ۱$ انتخاب می‌کنیم.

دست بدهد

مدل Skipgram

- یک بازه‌ی متن به طول $۲N + ۱$ انتخاب می‌کنیم.
- یک تعداد تصافی بین ۱ تا N انتخاب می‌کنیم به نام R

دست بدهد

مدل Skipgram

- یک بازه‌ی متن به طول $۲N + ۱$ انتخاب می‌کنیم.
- یک تعداد تصافی بین ۱ تا N انتخاب می‌کنیم به نام R
- R عدد سمت راست و همان مقدار عدد کلمه‌ی سمت راست را در نظر می‌گیریم

دست بدهد

مدل Skipgram

- یک بازه‌ی متن به طول $۲N + ۱$ انتخاب می‌کنیم.
- یک تعداد تصافی بین ۱ تا N انتخاب می‌کنیم به نام R
- R عدد سمت راست و همان مقدار عدد کلمه‌ی سمت راست را در نظر می‌گیریم
- هر کلمه‌ی $۲R$ را یک بار به عنوان خروجی در نظر می‌گیریم، ورودی کلمه‌ی میانی پاره کلمات از متن است

دست بدهد

مدل Skipgram

- یک بازه‌ی متن به طول $۲N + ۱$ انتخاب می‌کنیم.
- یک تعداد تصافی بین ۱ تا N انتخاب می‌کنیم به نام R
- R عدد سمت راست و همان مقدار عدد کلمه‌ی سمت راست را در نظر می‌گیریم
- هر کلمه‌ی $۲R$ را یک بار به عنوان خروجی در نظر می‌گیریم، ورودی کلمه‌ی میانی پاره کلمات از متن است
- این مدل یاد می‌گیرد که با ورودی گرفتن کلمه، توزیع کلمه‌های اطراف را به دست بدهد

درخت huffman

- برای لایه آخر دور راه حل داریم

- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

درخت huffman

- برای لایه آخر دور راه حل داریم
- که راه اول لایه‌ی خروجی درختی است

- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

درخت huffman

- برای لایه آخر دور راه حل داریم
- که راه اول لایه‌ی خروجی درختی است
- ساختن درخت هافمن به جای درخت دودویی باعث می‌شود پیچیدگی کمتر شود
- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

درخت huffman

- برای لایه آخر دور راه حل داریم
- که راه اول لایه‌ی خروجی درختی است
- ساختن درخت هافمن به جای درخت دودویی باعث می‌شود پیچیدگی کمتر شود
- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

- برای لایه آخر دور راه حل داریم

- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

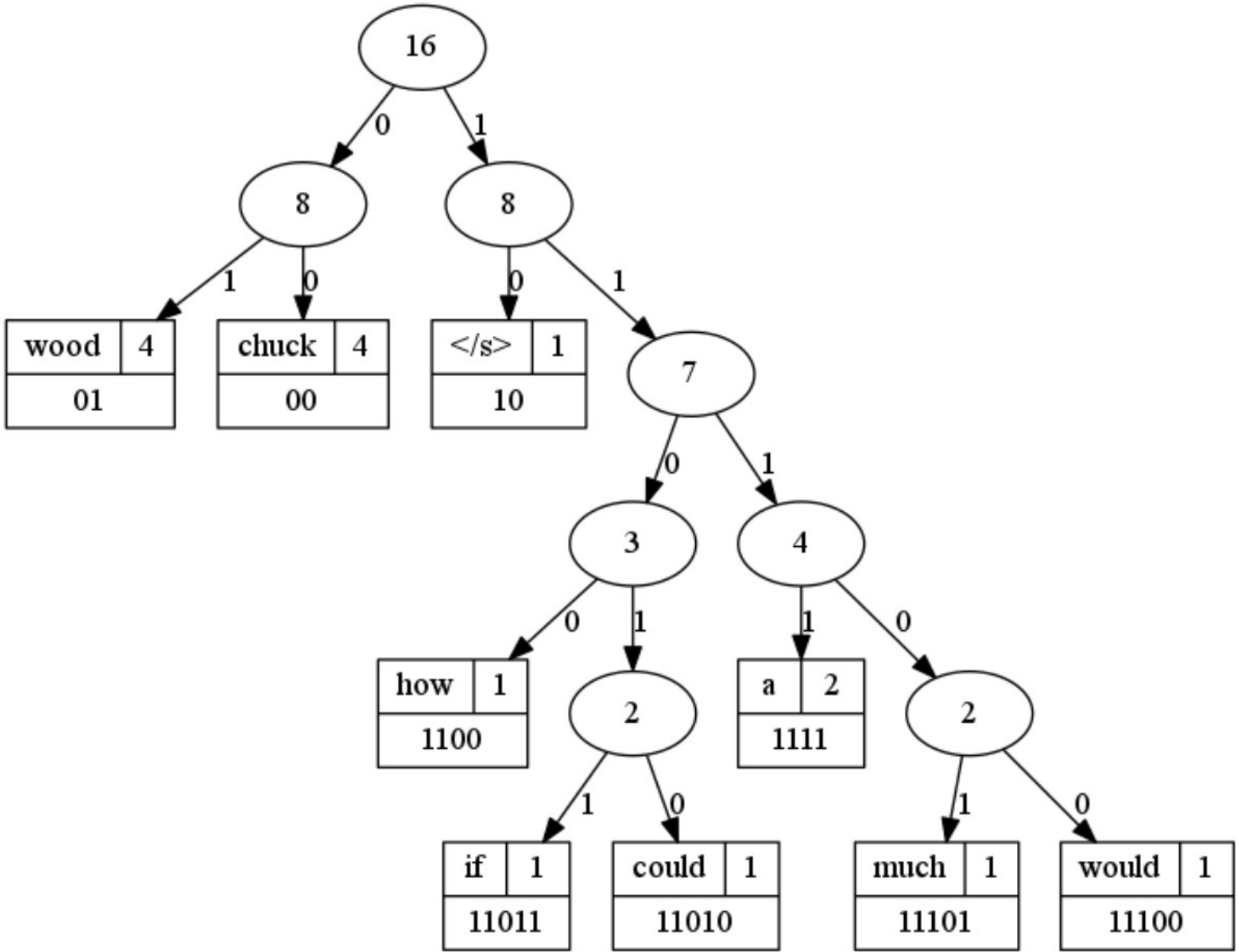
- برای لایه آخر دور راه حل داریم
- که راه اول لایه‌ی خروجی درختی است

- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

- برای لایه آخر دور راه حل داریم
- که راه اول لایه‌ی خروجی درختی است
- ساختن درخت هافمن به جای درخت دودویی باعث می‌شود پیچیدگی کمتر شود
- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

- برای لایه آخر دور راه حل داریم
- که راه اول لایه‌ی خروجی درختی است
- ساختن درخت هافمن به جای درخت دودویی باعث می‌شود پیچیدگی کمتر شود
- در یک داده‌گان با یک میلیون کلمه، باعث کاهش زمان به نصف می‌شود

درخت هافمن


16
0
1
8
8
1
0
0
1
wood 4
01
chuck 4
00
</s> 1
10
7
0
1
3
4
0
1
1
0
how 1
1100
2
a 2
1111
2
1
0
1
0
if 1
11011
could 1
11010
much 1
11101
would 1
11100

تابع هدف

هدف ماکسیمم کرد تابع زیر است:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-C\leq j\leq C, j\neq 0}\log P(w_{t+j}|w_t)$$

احتمال بالا به شکل زیر است:

$$p(w_O|w_I)=\frac{exp(v'^{\top}_{w_O}v_{w_I})}{\sum_{w=1}^{W}exp(v'^{\top}_{w}v_{w_I})}$$

اندازه‌گیری کارایی مدل

- کارای مدل در [Mik+13b] به صورت رابطه‌هایی بردارهای جادهی برقرار شده اندازه می‌گیرد

اندازه‌گیری کارایی مدل

- کارای مدل در [Mik+13b] به صورت رابطه‌هایی بردارهای جادهی برقرار شده اندازه می‌گیرد
- این رابطه‌ها به صورت شگفت انگیزی با محاسبات ساده جبر خطی بین بردارها فابل اندازه‌گیری هستند:

$$x = vector(\text{"}bigger\text{"}) - vector(\text{"}big\text{"}) + vector(\text{"}small\text{"})$$

اندازه‌گیری کارایی مدل

- کارای مدل در [Mik+13b] به صورت رابطه‌هایی بردارهای جادهی برقرار شده اندازه می‌گیرد
- این رابطه‌ها به صورت شگفت انگیزی با محاسبات ساده جبر خطی بین بردارها فابل اندازه‌گیری هستند:

$$x = vector(\text{"}bigger\text{"}) - vector(\text{"}big\text{"}) + vector(\text{"}small\text{"})$$

- برای پیدا کردن، کافی است فاصله‌ی کوسینوسی بردار x با همه‌ی بردارهای یادگیرفته شده توسط مدل اندازه بگیریم و نزدیک ترین بردار به این، بردار *smaller* هست.

اندازه‌گیری کارایی مدل

- کارای مدل در [Mik+13b] به صورت رابطه‌هایی بردارهای جادهی برقرار شده اندازه می‌گیرد
- این رابطه‌ها به صورت شگفت انگیزی با محاسبات ساده جبر خطی بین بردارها فابل اندازه‌گیری هستند:

$$x = vector(\text{"}bigger\text{"}) - vector(\text{"}big\text{"}) + vector(\text{"}small\text{"})$$

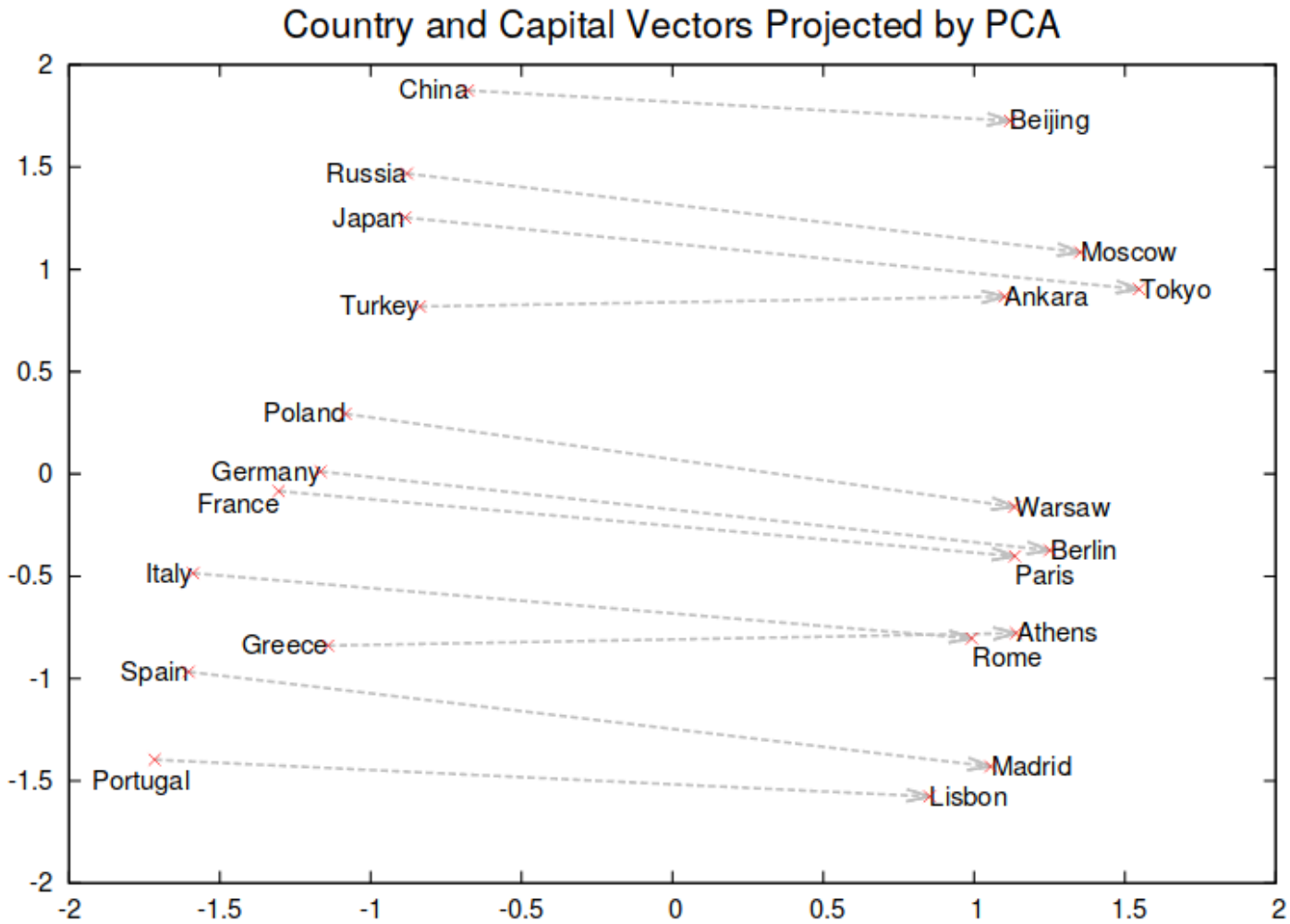
- برای پیدا کردن، کافی است فاصله‌ی کوسینوسی بردار x با همه‌ی بردارهای یادگیرفته شده توسط مدل اندازه بگیریم و نزدیک ترین بردار به این، بردار $smaller$ هست.
- رابطه‌های معنایی معناشناسانه نیز به این شکل برقرار است.

$$d_{cos}(a,b) = ۱ - \frac{a.b}{\|a\| \, \|b\|}$$

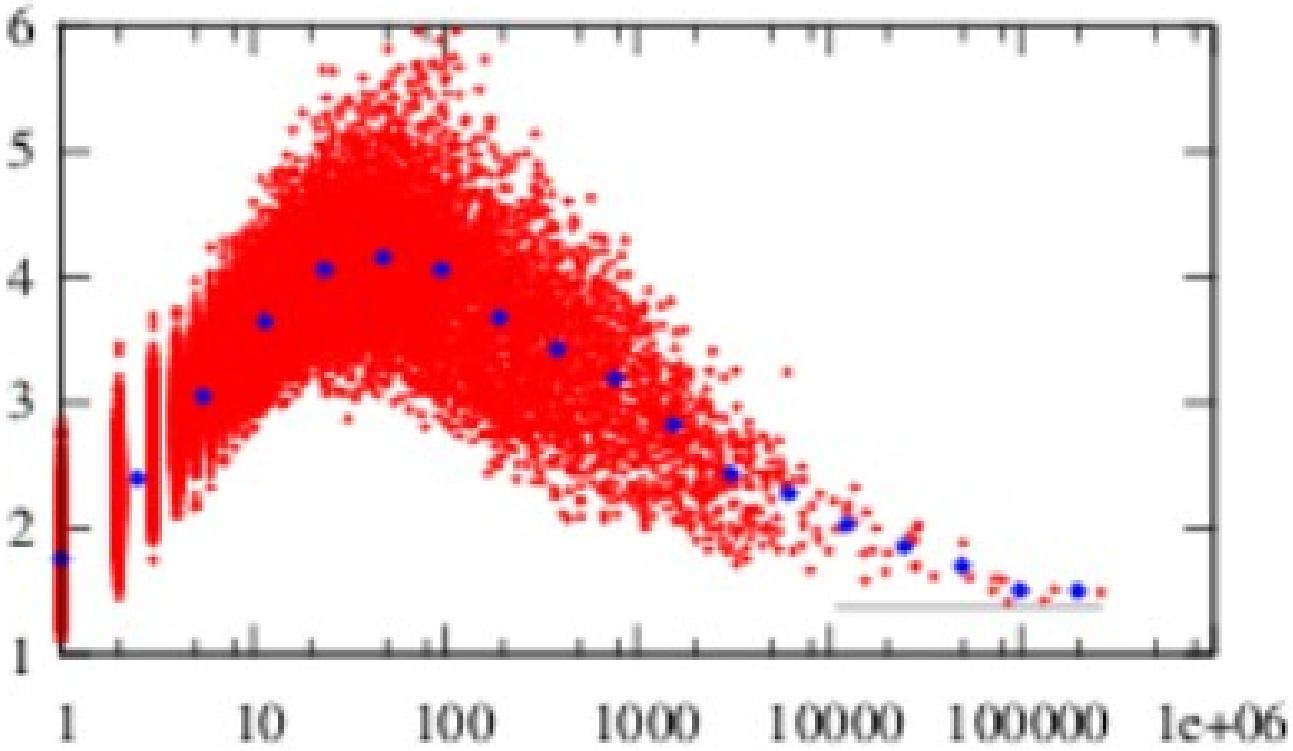
$$x = vector(\text{"}King\text{"}) - vector(\text{"}man\text{"}) + vector(\text{"}woman\text{"})$$

$$\arg\min_{y} d_{cos}(x,y) == vector(\text{"}Queen\text{"})$$

اندازه‌گیری کارایی مدل



چرا فاصله‌ی کوسینوسی



شرایط آزمایش‌ها

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

کارایی مدل‌ها

Dimensionality / Training words	24M	49M	98M	196M	391M	783M
50	13.4	15.7	18.6	19.1	22.5	23.2
100	19.4	23.1	27.8	28.7	33.4	32.2
300	23.2	29.2	35.3	38.6	43.7	45.9
600	24.0	30.1	36.5	40.8	46.6	50.4

کارایی مدل‌ها

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

بهتر کردن مدل

- از نظر محاسباتی کند است
- راه حل دوم نمونه‌گیری منفی

بهتر کردن مدل

- از نظر محاسباتی کند است
- راه حل اول زیر نمونه گیری
- راه حل دوم نمونه‌گیری منفی

بهتر کردن مدل

- از نظر محاسباتی کند است
- راه حل اول زیر نمونه گیری
- راه حل دوم نمونه‌گیری منفی

کم نمونه‌گیری

- کلمه‌هایی مثل «یا» «و» و ... اطلاعات زیادی ندارند اما تعداد زیادی دارند

کم نمونه‌گیری

- کلمه‌هایی مثل «یا» «و» و ... اطلاعات زیادی ندارند اما تعداد زیادی دارند
- با احتمال زیر مثال را در آموزش لحاظ نمی‌کنیم

$$P(w_i) = ۱ - \sqrt{\frac{t}{f(w_i)}} \qquad (۳)$$

نمونه‌گیری منفی

- مشکل محاصبه‌ی مخرج تابع softmax است

نمونه‌گیری منفی

- مشکل محاصبه‌ی مخرج تابع softmax است
- مخرج را در بعضی مدل‌های زبان با روش‌های مختلف تخمین می‌زنند

نمونه‌گیری منفی

- مشکل محاصبه‌ی مخرج تابع softmax است
- مخرج را در بعضی مدل‌های زبان با روش‌های مختلف تخمین می‌زنند
- تخمین دقیق در مدل زبان، کاربرد دارد اما اینجا فقط رابطه‌ی بین کلمات مورد نظر ماست

نمونه‌گیری منفی

- مشکل محاصبه‌ی مخرج تابع softmax است
- مخرج را در بعضی مدل‌های زبان با روش‌های مختلف تخمین می‌زنند
- تخمین دقیق در مدل زبان، کاربرد دارد اما اینجا فقط رابطه‌ی بین کلمات مورد نظر ماست
- تابع توزیع خروجی به نسبت ورودی را به شکل زیر می‌نویسیم

$$\log \sigma(v'^{\top}_{w_O} v_{w_I}) + \sum_{i=1}^{k} E_{w_i \sim P_n(w)} \left[\log \sigma(-v'^{\top}_{w_i} v_{w_I}) \right] \qquad (۴)$$

و به عنوان هر یک از $P(w_O|w_I)$ جایگزین می‌کنیم

نمونه‌گیری منفی

- مشکل محاصبه‌ی مخرج تابع softmax است
- مخرج را در بعضی مدل‌های زبان با روش‌های مختلف تخمین می‌زنند
- تخمین دقیق در مدل زبان، کاربرد دارد اما اینجا فقط رابطه‌ی بین کلمات مورد نظر ماست
- تابع توزیع خروجی به نسبت ورودی را به شکل زیر می‌نویسیم

$$\log \sigma(v'^{\top}_{w_O} v_{w_I}) + \sum_{i=1}^{k} E_{w_i \sim P_n(w)} \left[\log \sigma(-v'^{\top}_{w_i} v_{w_I}) \right] \qquad (۴)$$

و به عنوان هر یک از $P(w_O|w_I)$ جایگزین می‌کنیم

- نویز پیشنهادی در [Mik+13a]

$$P_n(w) = \frac{U(w)^{3/4}}{\sum_{j=1}^{W} (U(w_j)^{3/4})} \qquad (۵)$$

کارایی نمونه‌گیری منفی و کم‌نمونه گیری

Method	Time [min]	Syntactic [%]	Semantic [%]	Total accuracy [%]
NEG-5	38	63	54	59
NEG-15	97	63	58	**61**
HS-Huffman	41	53	40	47
NCE-5	38	60	45	53
The following results use 10^{-5} subsampling				
NEG-5	14	61	58	60
NEG-15	36	61	61	**61**
HS-Huffman	21	52	59	55

- مدل مرجع برای word2vec:

- مدل مرجع برای word2vec:
- داده‌گان ویکیپدیای فارسی، سال ۲۰۱۶، آموزش توسط آزمایشگاه یادگیری دانشگاه قم.

- مدل مرجع برای word2vec:
- داده‌گان ویکیپدیای فارسی، سال ۲۰۱۶، آموزش توسط آزمایشگاه یادگیری دانشگاه قم.
- پیش پردازش شامل: حذف اعراب، جایگزینی حروف عربی با فارسی، نرمال سازی حروف اضافه، حذف اعداد

- مدل مرجع برای word2vec:
- داده‌گان ویکیپدیای فارسی، سال ۲۰۱۶، آموزش توسط آزمایشگاه یادگیری دانشگاه قم.
- پیش پردازش شامل: حذف اعراب، جایگزینی حروف عربی با فارسی، نرمال سازی حروف اضافه، حذف اعداد
- آموزش Skipgram با Negative Sampling با ۱۰۰ بعد

```
w2v.most_similar("ریاضی")
```

```
[('0.8600557446479797 ,'ریاضیات),
 ('0.7547142505645752 ,'ریاضیاتی),
 ('0.7294837236404419 ,'جبری),
 ('0.7266465425491333 ,'ترکیبیات),
 ('0.7136733531951904 ,'محاسباتی),
 ('0.7023245096206665 ,'ترکیبیاتی),
 ('0.695412278175354 ,'جبر),
 ('0.6906285285949707 ,'آنالیز),
 ('0.6899727582931519 ,'فیزیک),
 ('0.6870720386505127 ,'چندمتغیره)]
```

```
w2v.most_similar("برلین")
```

```
[('0.7904973030090332 ,'هامبورگ'),
 ('0.7537745237350464 ,'مونیخ'),
 ('0.7492792010307312 ,'فرانکفورت'),
 ('0.748837947845459 ,'آلمان'),
 ('0.7343987226486206 ,'کلن'),
 ('0.7133944630622864 ,'لایپزیگ'),
 ('0.7073416113853455 ,'درسدن'),
 ('0.6995538473129272 ,'اشتوتگارت'),
 ('0.6772368550300598 ,'دوسلدورف'),
 ('0.6666285991668701 ,'مانهایم')]
```

```
w2v.most_similar_cosmul(positive=["شاه" ,"زن"], negative=['مرد'],
```

```
[('0.8861917853355408 ,'سلطان),
 ('0.8850266337394714 ,'قباچه),
 ('0.8567750453948975 ,'پادشاه),
 ('0.8547272682189941 ,'دربار),
 ('0.8440592288970947 ,'درباریان),
 ('0.8375454545021057 ,'احمدشاه),
 ('0.836741030216217 ,'ولیعهد),
 ('0.8310530185699463 ,'شاهنشاه),
```

```
w2v.most_similar(positive=["آلمان" ,"شهر"], negative=['کشور'])
```

```
[('0.5935142040252686 ,'درسدن),
 ('0.59113609790802 ,'دوسلدورف),
 ('0.5882521867752075 ,'برلین),
 ('0.5763571262359619 ,'لوبک),
 ('0.5709136128425598 ,'مونیخ),
 ('0.5681554079055786 ,'فرانکفورت),
 ('0.5676054954528809 ,'زاکسن),
 ('0.5672116279602051 ,'هامبورگ),
 ('0.566654813687134 ,'کلن),
 ('0.5638121366500854 ,'اولدنبورگ)]
```

```
w2v.most_similar_cosmul(positive=["ایران" ,"شهر"], negative=['کشور'])
```

```
[('1.1280869245529175 ,'کاشان'),
 ('1.0347970724105835 ,'کرج'),
 ('1.0239100456237793 ,'شیراز'),
 ('1.0002652406692505 ,'اردهال'),
 ('0.9985761046409607 ,'بروجرد'),
 ('0.9854745864868164 ,'ورامین'),
 ('0.9822379946708679 ,'اصفهان'),
 ('0.978914201259613 ,'شهرکرد'),
 ('0.9765846729278564 ,'سیرجان'),
 ('0.974379301071167 ,'برازجان')]
```

```
w2v.most_similar_cosmul(positive=["فرانسه" ,"شهر"], negative=['کشور'])
```

```
[('0.9347296357154846 ,'آوینیون'),
 ('0.8908623456954956 ,'ولانس'),
 ('0.8822587728500366 ,'آکویتین'),
 ('0.8812925815582275 ,'اورلئان'),
 ('0.8812360763549805 ,'تولوز'),
 ('0.8724114298820496 ,'استراسبورگ'),
 ('0.8702823519706726 ,'دازور'),
 ('0.8694525361061096 ,'آرل'),
 ('0.8692155480384827 ,'اوآز'),
 ('0.8659394979476929 ,'روشایدر')]
```

```
w2v.most_similar(positive=["پرسپولیس" ,"سپاهان"], negative=['استقلال'])
```

```
[('ذوب\u200c0.7490308284759521 ,'آهن'),
 ('0.640939474105835 ,'تراکتورسازی'),
 ('0.6299975514411926 ,'سایپا'),
 ('گیتی\u200c0.619982123374939 ,'پسند'),
 ('0.5834908485412598 ,'جباری'),
 ('0.5777952671051025 ,'صبا باتری'),
```

```
w2v.most_similar("چلسی")
```

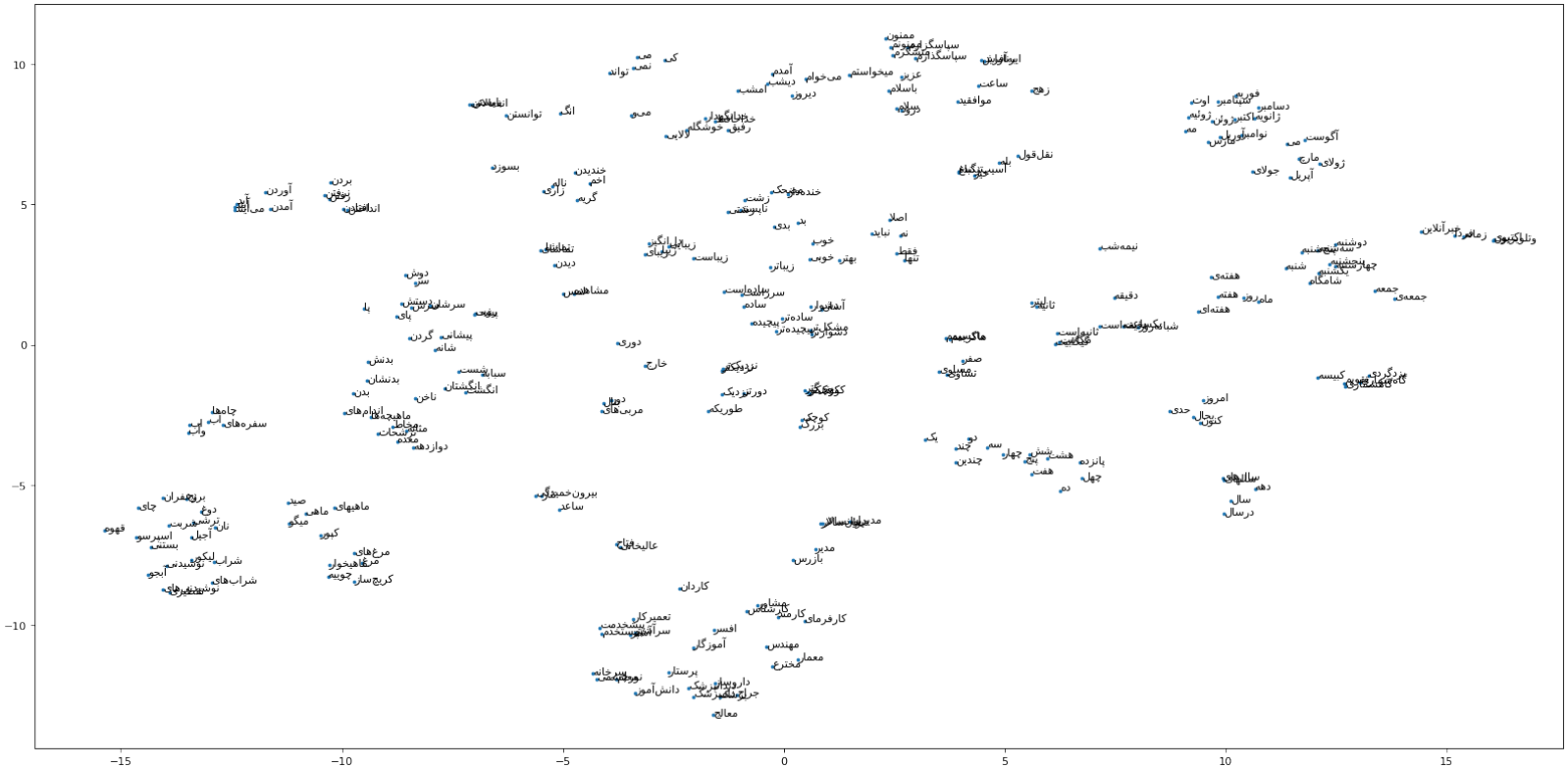
```
[('0.8139051198959351 ,'آرسنال'),
 ('0.7773184776306152 ,'لیورپول'),
 ('0.6983660459518433 ,'اورتون'),
 ('0.6974737644195557 ,'منچستریونایتد'),
 ('0.6824600100517273 ,'تاتنهام'),
 ('0.667028546333313 ,'فولهام'),
 ('0.6431244015693665 ,'فولام'),
 ('0.6123976111412048 ,'بلکبرن'),
 ('0.6099734306335449 ,'بولتون'),
 ('0.6097316741943359 ,'یونایتد')]
```

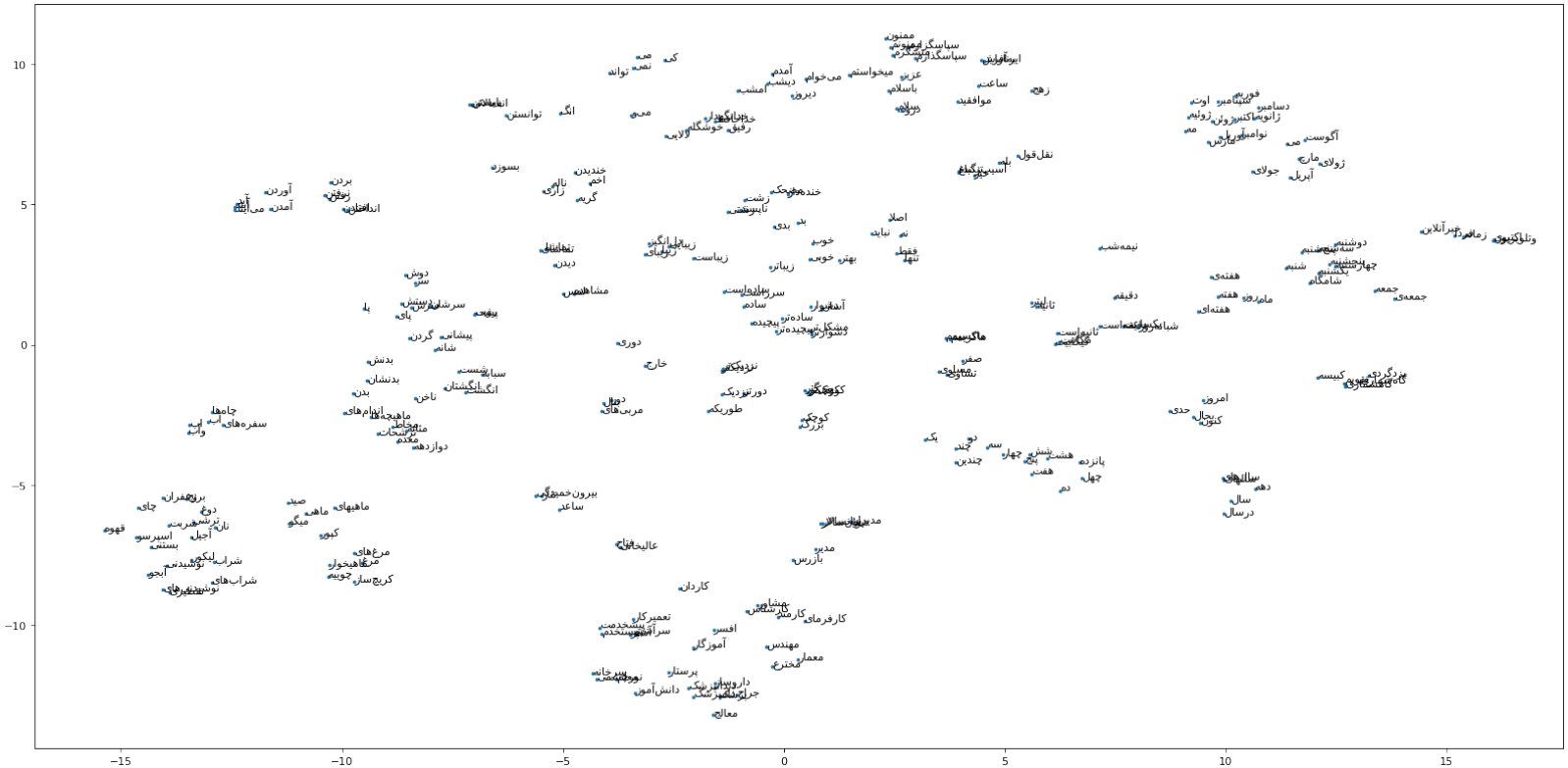
```
w2v.most_similar_cosmul(positive=["ملوان" ,"اصفهان"], negative=['انزلی'])
```

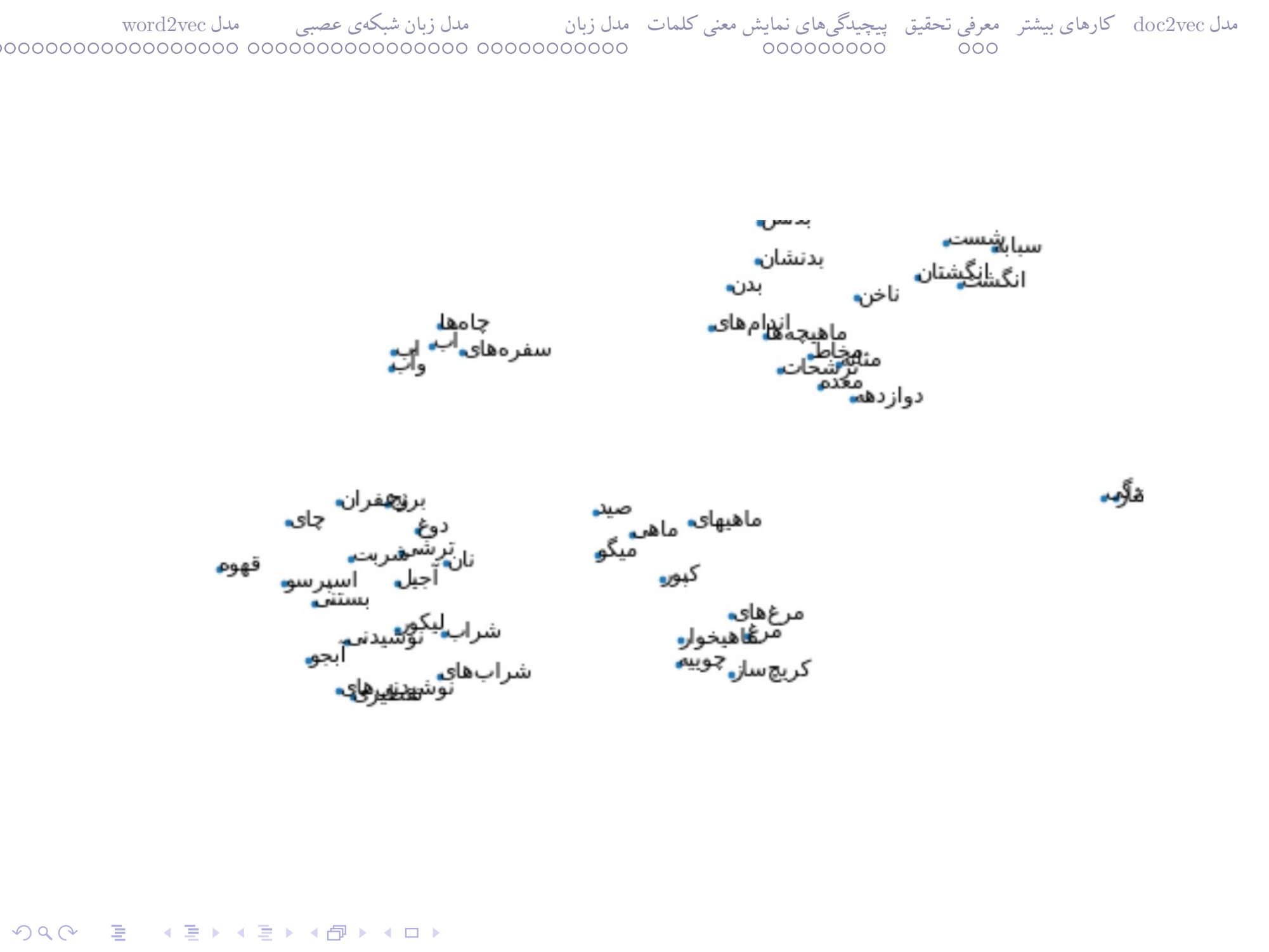
```
[('0.8954960107803345 ,'سپاهان'),
 ('ذوب\u200c0.8334643840789795 ,'آهن'),
 ('0.8287364840507507 ,'پرسپولیس'),
 ('0.8029134273529053 ,'جباری'),
 ('0.7976706027984619 ,'کرانچار'),
```

```
most_similar_cosmul(positive=["اصفهان" ,"تراکتورسازی"], negative=['تبریز'])
```

```
[('ذوب\u200c0.9527584910392761 ,'آهن'),
 ('0.9380269050598145 ,'سپاهان'),
 ('پلی\u200c0.8899718523025513 ,'اکریل'),
 ('گیتی\u200c0.8810611963272095 ,'پسند'),
 ('فولادماهان\u200c0.8765514492988586 ,'سپاهان'),
 ('0.8403027653694153 ,'فولادماهان'),
 ('0.8379800915718079 ,'اسکوچیچ'),
 ('0.8343930840492249 ,'باریج'),
 ('0.8309004306793213 ,'پرسپولیس'),
 ('0.8270416259765625 ,'کرانچار')]
```

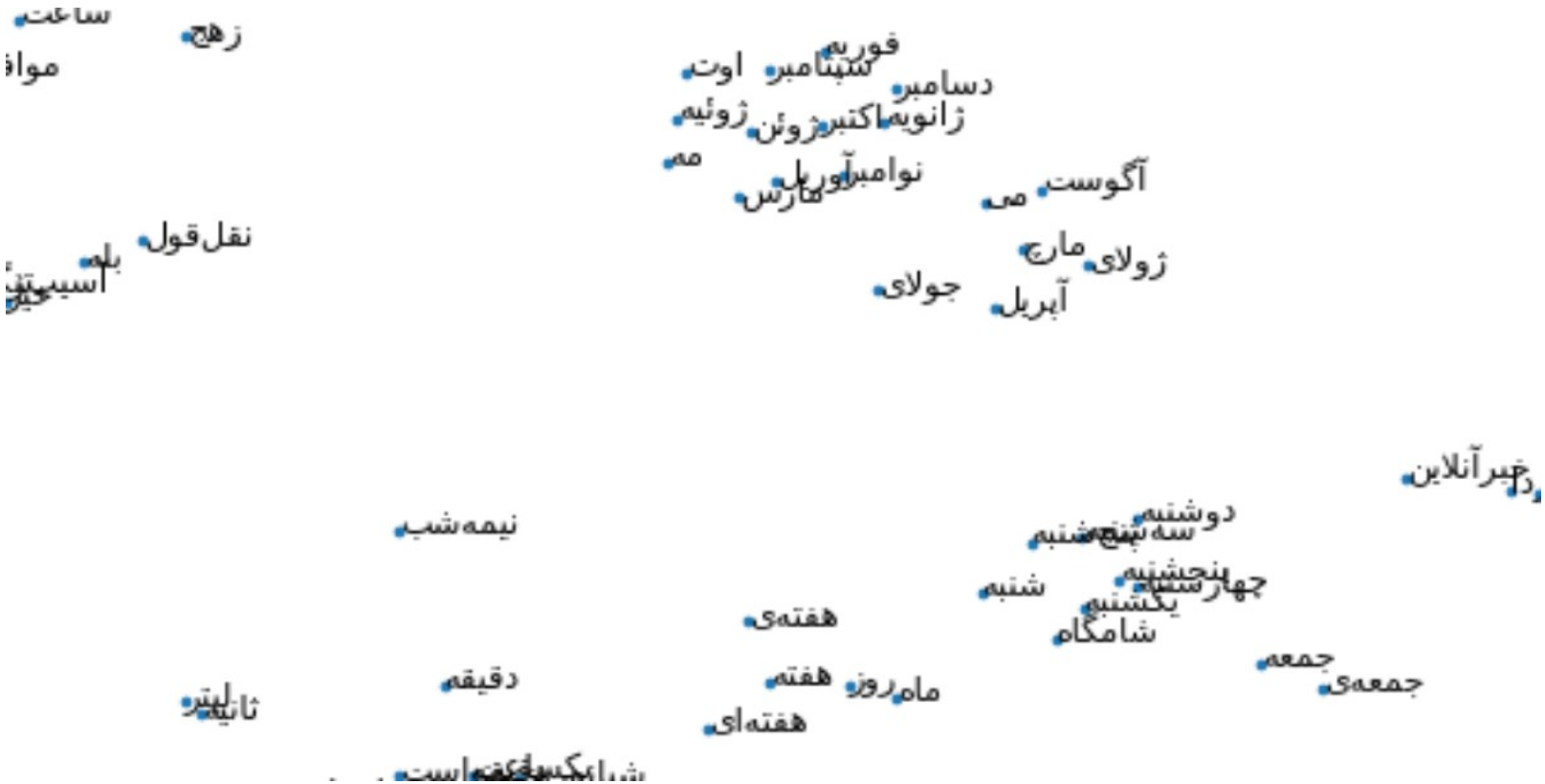

10
5
0
−5
−10
−15
−10
−5
0
5
10
15


10
5
0
−5
−10
−15
−10
−5
0
5
10
15


بدن
سیاه
بدنشان
انگشتان
انگشت
بدن
ناخن
چاه‌ها
اندام‌های
ماهیچه‌ها
سفره‌های
مخاط
منطقه
ترشحات
معده
دوازده
برنج
زعفران
چای
دوغ
صید
ماهیهای
ماهی
ترشی
شربت
نان
میگو
قهوه
اسپرسو
آجیل
کبور
بستنی
مرغ‌های
لیکور
شراب
نوشیدنی
مرغ
آبجو
چوبیه
کریج ساز
شراب‌های


جمعه
دقیقه
هفته
روز
ماه
هفته‌ای
ثانیه
لیتر
صفر
مساوی
دگردی
کبیسه
امروز
حدی
کنون
یک
دو
چند
سه
چهار
شش
هشت
پانزده
چندین
پنج
هفت
چهل
دهه
سال
درسال


جمعه
دقیقه
هفته
روز
ماه
هفته‌ای
ثانیه
لیتر
صفر
مساوی
دگردی
کبیسه
امروز
حدی
کنون
یک
دو
چند
سه
چهار
شش
هشت
پانزده
چندین
پنج
هفت
چهل
دهه
سال
درسال


ساعت
زهر
موا
فوریه
سپتامبر
اوت
دسامبر
ژانویه
اکتبر
ژوئیه
ژوئن
مه
نوامبر
آوریل
مارس
آگوست
می
نقل‌قول
بله
مارچ
ژولای
جولای
آپریل
دقیقه
ثانیه
نیمه‌شب
هفته‌ی
هفته
روز
ماه
هفته‌ای
دوشنبه
سه‌شنبه
پنج‌شنبه
چهارشنبه
یکشنبه
شنبه
شامگاه
جمعه
جمعه‌ی

یک دو : سه

دو سه : چهار

سه چهار : پنج

چهار پنج : شش

پنج شش : چهار (رتبه هفت ۲)

شش هفت : پنج (رتبه هشت ۱)

هفت هشت : پنج (نیست)

هشت نه : پنج (نیست)

```
w2v.most_similar('یکصد')
```

```
[('0.8516631126403809 ','دویست'),
 ('0.8497992753982544 ','سیصد'),
 ('0.8466461300849915 ','پانصد'),
 ('0.8063699603080075 ','چهارصد'),
 ('0.8051990866661072 ','دوهزار'),
 ('0.7979689240455627 ','یکهزار'),
 ('0.7950036525726318 ','پنجاه'),
 ('0.7919832468032837 ','صد'),
 ('0.7910781502723694 ','ششصد'),
 ('0.7847251892089844 ','هزار')]
```

یکصد دویست : سیصد

دویست سیصد : پانصد (رتبه چهارصد ۱)

سیصد چهارصد : ششصد (ربته پانصد ۱)

چهارصد پانصد : سیصد (رتبه ششصد ۲)

پانصد ششصد : سیصد (ربتبه هفتصد ۱)

ششصد هفتصد : هشتصد

هفتصد هشتصد : نهصد

هشتصد نهصد : هفتصد (نیست)

داده‌گان

- داده‌گان پرسیکا

ورزشي

داده‌گان

- داده‌گان پرسیکا
- آموزشي
  اجتماعي
  اقتصادي
  بهداشتي
  تاريخي
  سياسي
  علمي
  فرهنگي
  فقه و حقوق
  مذهبي
  ورزشي

عنوان:

وزير علوم درجمع استادان نمونه: سن بازنشستگي استادان نمونه به ۷۰ سال افزايش مي‌يابد دانشگاه بايد مهد چالشهاي گفتماني و خط دهنده و برنامه‌ريز جريانات سياسي باشد,

متن:

وزير علوم در جمع استادان نمونه كشور گفت: از استادان نمونه كشور انتظار مي‌رود كه رويكرد دانايي محوري و گفتمان علمي را به عنوان يك بحث فرهنگي در دانشگاهها توسعه و رونق بخشند. به گزارش سرويس صنفي آموزشي خبرگزاري دانشجويان ايران (ايسنا)، دكتر محمد مهدي زاهدي در اولين مجمع عمومي استادان نمونه دانشگاه‌هاي سراسر كشور كه در دانشگاه تهران برگزار شد، افزود: توصيه ما در جهت تلاش براي دانايي محوري و توسعه گفتمان علمي به معني عدم تمايل به مباحث سياسي نيست؛ بلكه برعكس، دانشگاه بايد مهد چالشهاي گفتماني باشد ولي اين امر، بدان معني نيست كه دانشگاه، ابزار دست سياسيون قرار بگيرد. وي تأكيد كرد: دانشگاه نه تنها نبايد تحت تأثير القائات سياسي قرار بگيرد؛ بلكه بايد خط دهنده و برنامه‌ريز جريانات سياسي باشد و مهمترين عنصر پياده شدن اين آرمان، دانشجويان و اعضاي هيات علمي دانشگاهها و در رأس آنها استادان نمونه هستند.

پیش پردازش

میانگین طول سند ۳.۴۴۴

پیش پردازش:

حذف کلمات اضافه:«و»، «با» ، «است» و ...

جایگزینی حروف عربی با فارسی و حذف اعراب

Embeddings	MLP	SVC	LR	۱-NN
Doc۲Vec(dbow,d۳۰۰,n۵)	۶۶	۷۴	۷۳	
Doc۲Vec(dm/m,d۱۰۰,hs,w۵)		۷۶		
Doc۲Vec(dbow,d۳۰۰,n۵) tagged	۷۳	۷۵	۷۶	
Doc۲Vec(dbow,d۳۰۰,n۲۰) tagged	۷۱	۷۰	۷۷	
Doc۲Vec(dbow,d۱۰۰,n۲۰) tagged	۷۷		۷۷	
TFIDF	۸۳	۷۶	۸۵	
word۲vec avg	۷۷		۷۷	
word۲vec avg TFIDF	۷۶		۷۸	
baseline(persica ۲۰۰۵)				۷۰

جادهی سندهای متنی

- برای جادهی جمله‌ها و پاراگراف‌ها در [LM14] ارائه شده.

جادهی سندهای متنی

- برای جادهی جمله‌ها و پاراگراف‌ها در [LM14] ارائه شده.
- هدف پیدا کردن D بردار جادهی پاراگراف‌هاست و W بردار جادهی کلمه‌هاست

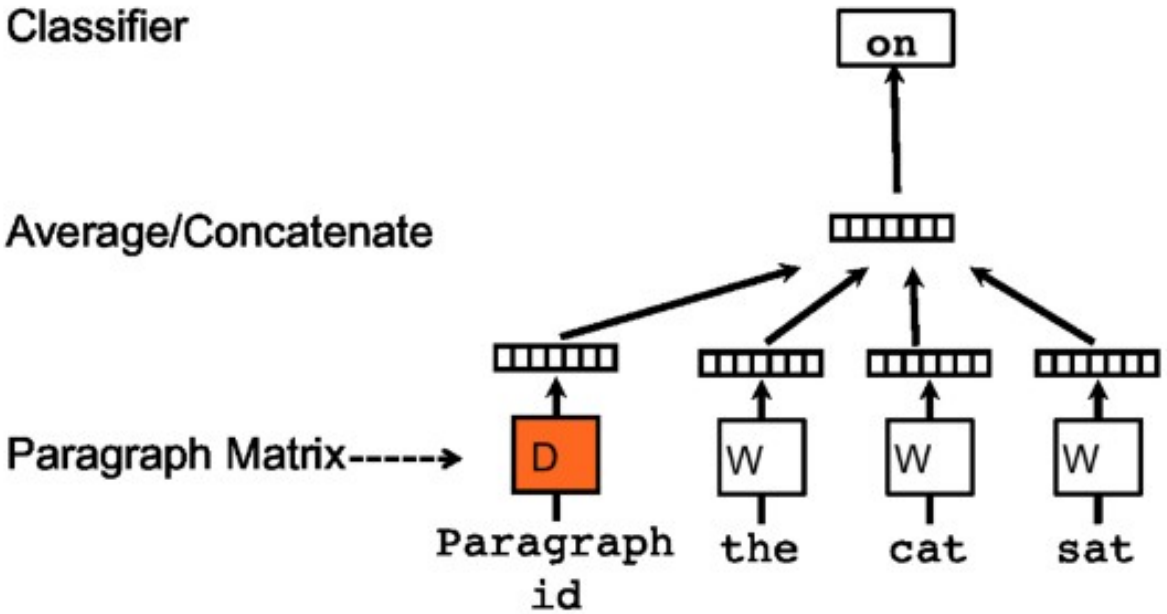
جادهی سندهای متنی

- برای جادهی جمله‌ها و پاراگراف‌ها در [LM14] ارائه شده.
- هدف پیدا کردن D بردار جادهی پاراگراف‌هاست و W بردار جادهی کلمه‌هاست
- دو مدل مانند CBOW و Skipgram ارائه شده

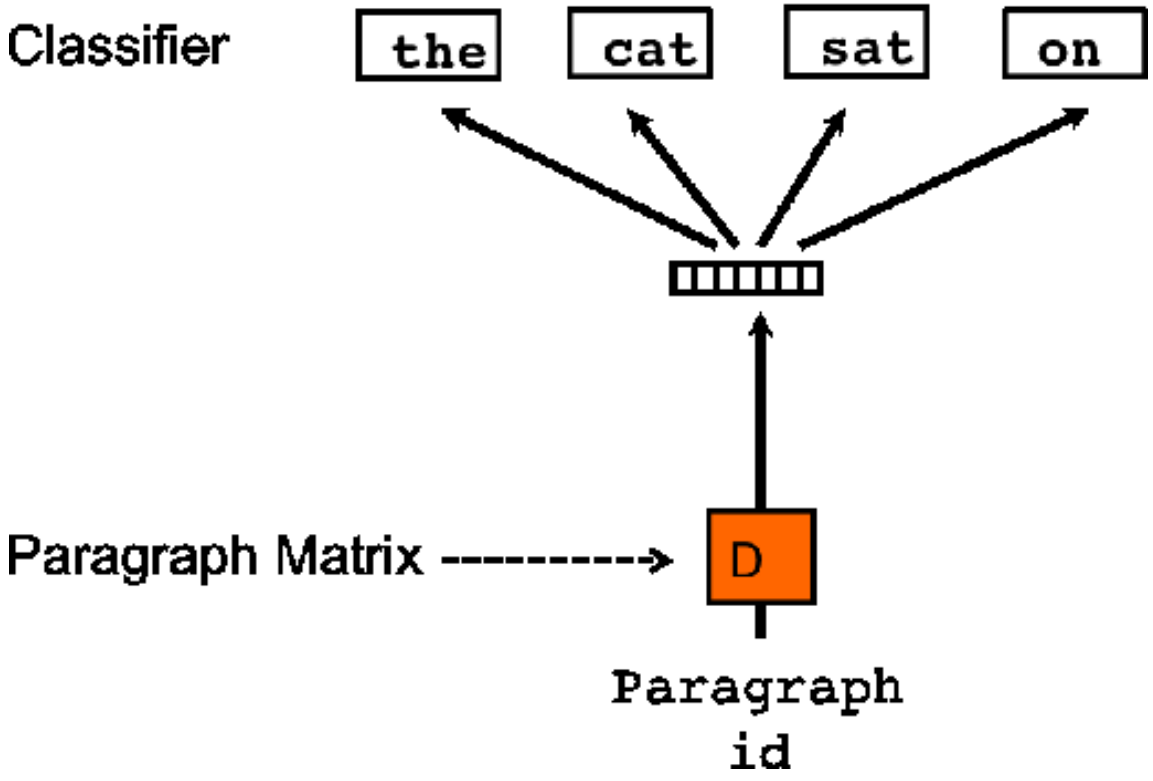
جادهی سندهای متنی

- برای جادهی جمله‌ها و پاراگراف‌ها در [LM14] ارائه شده.
- هدف پیدا کردن D بردار جادهی پاراگراف‌هاست و W بردار جادهی کلمه‌هاست
- دو مدل مانند CBOW و Skipgram ارائه شده
- روش دیگر مانند skipgram است

Paragraph Vector Distributed Memory مدل



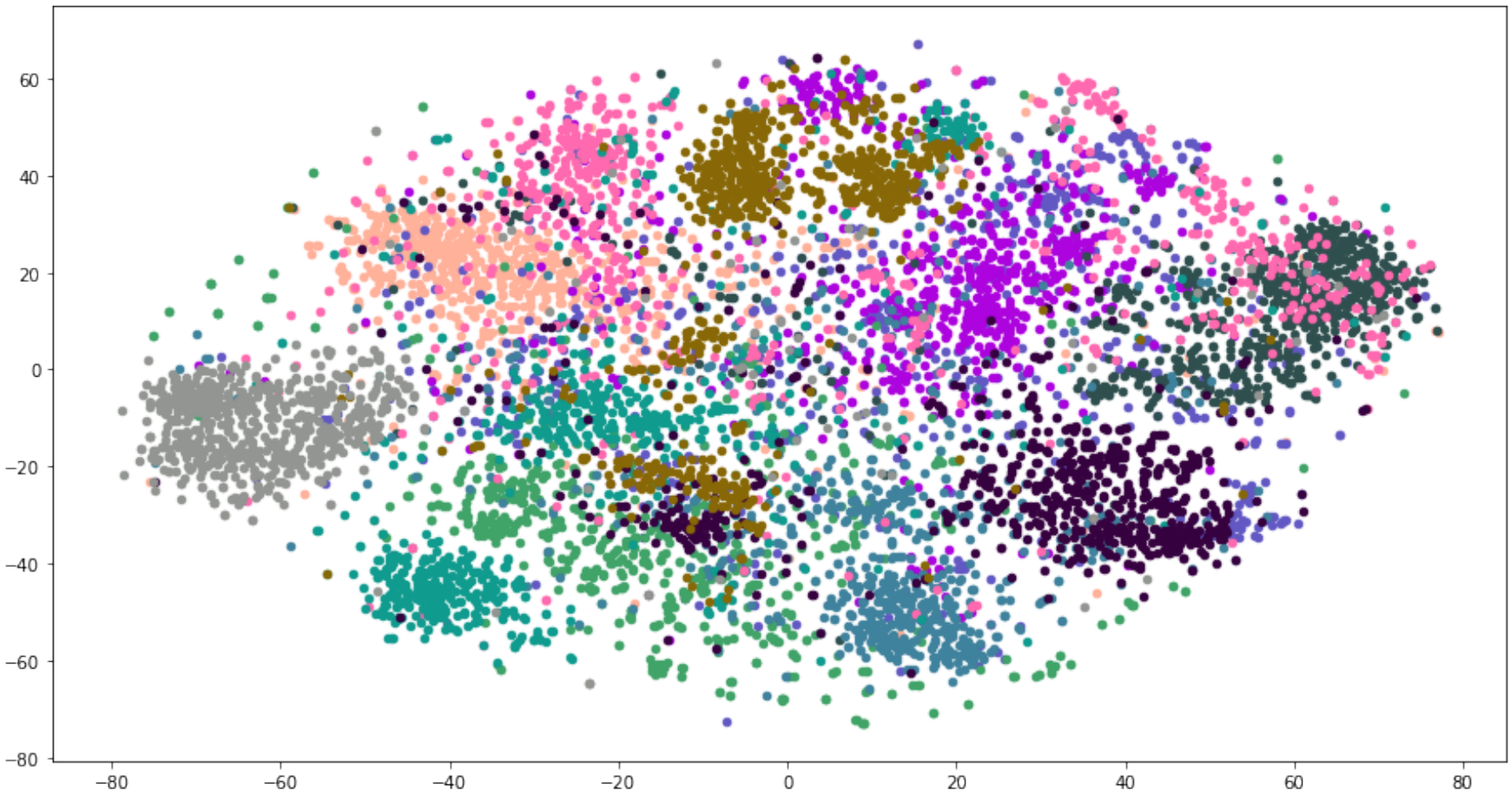
Paragraph Vector Distributed Bag of Words مدل



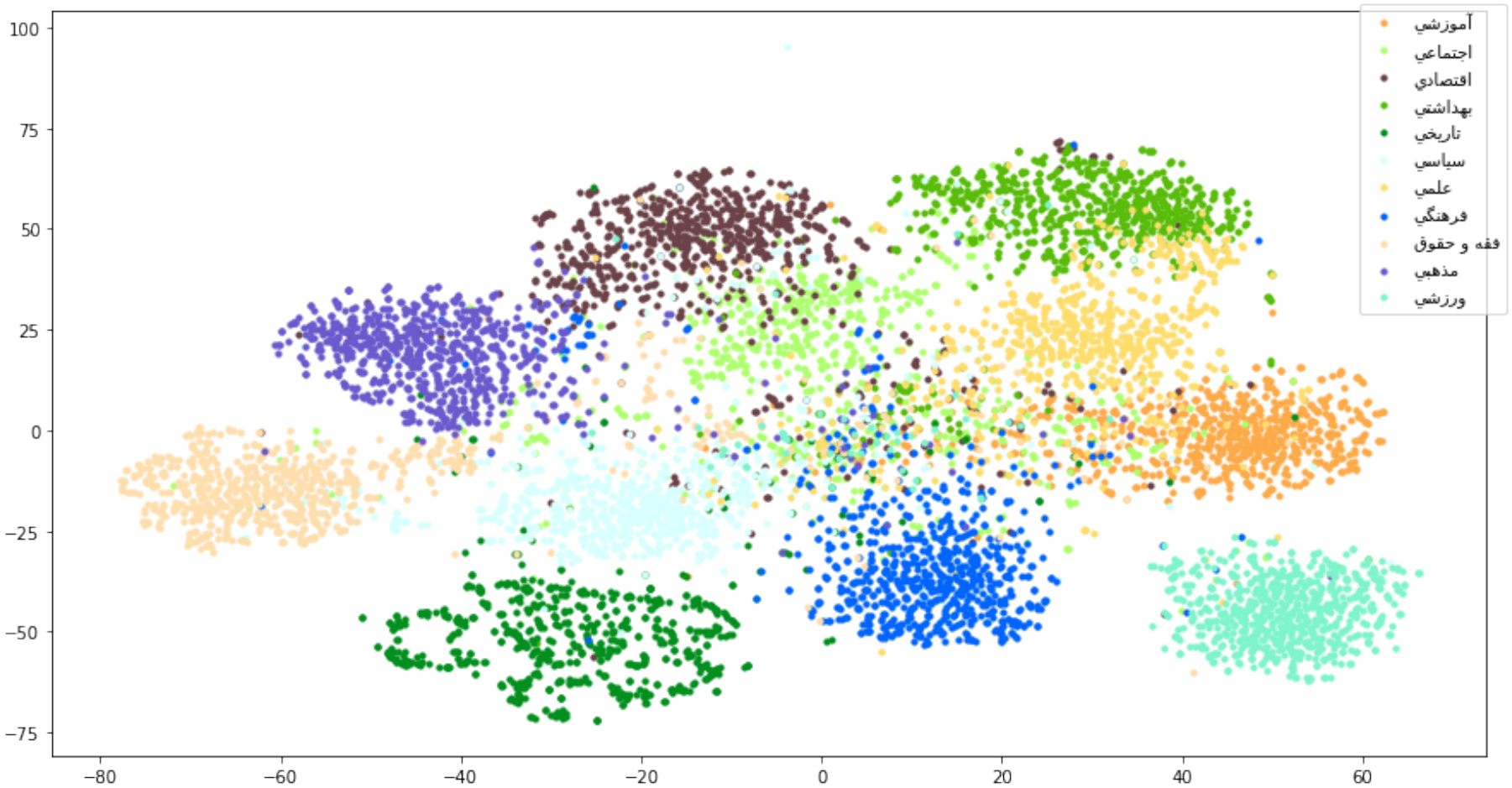
کارایی مدل doc2vec

MNB-uni (Wang & Manning, 2012)	16.45%
MNB-bi (Wang & Manning, 2012)	13.41%
SVM-uni (Wang & Manning, 2012)	13.05%
SVM-bi (Wang & Manning, 2012)	10.84%
NBSVM-uni (Wang & Manning, 2012)	11.71%
NBSVM-bi (Wang & Manning, 2012)	8.78%
Paragraph Vector	**7.42%**

نمایش تگ خورده



نمایش تگ نخورده



نمایش تگ مدل سبد کلمات