

数据集

人民日报：2020年10月04日-2021年10月04日

- 概况
 - 25590 articles
 - 742362 sentences
 - 0.021 billion words
 - 294730 tokens
 - 182004942 pairs (window size: 5)
 - 词云(120 words)



训练参数

vector dimension: 100

window size: 5

K: 5

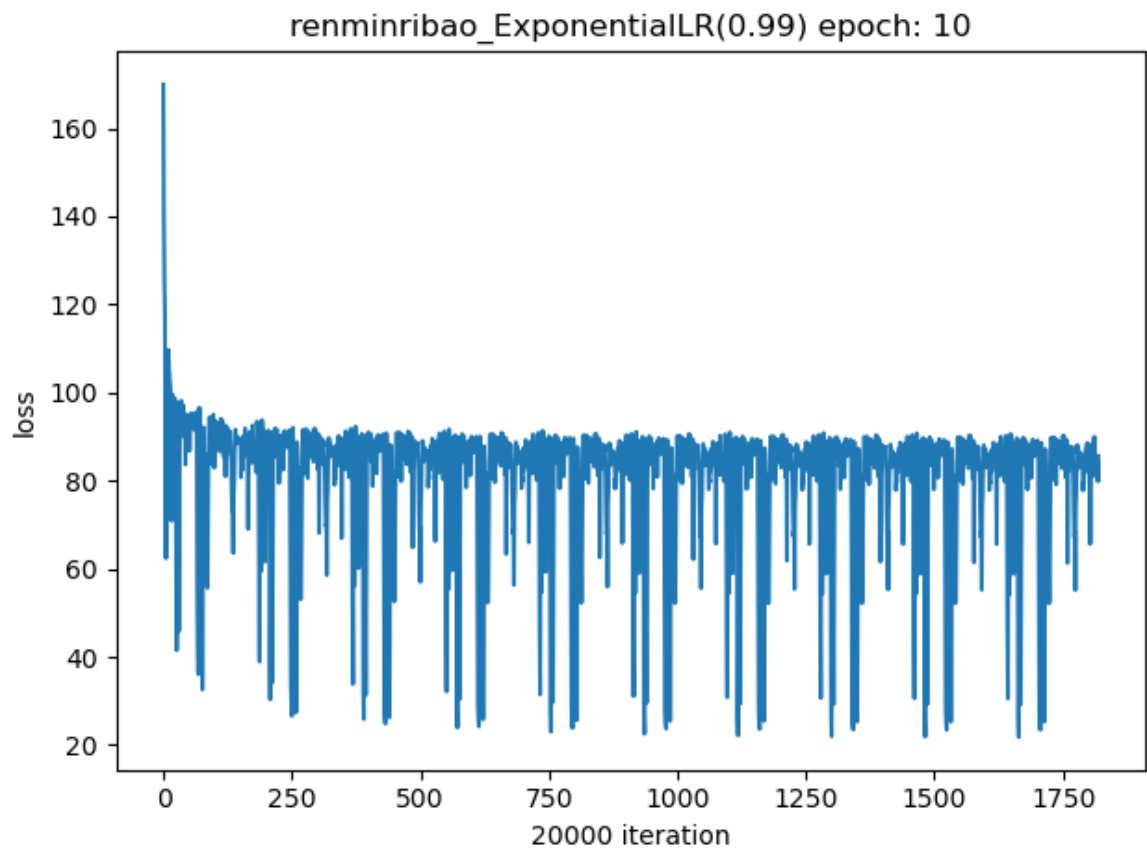
batch size: 50

epoch: 10






learning rate: 0.025

训练结果

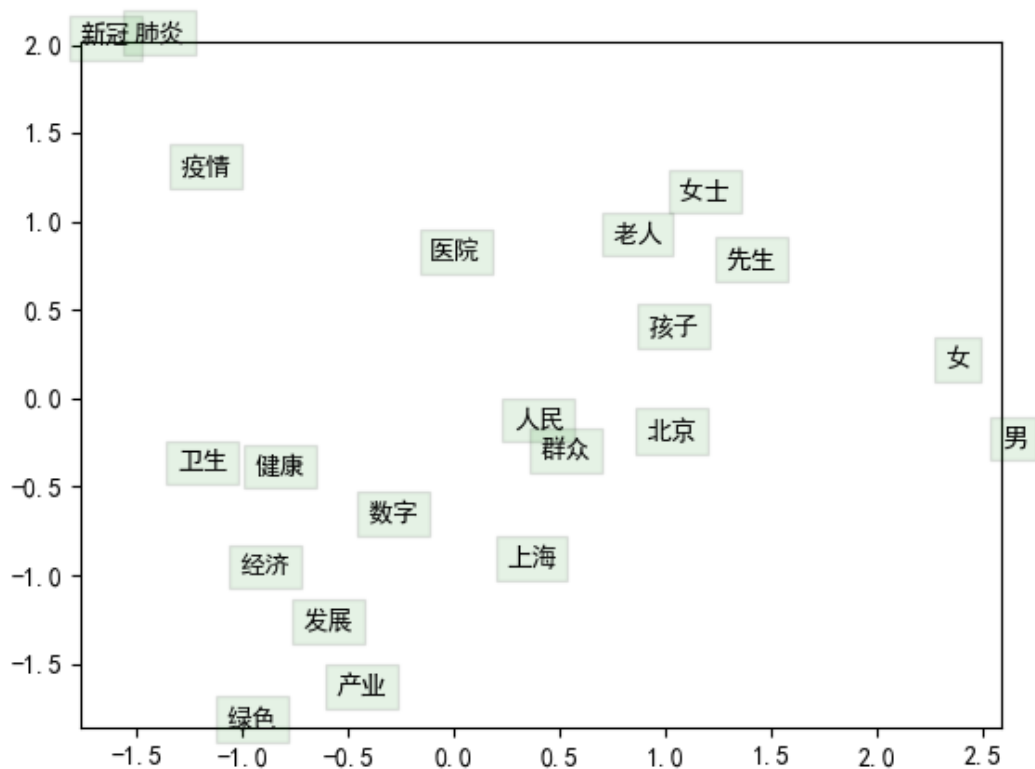
- loss下降曲线



- 词向量保存

 embeddings_1epoch_87.979loss.txt	2021/10/10 18:56	文本文档	369,541 KB
 embeddings_2epoch_85.086loss.txt	2021/10/11 7:25	文本文档	368,920 KB
 embeddings_3epoch_83.764loss.txt	2021/10/11 10:56	文本文档	368,447 KB
 embeddings_5epoch_82.469loss.txt	2021/10/11 20:40	文本文档	367,845 KB
 embeddings_9epoch_81.413loss.txt	2021/10/13 17:19	文本文档	367,189 KB

- 词向量可视化



- 相似性测试

```
words: 新冠
similar words:
[('肺炎', 0.8603816032409668),
 ('抗击', 0.7354764342308044),
 ('病毒', 0.7166163921356201),
 ('延宕', 0.6780088543891907),
 ('疫情', 0.6722403764724731),
 ('内新冠', 0.6578879356384277),
 ('疫苗', 0.6544620990753174),
 ('接种', 0.6023126244544983),
 ('同新冠', 0.5994764566421509),
 ('冠状病毒', 0.5871719121932983)]

words: 人民
similar words:
[('印度政府', 0.5914909839630127),
 ('衷心', 0.5727487802505493),
 ('发自内心', 0.5707407593727112),
 ('位置中国共产党', 0.5555098652839661),
 ('依靠人民', 0.5408075451850891),
 ('美好生活', 0.528394341468811),
 ('幸福', 0.5198014378547668),
 ('放在心上', 0.5187814235687256),
 ('群众', 0.5183557868003845),
 ('同胞', 0.5103503465652466)]

words: 发展
similar words:
[('旅游业', 0.6072622537612915),
 ('创新性', 0.5869481563568115),
 ('经济社会', 0.5864200592041016),
 ('跨越式', 0.5672993659973145),
```

('体育事业', 0.5610668063163757),
('文旅', 0.5601040720939636),
('高质量', 0.554404616355896),
('黄河流域', 0.5506658554077148),
('繁荣', 0.5431921482086182),
('进步', 0.5419984459877014)]

words: 绿色

similar words:

[('低碳', 0.7634264230728149),
('环保', 0.6420109272003174),
('转型', 0.6213286519050598),
('环境友好', 0.6132680773735046),
('讲究卫生', 0.6122341156005859),
('都市型', 0.6091085076332092),
('绿色革命', 0.5904538035392761),
('能源', 0.5856208205223083),
('碳循环', 0.5794229507446289),
('清洁', 0.5748923420906067)]

words: 北京

similar words:

[('未完待续', 0.6831703186035156),
('专栏(', 0.673279345035553),
('发本报', 0.6611239910125732),
('汝新华社', 0.6597181558609009),
('汪哲平本报', 0.6554037928581238),
('张丹峰新华社', 0.6540434956550598),
('杨文斌', 0.6522436141967773),
('第比利斯', 0.652116596698761),
('张芳曼本报', 0.6461266279220581),
('沈亦伶本报', 0.6431300044059753)]

words: 先生

similar words:

[('孙道临', 0.6072688102722168),
('女士们', 0.60660320520401),
('读一读', 0.6045931577682495),
('邓小平', 0.5893477201461792),
('合照', 0.5786231756210327),
('同志', 0.5731707215309143),
('查韦斯', 0.5671159029006958),
('这封', 0.5663946866989136),
('馆员', 0.5633106231689453),
('诗人', 0.5618208646774292)]

words: 数字

similar words:

[('数字化', 0.6296783685684204),
('人工智能', 0.6016441583633423),
('虚拟现实', 0.5961554646492004),
('飞桨', 0.591392457485199),
('开发者', 0.5806708335876465),
('交互式', 0.5730010271072388),
('裸眼', 0.5715974569320679),
('及物', 0.5713117122650146),
('全息', 0.5710457563400269),
('信息技术', 0.5665836930274963)]

words: 产业

similar words:

[('生态旅游', 0.6470369696617126),
('新兴产业', 0.6328831911087036),

('优势产业', 0.6275804042816162),
('委员会洛川县', 0.6137779355049133),
('特色产业', 0.6064068675041199),
('传统产业', 0.6043283939361572),
('转型', 0.603561520576477),
('一二三', 0.601521372795105),
('产业化', 0.6010552048683167),
('信创', 0.5984545946121216)]

words: 经济

similar words:

[('复苏', 0.5899767279624939),
('贸易', 0.5835217237472534),
('拉动', 0.5685859322547913),
('疫后', 0.5623974800109863),
('主动力', 0.5564263463020325),
('主任医师孙', 0.5527517795562744),
('主任吉林省田秋', 0.5423448085784912),
('中巴', 0.5394772291183472),
('服务业', 0.5384839177131653),
('腾格里', 0.5380746126174927)]

• 类比测试

男--博士,

女--?

[('硕士', 0.6023393869400024),
('香港科技大学', 0.5548747181892395),
('计算机系', 0.5405473709106445),
('剑桥大学', 0.534896969795227),
('李传锋', 0.5326829552650452),
('南京农业大学', 0.5315748453140259),
('旁听生', 0.5241110920906067),
('郭光灿', 0.5227930545806885),
('海归', 0.522708535194397),
('药学院', 0.5192795395851135)]

女--女士,

男--?

[('刘先生', 0.5702798366546631),
('李先生', 0.5457801818847656),
('某某', 0.540930986404419),
('感谢信', 0.5367707014083862),
('陈女士', 0.521395742893219),
('胡锡恩', 0.5200252532958984),
('领养', 0.5141705870628357),
('奶奶', 0.5098757147789001),
('夏行', 0.5089772939682007),
('一位', 0.5088501572608948)]

城市--建设,

农村--?

[('文化公园', 0.5251893997192383),
('拔地而起', 0.5027098655700684),
('廊道', 0.49922025203704834),
('外环', 0.49078091979026794),
('这座', 0.4868094027042389),
('园林景观', 0.47893524169921875),
('世界级', 0.4766378700733185),
('长隆', 0.47415652871131897),

('城市公园', 0.4703103303909302),
('大运河', 0.4702015519142151)]

经济--发展,

生态环境--?

[('业态', 0.5431914329528809),
('注入', 0.5175008177757263),
('旅游业', 0.5149070024490356),
('拉动', 0.507580041885376),
('引擎', 0.503610372543335),
('崛起', 0.5028665661811829),
('文旅', 0.5023772716522217),
('增长极', 0.5019749402999878),
('重振', 0.49992120265960693),
('增长点', 0.49920737743377686)]