doc2vec在开放题评分模型中的应用研究

驰声研发

方敏

要点

§ doc2vec特点及对开放题研究的意义
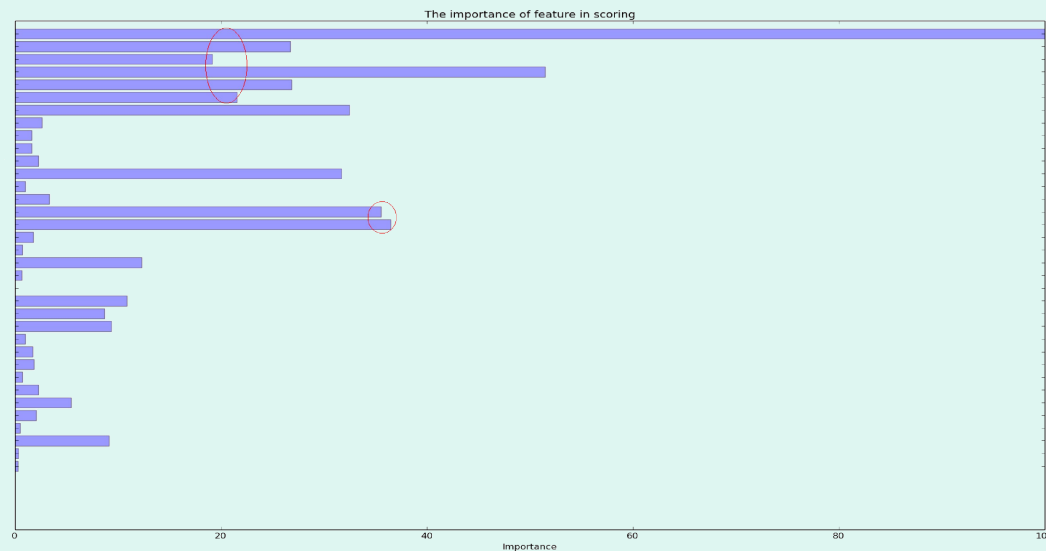
§ doc2vec简介

§ 应用思路及方案

§ 相关实验结果

§ doc2vec特点及对开放题研究的意义

- 开放题
  —题型范围：口头作文oesy、故事复述prt1、看图作文 pict
- 开放题的评分特征中关于文本的特征

The importance of feature in scoring

Importance

- doc2vec特点
    - 用连续的分布式向量表示文本，向量长度固定
    - 包含词序、上下文关系等信息
    - 无监督训练
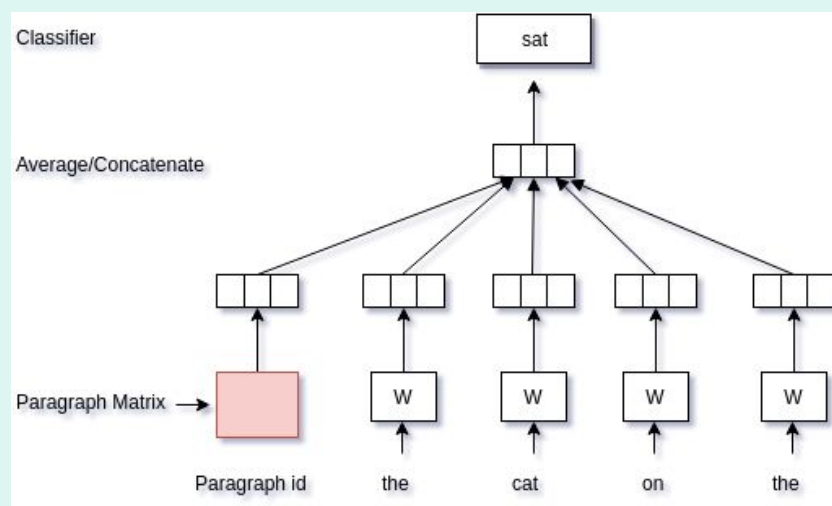
- 对开放题的研究意义
    - 强化评分特征在文本词序、语义上的信息表达
    - 构建语义特征

§ doc2vec简介

Tomas Mikolov

- Tomas Mikolov
  2012.6，Microsoft，Research Intern
  2012.10，Google，Research Scientist
      Google Brain，参与了 word2vec 项目的开发
  2014.5，Facebook，Research Scientist
      Facebook 个人页面上写到他的长期研究目标是"开发能够使用自然语言进行学习和与人类交流的智能机器"。
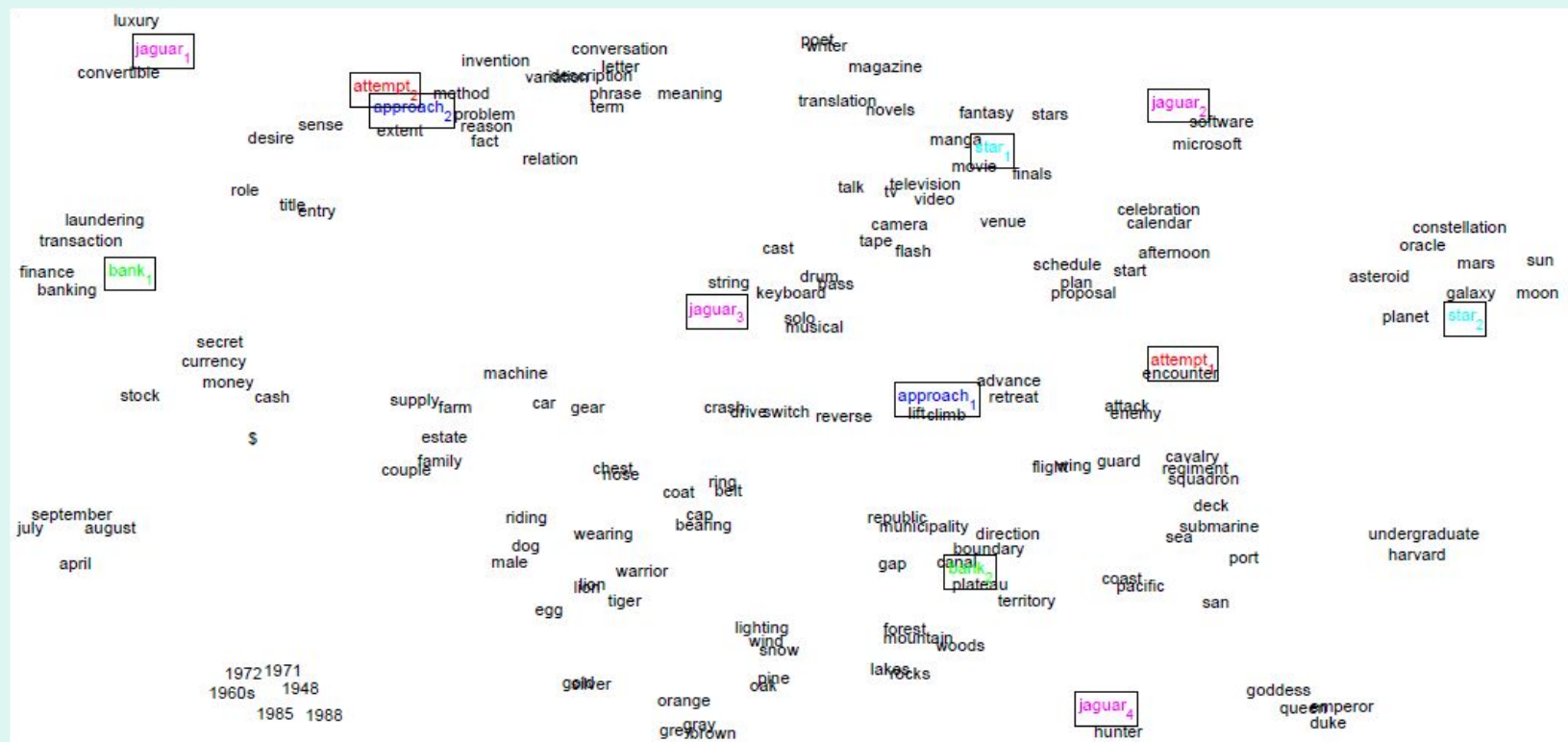
- 对比模型bag-of-word

- king－man＋woman ≈queen

$$w_1, w_2, w_3, ..., w_T$$

$$\frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, ..., w_{t+k})$$

$$p(w_t | w_{t-k}, ..., w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$$

y = b + Uh(w_{t-k},...w_{t+k};W,D)

luxury
jaguar₁
convertible

conversation
poet
writer
magazine

invention
letter
description
phrase meaning
variation
method
attempt₂
approach₂
problem
reason
term
translation novels
fantasy stars
manga
star₁
movie finals
microsoft

sense
desire
extent
fact
relation

role
title entry
talk
television
tv video
celebration
calendar

laundering
transaction
camera
venue
constellation
oracle
mars sun
finance bank₁
cast
tape flash
schedule
plan
start
afternoon
asteroid
galaxy moon
banking
string
drum
bass
keyboard
proposal
planet star₂

secret
currency
money
jaguar₃
solo
musical
attempt₄
encounter

stock cash
machine
advance
retreat
attack
enemy

$
supply farm
car gear
crash
drive switch reverse
approach₁
lift climb
flight king guard
cavalry
regiment
squadron

estate
family
chest
nose
ring
coat belt
deck
submarine
sea port

september
july august
riding
wearing
cap
bearing
republic
municipality
direction
boundary
gap canal
bank₂
plateau
coast
pacific
san
undergraduate
harvard

april
dog
male
warrior
lion
tiger
territory

egg
lighting
wind
snow
forest
mountain
woods
lakes
rocks

1972 1971
1960s 1948
1985 1988
gold
colder
pine
oak
jaguar₄
hunter
goddess
queen emperor
duke
orange
gray
grey brown

§ 应用思路及方案

- 文本内容与doc2vec的关系

- 评分特征构建
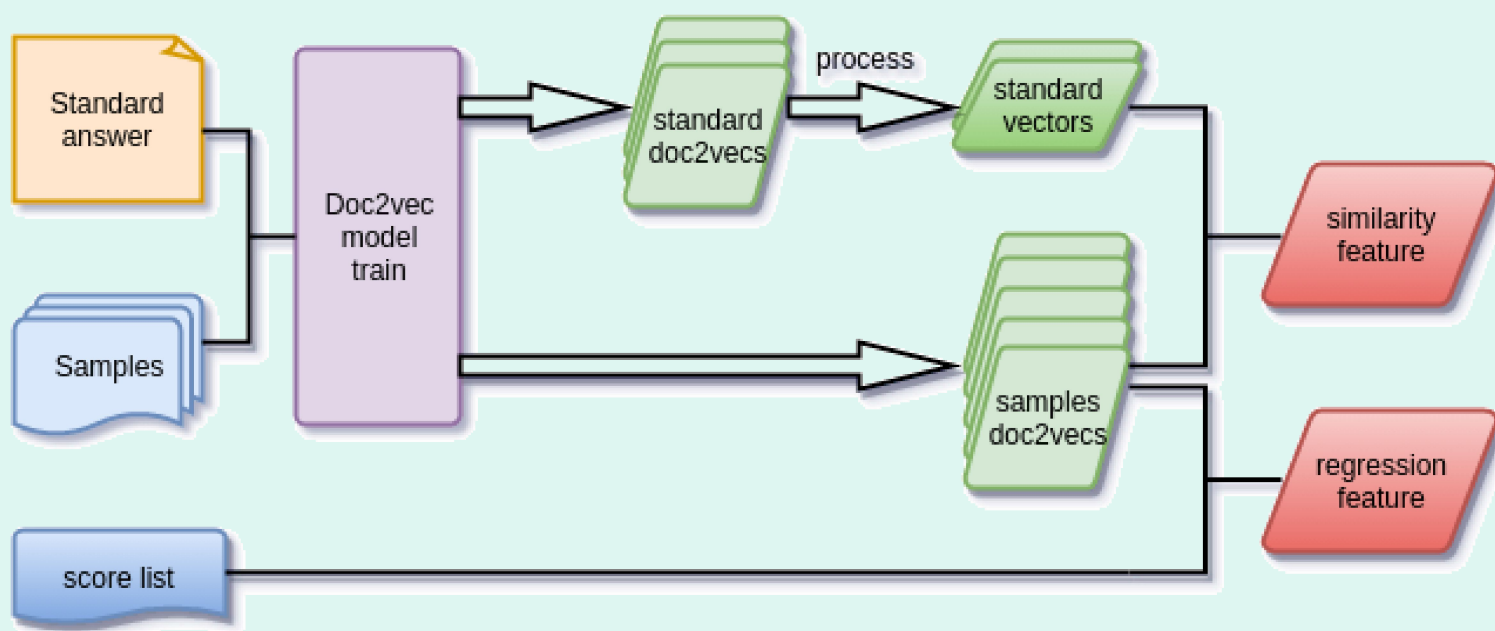
§ 相关实验结果

· 文本内容与doc2vec的关系

PAPER-000001-QT-000002_02(0.9689) ：

1 **0.8805 0.91** car bring many convenience we can go wherever we want and quickly get to our destination and it bring many problems either

2 **0.7337 0.82** it cause environment pollution and make our living around wrong and even bring many disease that do harm to our health

3 **0.8188 0.80** because too many many cars in the city it cause traffic jam

4 **0.6109 0.66** in my opinion we can do something to deal with this

5 **0.7973 0.75** for example we can go out for a walk instead of driving cars

6 **0.8866 0.83** we can use public traffic transportation instead of cars and plant many trees to improve our quality

7 **0.9298 0.92** i think the car bring our convenience and we should care more about the environment so lets live in green life

表1 中心句的doc2vec

doc	sim(vec_center, vec_standard)	乱序操作
1	0.92	0.57
2	0.84	0.57

表2 重复中心句构成的文本

paper	1 × sent	2 × sent	3 × sent	5 × sent
1	0.93	0.92	0.94	0.94
2	0.86	0.87	0.87	0.84

表3 特征加入到svr特征集后的评分性能变化			
	dim	pcoef dev,tst	评分性能
baseline	35	-----------	91.52%/0.8169
PV-DBOW_ridge	36	0.75 0.70	92.38%/0.8239
PV-DMc_mean	36	0.65 0.63	92.58%/0.8345
PV-DMc_Kmean2	36	0.67 0.62	92.63%/0.8313

结束

谢谢！！！