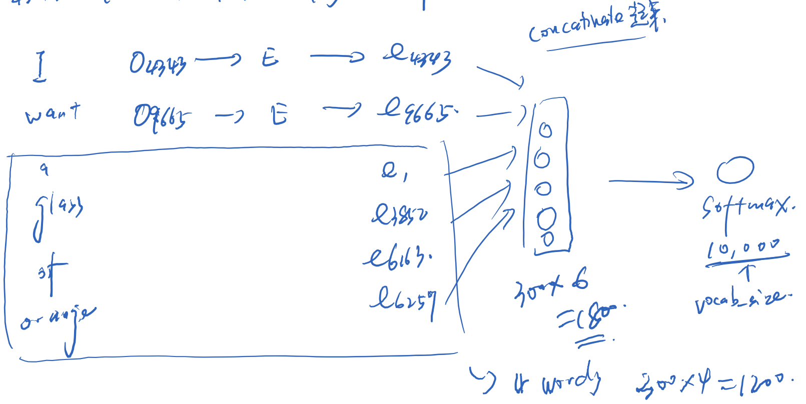
Learning word embeddings: word2vec & GloVe

Friday, June 8, 2018 1:58 PM

Some more complicated algorithms

I want a glass of orange ↓ juice.

4343 9665 1 3852 6163 6257



or only look at the previous 4 words

use a fixed history ~~ccc~~ ~~en~~ enables you to work with arbitraily long sentences.

This will do a descent job.

other context / target pairs.

I want [a glass of orange] juice [to go along with] my careal

4 words on left and right. a glass of orange — to go with.

☆ [Last 1 word] nearby one word.

↳ skip gram model

simplier ~~a~~ algorithm — just one word.

if your main goal is just word embedding, one word is enough!

Word2Vec Algorithm

Thomas Mikolov. ~~et al.~~ Kai Chen, Greg Corrado, Jeff Dean

I want a glass of orange juice to go along with my cereal.

skip gram

randomly pick a context word →

Context	Target
orange	juice
orange	glass
orange	my

← randomly pick a word in some window as the target

The goal is good word embedding.

O = one-hot encoding

Model. vocab size = 10,000 k.

Context c("orange") → Target t("juice").

$O_c \to E \to e_c \to \bigcirc \xrightarrow{} \hat{y}$

softmax

Softmax $$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10000} e^{\theta_j^T e_c}}$$

θ_t parameter associated with output t.

가명? identifiable?

식별가능.

$$y = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} \longleftarrow 4834$$

$$\sum_{j=1}^{10000} e^{\theta_j^T e_c}$$

对吗? identifiable?

learn 的是啥.

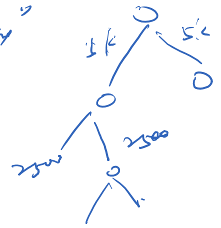
$$y = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow 4834$$

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{10000} y_i \log \hat{y}_i$$

The Skip gram Model

Problem: Computationally taxing. — Sum over 10,000 in softmax

solution: "hierarchical softmax"



← not ~~red~~ really symmetric ~~symmetric~~ balanced tree in applications.

P(c).

read the orignal paper

2 versions

alternative vesin

1) surrounding words → middle word

Negative Sampling → computationally cheap ~~method~~. ~~version~~ method ~~of~~ for skip gram

Mikolov, Sutskever, Chen, Corrado. Dean.

I want a glass of orgauge juice to go along with my cereal.

Context	Target	Target?
		y
orange	juice	1
	king	0
	book	0
	the	0
	of	0

x, k

Which k to choose?

k = 5–20 for small data ~~set~~

k = 2–5 for larger datasets.

① pick a context word, pick a target word. — positive example

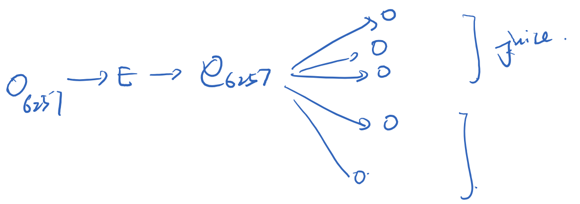
② for k time pick random words in the dictionary. label as 0.

It is ok if one of the randomly picked word is actually in the window.

$$\text{Softmax}: \quad p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}} \leftarrow 10,000 \text{ way softmax}$$

$$p(y=1|c,t) = \sigma(\theta_t^T e_c).$$

sigmoid. — with negative sampling



10,000 binary classification problem

~~use~~ Computational power!

Selecting negative example?

Selecting negative example?

- proportional to word freq → end up with lots of "a" "the"
- equal prob? Not good either

use something in between. [illegible]

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=1}^{10000} f(w_j)^{3/4}}$$

$\boxed{\text{GloVe word vector}}$ — not used as much as
- wordevec
- skip gram

Global Vector for word representation.

I want a glass of orange juice to go along with my cereal.

c, t.

$\underset{c \quad t}{X_{ij}} = \#$ times $\underset{t}{i}$ appears in context of $\underset{c}{j}$

For GloVe define $X_{ij} = X_{ji}$

X_{ij} a count of how much i and j appear with each other.

$\boxed{\text{Minimize}}$ $\sum_{i=1}^{10000} \sum_{j=1}^{10000} f(X_{ij}) (\underset{t}{\theta_i^T} \underset{c}{e_j} + \overset{t}{b_i} + \overset{c}{b_j'} - \log X_{ij})$

$\theta_t^T e_c$

Weighting term $f(X_{ij}) = 0$ if $x_{ij} = 0$ "0 log 0 = 0" 防止计算log0.

frequent word. this, is, a.

infrequent word. durian

weight - not give frequent words too much weight

not give infrequent words too little weight.

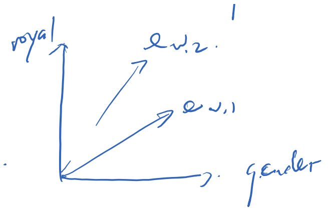
θ_i e_j are symmetric

$$e_w^{(final)} = \frac{e_w + \theta_w}{2}.$$

About the featurization: You cannot guarantee that each dimension of featurization is interpretable.

royal $e_{w,2}$.

is interpretable.



You can't guarantee the features are human-interpretable.