Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec

Christopher Moody

Presenter : 이종진

Seoul National University

ga0408@snu.ac.kr

January 07, 2020

lda2vec

- Dense vector representations + interpretable representaions over documents(topics)
- Dense vector representations of words, documents and topics are in same embedding spaces.
- Document vectors are expressed mixture of topic vectors ($d_1 = 0.9t_1 + 0.1t_2$)

word2vec(skip gram)

- Using pivotal word, predict words in fixed window size.
- n : number of words, d : embedding dimension.
- Two weight matrix $W_{n\times d}$, $U_{d\times n}$
- Rows of the W are dense representations of words.

word2vec(skip gram)

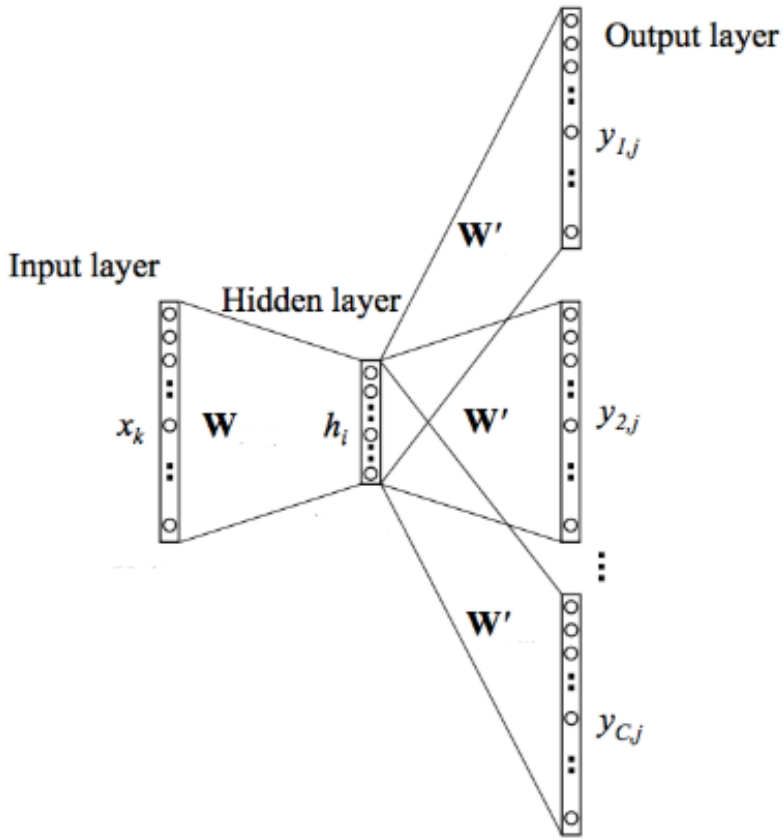


Figure: word2vec

word2vec(skip gram)

- Two weight matrix $W_{n\times d}$, $U_{d\times n}$
- The training objective of the Skip-gram model

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{-c\leq t\leq c, t\neq 0}\log p(w_{i+t}|w_i)$$

- $P(w_j|w_i)$ is defined as

$$P(w_j|w_i) = \frac{\exp(\vec{w}_i \cdot \vec{u}_j)}{\sum_{j=1}^{n}\exp(\vec{w}_i \cdot \vec{u}_j)}$$

,$\vec{w}_i, \vec{u}_j$: ith row of W, jth column of U.

word2vec(skip gram with negative sampling)

- When number of word is large, skipgram formulation is impractical.
- Sample N(5∼20) number of words which are not in window size.
- The word w_i is drawing with probability,

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{i:w_i \notin \text{window}} f(w_i)^{3/4}}$$

,$f(w_i)$ = frequency of word w_i / total frequency of words

- The training objective function :

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{-c \leq t \leq c, t \neq 0} \mathcal{L}_{i.i+t}$$

- $\mathcal{L}_{ij}$ is defined as

$$\log \sigma(\vec{w}_i \cdot \vec{u}_j) + \sum_{l=1}^{N} \log \sigma(-\vec{w}_i \cdot \vec{u}_{neg,l})$$

Word vectors

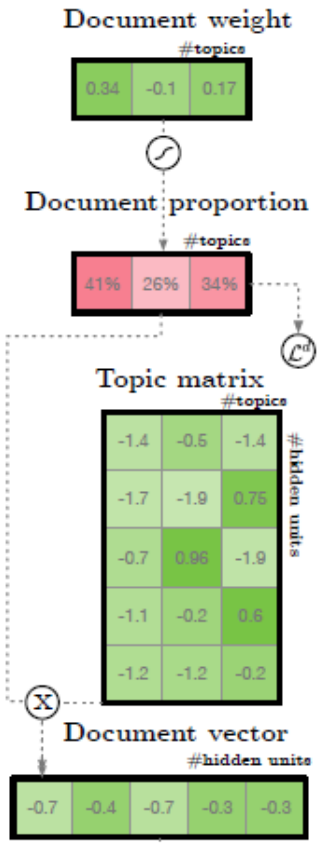
- Similary to word2vec, using pivot word predict target word in window size.
- $\vec{w}_{ij}$ indicates certain row of the matrix W which corresponding to ith words in jth document.
- $\vec{u}_{ij}$ indicates certain column of the matrix U which corresponding to ith words in jth document.
- It uses a context vector, which is a sum of pivotal word vectors and document vector ($\vec{c}_{ij} = \vec{w}_{ij} + \vec{d}_j$)

Document vectors

- Document vectors are expressed mixture of topic vectors
- K : number of topics, J : number of documents
- Each documents has document weight $(h_{j1}, \ldots, h_{jK})$
- Through softmax transform, $(p_{j1}, \ldots, p_{jn}), \sum_{k=1}^{K} p_{jk} = 1$
- Document vector $(\vec{d_j})$:

$$\vec{d_j} = p_{j0}\vec{t_0} + p_{j1}\vec{t_1} + \ldots + p_{jk}\vec{t_k} + \ldots + p_{jK}\vec{t_K}$$

Topic vectors



- Columns of embedding matrix are dense representation of topics.
- Similarity between word and topic can be calculated.

SGNS loss

- The total loss

$$\mathcal{L} = \mathcal{L}^d + \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{-c \leq t \leq c, t \neq i} \mathcal{L}_{ijt}^{neg}$$

- The sum of the Skipgram Negative Sampling Loss(SGNS) $\mathcal{L}_{ijt}^{neg}$

$$\mathcal{L}_{ijt}^{neg} = \log \sigma(\vec{c}_{ij} \cdot \vec{u}_{tj}) + \sum_{l=1}^{N} \log \sigma(-\vec{c}_{ij} \cdot \vec{u}_{neg,l})$$

  - n_j : number of words in jth document
  - c is window size.
  - N : Negative sample size.
  - $\vec{c}_{ij}$: a context vector, sum of pivot word vectors and document vector ($\vec{c}_{ij} = \vec{w}_{ij} + \vec{d}_j$)

Dirichlet-likelihood loss

- Dirichlet-likelihood $\mathcal{L}^d$:

$$\mathcal{L}^d = \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} (\alpha - 1) \log p_{jk}$$

- This simple likelihood encourages
  - Document has sparse $p_{jk}, k = 1, \ldots, K$ when $\alpha < 1$

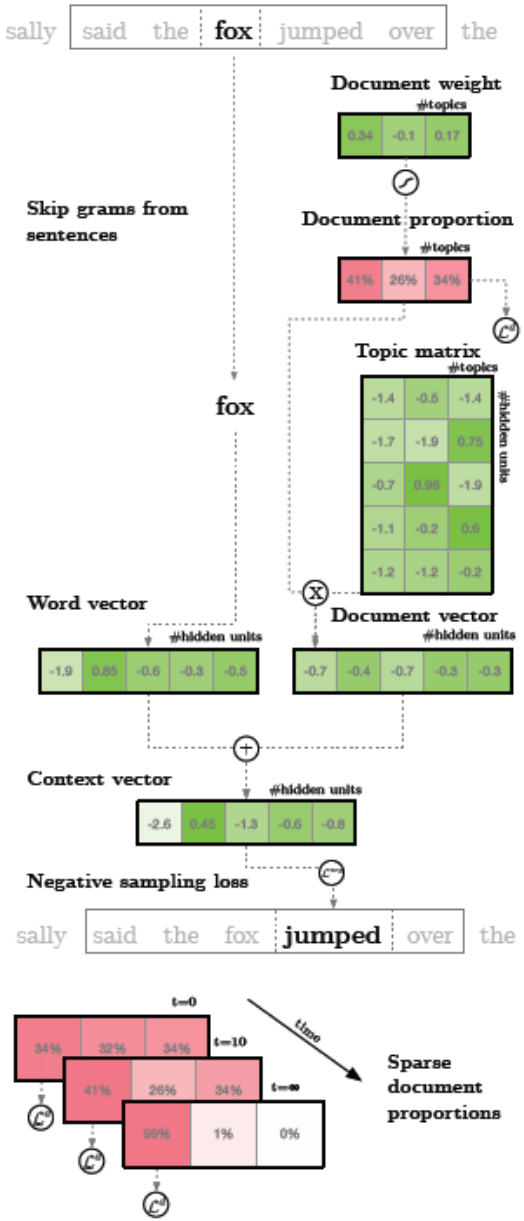


Figure: architecture

Effect of sum of word vector and document vector

- For word 'Germany', neighboring words are predicted similar such as 'France', 'Spain', and 'Austria'.
- if the document about airlines, neighboring words could be 'Lufthansa', 'Condor Flugdienst', and 'Aero Lloyd'
- This models can consider document-wide relationships, while still leveraging local inter-word relationships.

Experiments

- Twenty Newsgroups

Topic Label	"Space"	"Encryption"	"X Windows"	"Middle East"
Top tokens	astronomical Astronomy satellite planetary telescope	encryption wiretap encrypt escrow Clipper	mydisplay xlib window cursor pixmap	Armenian Lebanese Muslim Turk sy
Topic Coherence	0.712	0.675	0.472	0.615

- The most similar words are listed
- Closely related newsgroup, 'sci.space', 'sci.crypt', 'comp.windows.x' and 'talk.politics.mideast'

Experiments

- Hacker News Comments corpus
- Social content-voting website and community whose focus is largely on technology and entrepreneurship.

"Housing Issues"	"Internet Portals"	"Bitcoin"	"Compensation"	"Gadget Hardware"
more housing	DDG.	btc	current salary	the Surface Pro
basic income	Bing	bitcoins	more equity	HDMI
new housing	Google+	Mt. Gox	vesting	glossy screens
house prices	DDG	MtGox	equity	Mac Pro
short-term rentals	iGoogle	Gox	vesting schedule	Thunderbolt

Figure: The inferred topic labed is shonw in the first row

Experiments

Artificial sweeteners	Black holes	Comic Sans	Functional Programming	San Francisco
glucose	particles	typeface	FP	New York
fructose	consciousness	Arial	Haskell	Palo Alto
HFCS	galaxies	Helvetica	OOP	NYC
sugars	quantum mechanics	Times New Roman	functional languages	New York City
sugar	universe	font	monads	SF
Soylent	dark matter	new logo	Lisp	Mountain View
paleo diet	Big Bang	Anonymous Pro	Clojure	Seattle
diet	planets	Baskerville	category theory	Los Angeles
carbohydrates	entanglement	serif font	OO	Boston

Figure: The most similar words for given tokens

Query	Result
California + technology	Silicon Valley
digital + currency	Bitcoin
Javascript - browser + server	Node.js
Mark Zuckerberg - Facebook + Amazon	Jeff Bezos
NLP - text + image	computer vision
Snowden - United States + Sweden	Assange
Surface Pro - Microsoft + Amazon	Kindle

Figure: Examples of linear relationships