Регуляризация тематических моделей для векторных представлений слов

Выпускная квалификационная работа

Попов Артём Сергеевич
Научный руководитель: Воронцов Константин Вячеславович

МГУ имени М.В. Ломоносова, ВМК

7 июня 2017 г.

Постановка задачи

Дано: текстовая коллекция D, словарь W

Найти: для каждого слова $w \in W$ векторное представление (word embedding) $v_w \in \mathbb{R}^m$, $m \ll |W|$

Критерии качества:

- Семантически/синтаксически близким словам должны соответствовать близкие вектора (задача близости слов)
- В векторном пространстве можно производить интерпретируемые алгебраические операции (задача аналогий слов)
- Компоненты векторов должны быть интерпретируемы
- Компоненты векторов должны быть разреженными
- Векторные представления должны легко переноситься с отдельных слов на документы и другие модальности

Модели векторных представлений слов (word embeddings)

Гипотеза дистрибутивности (Harris, 1954): слова, встречающиеся в схожих контекстах, имеют схожее значение.

Известные модели, основанные на гипотезе:

- Нейросетевые подходы (семейство word2vec — cbow, skip-gram . . .)
- Матричные разложения (Glove, SVD sPPMI)

Преимущества и недостатки моделей:

+ Хорошее решение задачи близости слов
+ Хорошее решение задачи аналогий слов
– Компоненты неинтерпретируемы
– Компоненты неразрежены

Вероятностное тематическое моделирование

Мультимодальная тематическая модель ARTM:

T — набор скрытых переменных (тем)

M — множество модальностей, у каждой словарь W^m, $W = \bigcup_m W^m$

n_{dw} — сколько раз термин w встретился в документе d

$$\begin{cases} L(\Phi,\Theta) = \sum\limits_{m\in M} \alpha_m \sum\limits_{d\in D} \sum\limits_{w\in W^m} n_{dw} \ln p(w|d) + \sum\limits_i \tau_i R_i(\Phi,\Theta) \to \max\limits_{\Phi,\Theta} \\ p(w|d) = \sum\limits_{t\in T} p(w|t)p(t|d) = \sum\limits_t \phi_{wt}\theta_{td} \\ \phi_{wt} \geq 0, \quad \sum\limits_{w\in W^m} \phi_{wt} = 1, \quad \theta_{td} \geq 0, \quad \sum\limits_{t\in T} \theta_{td} = 1 \end{cases}$$

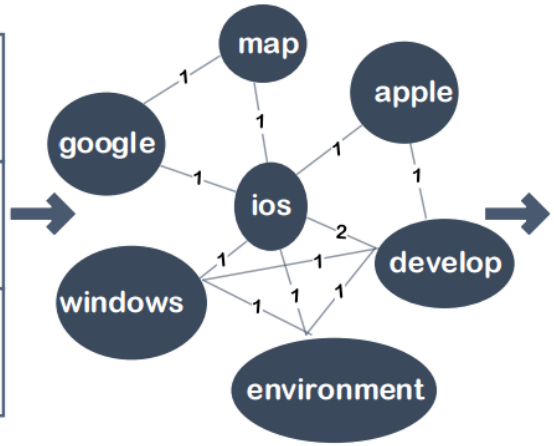
Преимущества и недостатки для построения word embeddings:

- \- Плохое решение задачи близости слов
- \- Плохое решение задачи аналогий слов
- \+ Компоненты интерпретируемы
- \+ Компоненты разрежены

Создание псевдо-документов

Псевдо-коллекция[1]: пусть n_{wc} — совстречаемость слов w и c, для каждого слова $w \in W$ создадим псевдо-документ d': $n_{d'c} = n_{wc}$

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Функционал изменённой одномодальной модели:

$$\begin{cases} L(\Phi,\Theta) = \sum\limits_{c\in W}\sum\limits_{w\in W} n_{cw} \ln \sum\limits_{t\in T} \phi_{wt}\theta_{tc} + \sum\limits_{i} \tau_i R_i(\Phi,\Theta) \to \max\limits_{\Phi,\Theta} \\ \phi_{wt} \geq 0, \quad \sum\limits_{w\in W} \phi_{wt} = 1, \quad \theta_{tc} \geq 0, \quad \sum\limits_{t\in T} \theta_{tc} = 1 \end{cases} \tag{1}$$

[1] *Zuo Y. et al.* Word Network Topic Modeling, 2016

Мультимодальная псевдо-коллекция

Естественное расширения для мультимодальной коллекции: учитывать слова разных модальностей, находящиеся в одном документе

doc1	Google map for iOS	Jobs
doc2	iOS is developed by Apple	Jobs
doc3	iOS develop enviroment for Windows	Gates

word	1 modality	2 modality
google	map ios	jobs
map	google ios	jobs
ios	google map develop apple develop enviroment windows	jobs jobs gates
develop	apple ios ios windows enviroment	jobs gates
apple	ios develop	jobs
enviroment	ios develop windows	gates
windows	ios develop enviroment	gates
jobs	google map ios ios develop apple	-
gates	ios develop windows enviroment	-

Не обязательно составлять псевдо-документы для токенов не основных модальностей.

Сравнение моделей на задаче близости слов

Эксперименты проводились на коллекции англоязычной Википедии.

Задача близости

Вход: список троек w_1, w_2 — слова, x — близость между ними
Измерим близость между векторами модели по всем парам
Выход: корреляция Спирмена между двумя списками близостей

Качество решения на разных датасетах:

	близость	WordSim Sim.	WordSim Rel.	WordSim	Bruni MEN	Rad. Turk
LDA	$-H(u,v)$	0.553	0.478	0.493	0.583	0.51
SVD (PPMI)	$\cos(u,v)$	0.711	0.648	0.672	0.236	0.616
SGNS	$\cos(u,v)$	**0.752**	0.633	0.665	0.**744**	**0.661**
ARTM (offline EM)	$\langle u,v\rangle$	0.71	0.62	0.65	0.67	0.59
ARTM (hybrid EM)	$\langle u,v\rangle$	0.723	**0.675**	**0.682**	0.672	0.642

Вывод: В решении задачи близости ARTM сопоставимо с SGNS

Особенности инициализации EM-алгоритма

Задача тематического моделирования, как матричное разложение:

$$F = \{n_{wc}\}_{w \in W, c \in W} \approx \Phi\Theta$$

Если входные данные симметричные можно производить разложение:

$$F = \{n_{wc}\}_{w \in W, c \in W} \approx \Phi\Phi^{bayes} \qquad \Phi^{bayes}_{tw} \propto \Phi_{wt} p(t)$$

Теорема

Пусть $F = F^T$, Φ_0, Φ_0^{bayes} — начальные инициализации матриц Φ и Θ. Тогда, матрицы Φ и Θ, полученные при решении задачи (1) с $R \equiv 0$ оффлайновым EM-алгоритмом, связаны соотношением $\Theta = \Phi^{bayes}$

Качество на задаче близости эквиваленто несимметричному случаю:

	близость	WordSim Sim.	WordSim Rel.	WordSim	Bruni MEN	Rad. Turk
ARTM (offline EM)	$\langle u, v \rangle$	0.71	0.62	0.65	0.67	0.59
ARTM (symmetric)	$\langle u, v \rangle$	0.71	0.62	0.64	0.66	0.59

Сравнение моделей на задаче аналогий слов

Задача аналогий

Вход: список четвёрок слов w_1, w_2, w_3, w_4 : $v_{w_3} - v_{w_1} + v_{w_2} = v_{w_4}$
Проверим свойство $v_{w_3} - v_{w_1} + v_{w_2} \approx v_{w_4}$ для векторов модели
Выход: Доля правильно найденных слов

Качество решения на разных датасетах:

	google	msr
LDA	0.159	0.146
SVD (PPMI)	0.340	0.135
SGNS	**0.688**	**0.375**
ARTM (offline EM)	0.385	0.174
ARTM (hybrid EM)	0.377	0.207

Вывод: Качество решения задачи аналогий хуже чем SGNS, но лучше традиционных тематических моделей

Примеры векторных операций для ARTM

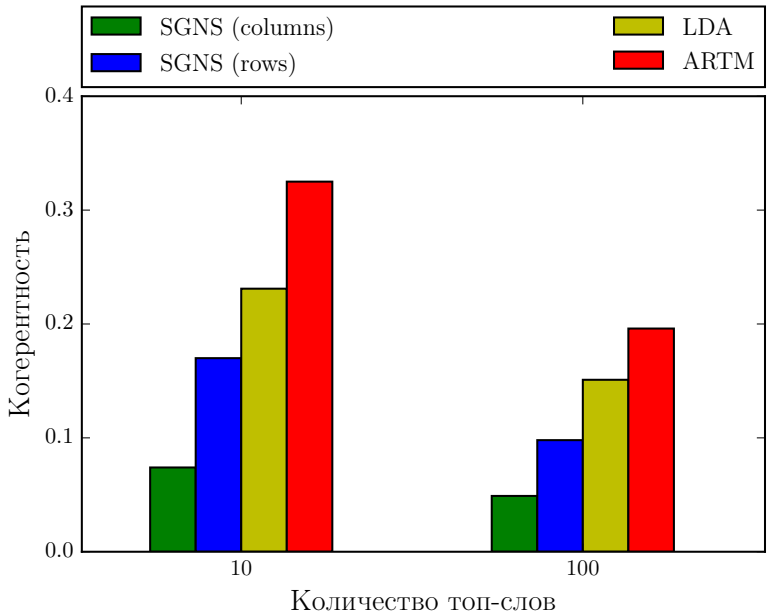
Операция	Результат ARTM	Результат SGNS
king + boy – girl	queen, princess, lord, prince	queen, princess, regnant, kings
moscow + spain – russia	madrid, barcelona, aires, buenos	madrid, barcelona, valladolid, malaga
india + ruble – russia	rupee, birbhum, pradesh, madhaya	rupee, rupiah, devalued, debased
better + bad – good	really, something, thing, nothing	worse, easier, prettier, funnier
cars + computer – car	computers, software, servers, implementations	computers, software, hardware, microcomputers

Интерпретируемость компонент

В статье[2] показана связь когерентности с интерпретируемостью

w_i — i-ый токен в порядке убывания значений координаты t векторов

$$Coherence(t) = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} PPMI(w_i, w_j)$$



SGNS		ARTM	
rana	membrane	art	moscow
rashid	oral	painting	dynamo
malek	dental	museum	sofia
aziz	absorbed	painters	spartak
khalid	gre	gallery	levski

Вывод: интерпретируемость компонент в ARTM лучше SGNS и LDA

[2] *Newman D. et al.* Automatic Evaluation of Topic Coherence, 2010

Разреженность компонент

Регуляризаторы разреживания Φ и/или Θ (α_t, β_w — равномерные)

$$R_i(\Phi,\Theta) = \sum_t KL(\beta_w \parallel \phi_{wt}) \qquad R_i(\Phi,\Theta) = \sum_t KL(\alpha_t \parallel \theta_{tw})$$

	SGNS	ARTM	ARTM (+разр.)
Разреженность	0	0.8	0.93

Сравнение близости:

	WordSim Sim.	WordSim Rel.	WordSim	Bruni MEN	Radinsky M. Turk
ARTM	0.723	**0.675**	**0.682**	0.672	**0.642**
ARTM (+разр.)	**0.728**	0.672	0.68	**0.675**	0.635

Вывод: используя регуляризацию, можно получить сильно разреженные вектора, эквивалентные исходным на задаче близости

Мультимодальные коллекции

Коллекция Lenta.ru: текст + время + категория + подкатегория

Качество на задаче близости по разным датасетам ($cos(u, w)$ / $\langle u, w \rangle$):

	WS-sim	WS-rel	Miller	Rubenstein
SGNS	0.63	0.53	0.377	0.415
ARTM	0.612/0.649	0.54/0.565	0.648/0.605	**0.63**/0.594
Multi-ARTM (с дополнительными псевдо-документами)	0.646/0.677	0.537/0.576	0.668/0.612	0.618/0.585
Multi-ARTM (без дополнительных псевдо-документов)	0.646/**0.682**	0.55/**0.58**	**0.675**/0.607	0.617/0.584

Вывод: на небольших коллекциях, ARTM превосходит SGNS.
С помощью использования модальностей можно улучшать качество предложенного подхода.

Векторные представления для временных промежутков

Использование дополнительных псевдо-документов необходимо для получения интерпретируемых представлений токенов модальностей:

премьера star wars 2015-12-18	церемония оскар 2016-02-29	9 мая 2015-05-09
джедай	статуэтка	великий
ситх	кинонаграда	годовщина
фетт	номинироваться	фотопортрет
энакин	кинопремия	нормандия
чубакка	линклейтер	парад
киносага	оскар	демонстрация
хэмилл	бёрдмен	шестие
кэрри	удостоиться	владимир
приквел	award	празднование
соло	критик	концентрационный
пробуждение	отрочество	освенцим

На защиту выносятся результаты:

- Предложен способ получения интерпретируемых и разреженных векторных представлений слов, сопоставимых с моделью SGNS по качеству решения задачи семантической близости слов
- Предложен способ улучшения предложенного подхода за счёт мультимодальности и способ получения векторных представлений для токенов дополнительных модальностей
- Доказана теорема о сводимости задачи поиска разложения симметричной матрицы в виде ФФ^{bayes} к общей задаче при специальной инициализации EM-алгоритма