Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

敏感文本数据脱敏方法

硕士研究生　关业礼

2022年05月29日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1. 了解文本脱敏的基本概念
  - 2. 了解文本泛化脱敏方法
  - 3. 了解文本差分隐私噪声扰动脱敏方法

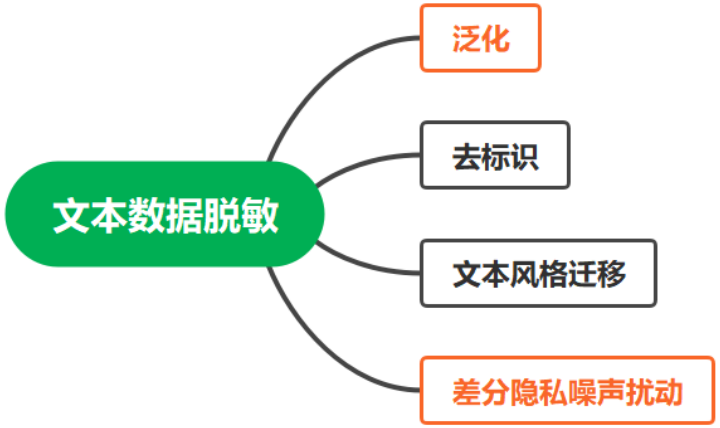
背景简介

- 当今社会的进步离不开数据的交互和共享，但广泛的数据流通也容易导致敏感甚至机密的数据遭到泄露

- 根据 datalossdb.org网站披露的信息显示，2014年美国有记录的数据泄露事件约50%发生在商界，约20%发生在政府部门，约30%发生在卫生和教育部门，而个人信息泄露无法估算

背景简介

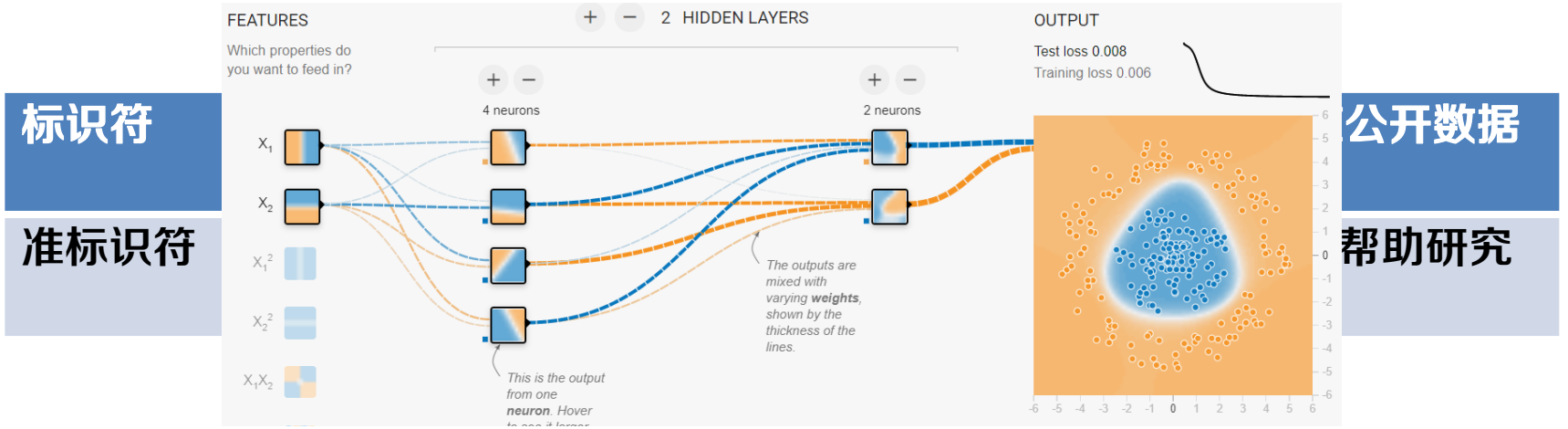
- 数据脱敏是指对敏感信息按照预设的规则或者变换算法进行数据变形，从而使得个人身份无法识别或者直接隐去敏感信息，达到保护隐私的目的
- 如何脱敏文本中的敏感信息并尽最大可能保留原始语义成为了一个挑战



基本概念

基本概念

- 隐私保护——从敏感数据中保护敏感信息不被披露
  - 敏感信息：电话号码、身份证号、人名等标识符和邮编、年龄等准标识符
  - 敏感数据：包含敏感信息的多元异构数据，如包含人名和疾病名称的医疗表格数据、包含人名和机构名称的文本数据、包含人脸的视频数据等
- 数据可用性——保留数据重要特征/属性的多少
  - 数据训练出的机器学习模型泛化能力越强，数据可用性越高



基本概念

- 脱敏方法中的隐私和数据可用性权衡
  - 脱敏技术不是单方向的保护隐私或保留数据可用性，而是对抗式的让隐私和可用性达到平衡
  - 原始数据可用性最高，隐私保护最差；加噪数据可用性和隐私保护中等；噪声数据可用性最低，隐私保护最好

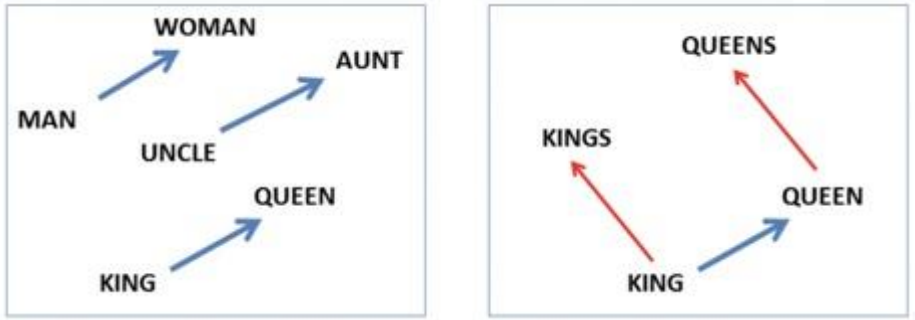
基本概念

- 词嵌入向量
  - 定义：概念上是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量。

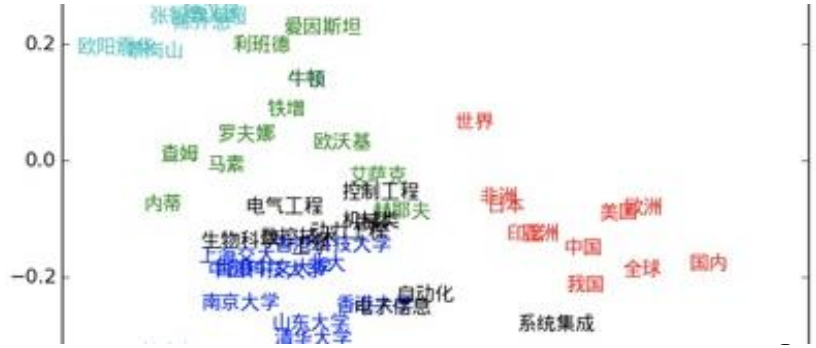
$$\begin{pmatrix} the \\ cat \\ sat \\ on \\ the \\ mat \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Word2Vec

类比性：



相似性：



基本概念

- 词向量相似度计算
  - 欧式距离
    - 欧式距离指在n维空间中两个点之间的真实距离。n维词向量$\boldsymbol{x} = (x_1, \dots, x_n)$和$\boldsymbol{y} = (y_1, \dots, y_n)$之间的欧式距离计算公式为：

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  - 余弦距离
    - 余弦距离指在n维空间中两个向量方向的差异。n维词向量$\boldsymbol{x} = (x_1, \dots, x_n)$和$\boldsymbol{y} = (y_1, \dots, y_n)$之间的余弦距离计算公式为：

$$\cos(\theta) = \frac{\sum_{i=1}^{n} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}}$$

词向量距离越近，单词相似度越高

基本概念

- 差分隐私
  - 定义：查询一对相邻的数据集D和D’，获得相同结果的概率非常接近，两者概率之比满足：

$$e^{-\varepsilon} \leq \frac{\Pr(M(D)=c)}{\Pr(M(D')=c)} \leq e^{\varepsilon}$$

- $d_x - privacy$
  - 定义：存在单词替换机制$M: X \to Y$，对于输出单词 $y \in Y$，任意的输入单词$x, x' \in X$的输出分布满足：

$$\frac{\Pr(M(x)=y)}{\Pr(M(x')=y)} \leq e^{\varepsilon \cdot d(x,x')}$$

x和x'的词向量距离

算法原理

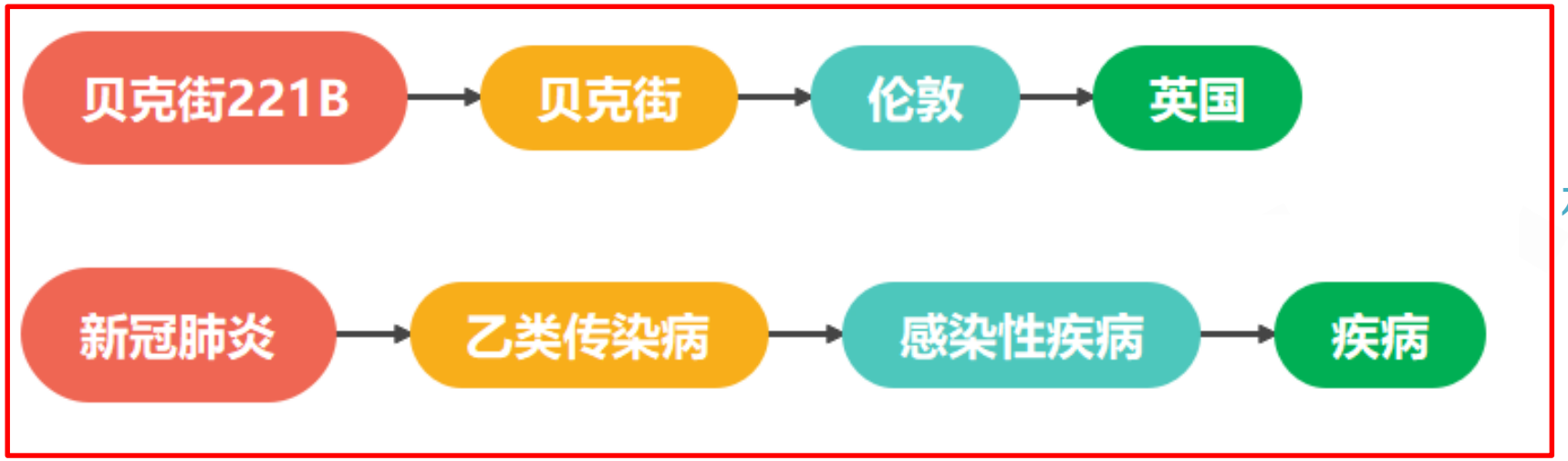
算法原理

T	对非结构化文本文档进行脱敏
I	包含敏感短语的文本文档
P	1、用大型语料库训练词嵌入模型 2、用词嵌入模型识别文档中的准标识符短语 3、对识别出的短语用泛化方法进行脱敏
O	脱敏文本文档

P	命名实体识别技术需要大量手动标注的数据且检测召回率低
C	需要构建一个包含各个准标识符短语的泛化版本的知识库
D	模型难以捕获文档中的准标识符短语
L	IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 2021 中科院2区

算法原理

- 泛化方法
  - 捕获文本数据中的敏感短语/单词并用更泛化的概念短语/单词替代

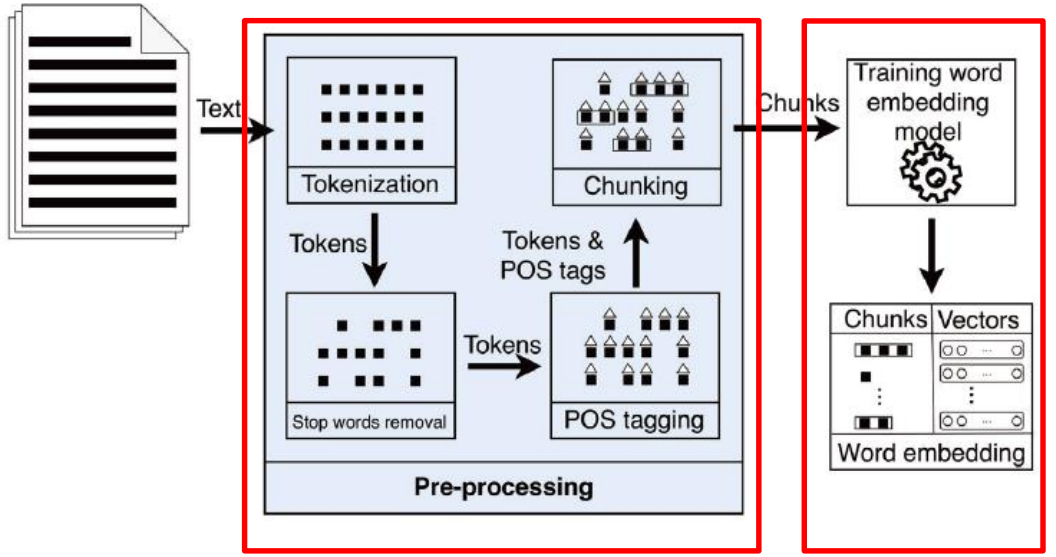


本体（ ontology ）：包含所有单词及其泛化版本的知识库

- 算法流程
  - 文本数据预处理与模型训练
  - 检测准标识符短语
  - 脱敏准标识符短语

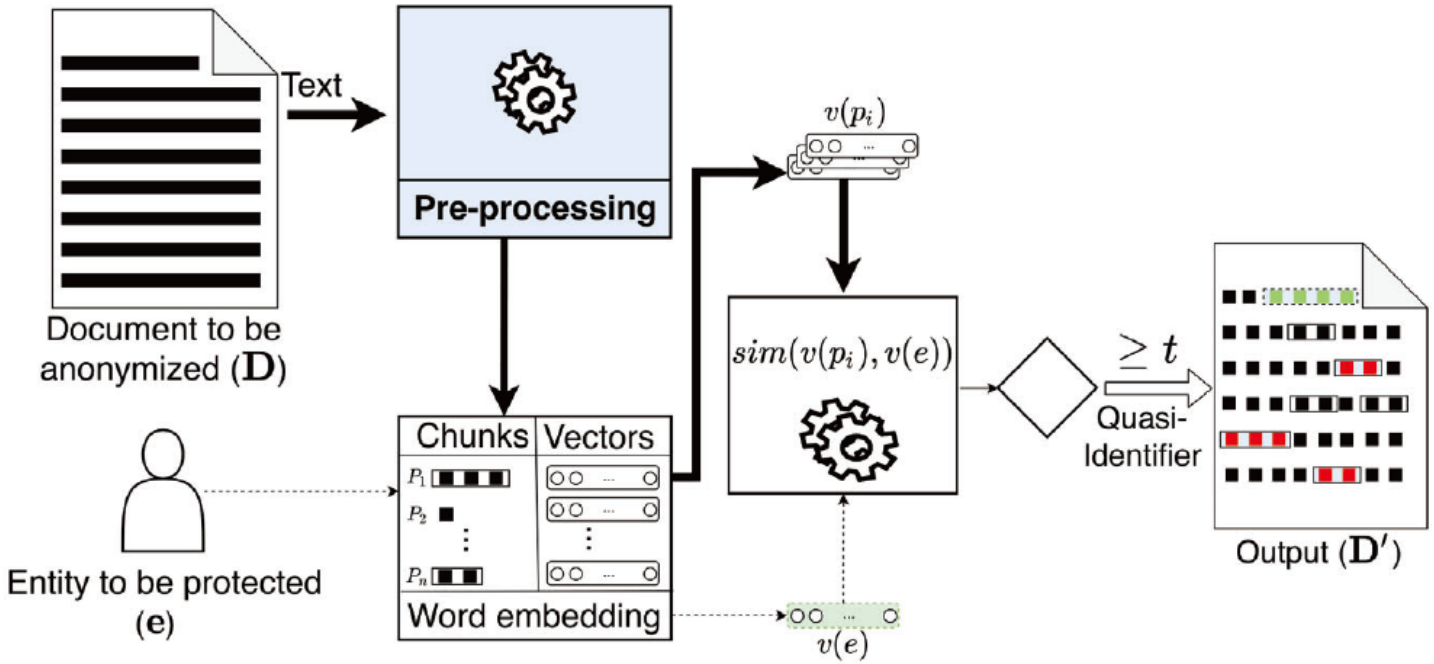
算法原理

- 文本数据预处理与模型训练
  - 分词（Tokenization）
    - 把一段完整的语句切分成具有独立含义的单词
  - 去除停用词（Stop words removal）
    - 去除对句子意义不大的单词，如“to”、“the”等
  - 词性标注（POS tagging）
    - 为分词结果中的每个单词标注一个正确的词性，如名词、动词、形容词等
  - 组块分析（Chunking）
    - 根据单词的性质将相似单词分组在一起变成短语



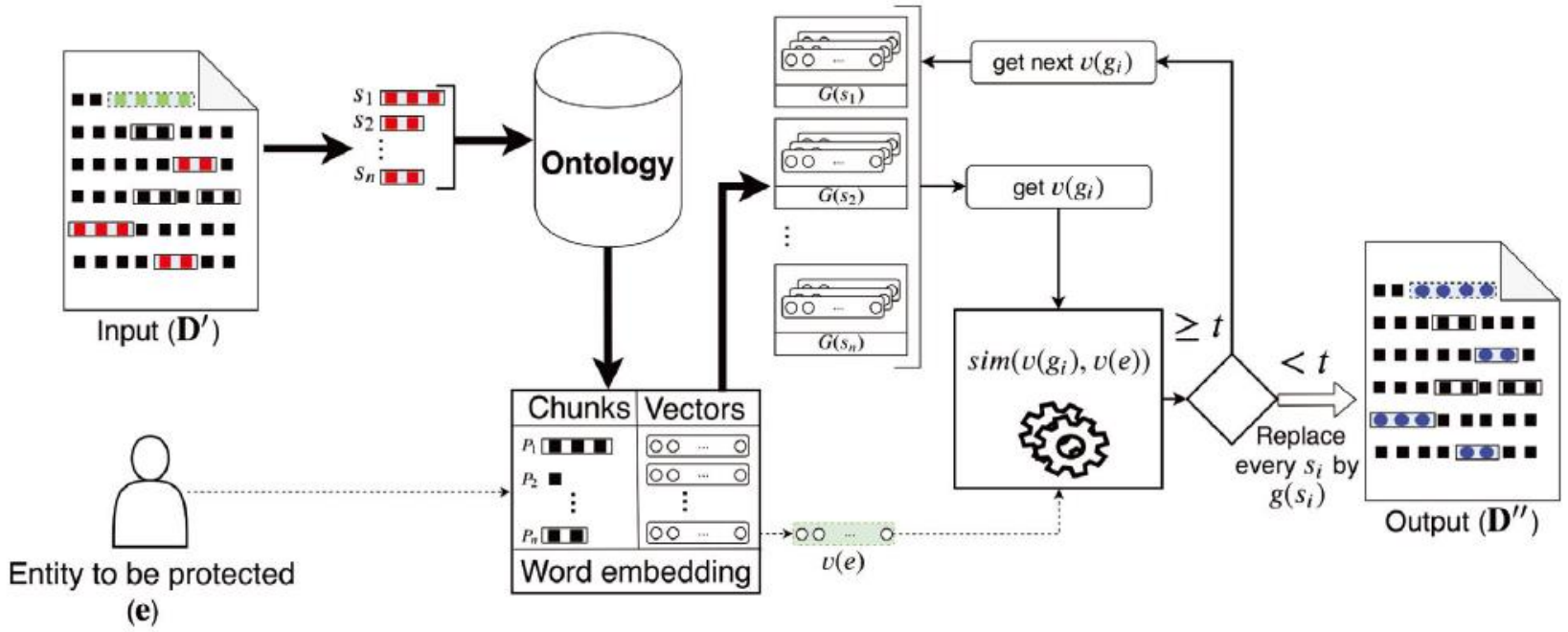
算法原理

- 检测准标识符短语
  - 给定文档D和敏感实体e，迭代地计算每个短语p_i的词向量与e的词向量的余弦相似度，若相似度超过阈值t，则认为p_i是准标识符短语s_i



算法原理

- 脱敏准标识符短语
  - 对每个准标识符短语s_i，通过将s_i与本体（ontology）中的概念单词相匹配，从本体中获得一组泛化单词序列$G(s_i)$
  - 迭代地计算$G(s_i)$中g_i的词向量与e的词向量的余弦相似度，若相似度超过阈值t，则使用g_i替代s_i



算法原理 实验数据

- 数据集
  - 维基百科“20世纪演员”类别下的19000篇文章的摘要组成的数据集，其中可能揭示演员身份的短语为准标识符
- 多维对比
  - 检测评估：检测准标识符短语的准确性
  - 可用性评估：文档在脱敏之后保留的可用性
  - 隐私性评估：针对重识别攻击的隐私保护效果
- 检测评估
  - 实验过程
    - 从数据集中随机取50份摘要，检测文档中的准标识符，并与实体识别模型NER3、NER4、NER7和Presidio对比检测的准确性

算法原理 实验结果

- 检测评估
  - 评估指标
    - 精度：检测到的敏感短语/检测到的短语
    - 召回率：检测到的敏感短语/总敏感短语
    - F1值：$\mathrm{F1} = 2 \cdot \frac{\text{精度} \times \text{召回率}}{\text{精度} + \text{召回率}}$
  - 实验结果
    - 该方法具有最高的召回率，隐私保护效果最好

Method	*Precision*	*Recall*	F_1
NER3	96.09%	19.59%	32.07%
NER4	97.59%	34.25%	49.72%
NER7	**98.32%**	27.89%	42.77%
Presidio	98.06%	27.07%	41.12%
Our method	82.69%	**81.24%**	**81.66%**
Our method (no pre-process)	83.48%	59.79%	69.00%

值越高，隐私保护效果越好

算法原理 实验结果

- 可用性评估
  - 实验过程
    - 用不同模型检测准标识符短语，使用抑制和泛化对短语进行脱敏，计算脱敏后文档的信息内容和脱敏准标识符短语的平均个数
  - 评估指标
    - 信息内容（Information Content）：用于量化一段文本的语义信息
  - 实验结果
    - 对命名实体识别模型来说，脱敏的准标识符短语越多，数据可用性越低
    - 该方法能够在准标识符短语被大量脱敏的情况下保持数据的可用性

值越高，可用性越好

Method	*Suppression*	*Generalization*	*Avg. masked terms*
NER3	**66.98%**	**85.44%**	33.8
NER4	39.70%	74.73%	64.08
NER7	54.00%	81.29%	54.16
Presidio	54.04%	81.00%	61.12
Our Method	48.48%	85.01%	**86.04**

算法原理 实验结果

- 隐私性评估
  - 实验过程
    - 用不同模型对文档进行脱敏后，使用重识别攻击模型攻击脱敏文档，计算攻击成功率
  - 评估指标
    - 攻击成功率：重识别攻击模型预测文档中演员名字的准确率
  - 实验结果
    - 该方法防御重识别攻击的效果很好，仅在人工注释之下

值越低，防御效果越好

Input	*Correct predictions*
Original summary	84.67%
Manual annotation	2.00%
NER3	18.00%
NER4	16.00%
NER7	37.33%
Presidio	18.67%
Our Method	10.00%

算法原理 优劣分析

- 优势
  - 将敏感短语的检测和脱敏有效地结合
  - 在隐私保护和数据可用性中表现出了良好的平衡性
- 局限性
  - 对比其他命名实体识别模型检测敏感短语的精度较低
  - 攻击者可以通过本体反向推导出被脱敏的短语

算法原理

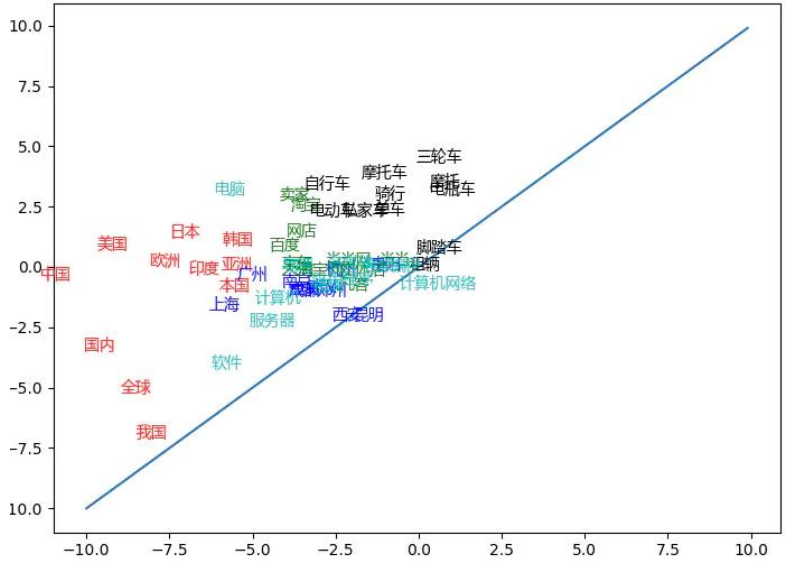
算法原理 TEM

T	对文本语句进行差分隐私脱敏
I	包含敏感信息的文本语句
P	1、获取语句中的首个单词 2、计算该单词与其他单词的词向量距离，按照距离阈值γ区分距离内单词和距离外单词 3、取词向量距离的负数为分数，对分数加噪 4、用加噪分数最大的单词替换原单词 5、取下一个单词，循环步骤2-5
O	脱敏文本语句

P	差分隐私单词替换机制的噪声太大导致“高维诅咒”
C	证明单词替换机制满足$d_x - privacy$
D	差分隐私单词扰动机制随机性高，文档可用性损失大
L	ICLR 2021

算法原理 TEM

- Madlib单词脱敏方法的问题
  - 高维诅咒
    - 已知初始n维词嵌入向量，对该向量加差分隐私噪声，则噪声也是n维，加噪词向量与初始词向量距离偏差过大
  - 自适应噪声
    - 对词向量添加的噪声没考虑词向量空间密度，在向量空间密集区域噪声过大容易输出不相关单词，在稀疏区域噪声过小容易输出原单词



算法原理 TEM

- **TEM的解决方案**
  - **对词向量之间的距离加噪，绕过“高维诅咒”**

Movie	0.5	0.6	0.1
Cat	0.2	0.2	0.4
Dog	0.3	0.3	0.4
Bird	0.4	0.5	0.1

词向量欧氏距离

	Movie	Cat	Dog	Bird
Movie	0	0.58	0.47	0.2
Cat	-	0	-	-
Dog	-	-	0	-
Bird	-	-	-	0

  - **设定距离阈值γ，依靠阈值γ来修改词向量空间密度，从而使噪声自适应**
    - **与输入单词的距离超过γ的候补单词，设定候补单词的距离恒为γ**

	Movie	Cat	Dog	Bird
Movie	0	0.58	0.47	0.2
Cat	-	0	-	-
Dog	-	-	0	-
Bird	-	-	-	0

假设γ=0.5

	Movie	Cat	Dog	Bird
Movie	0	0.5	0.47	0.2
Cat	-	0	-	-
Dog	-	-	0	-
Bird	-	-	-	0

算法原理 TEM

- 算法流程
  - 预处理
    - 输入单词x，计算其他单词与x的词向量欧式距离，距离小于γ的单词放入$L(w)$中，大于则放入$\bar{L}(w)$中，并设定一个分数$f_{\perp}(w)$
  - 单词替换
    - 计算x与$L(w)$中单词的词向量欧氏距离，对距离取负得到距离分数，对距离分数和$f_{\perp}(w)$加上Gumbel噪声，输出最大分数对应的单词，若最大分数为$f_{\perp}(w)$，则输出$\bar{L}(w)$中任意单词

Algorithm 3 Metric Truncated Exponential Mechanism

Input: Finite domain $\mathcal{W} = \{1, ..., m\}$ of elements, element index $x \in \mathcal{W}$, truncation threshold γ, metric $d_{\mathcal{W}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$, and privacy parameter ε.
Output: Element index.

Pre-processing:
for each $x \in \mathcal{W}$ **do**
  Create a list $L(w)$ of elements where each element i satisfies $d_{\mathcal{W}}(i, w) \leq \gamma$ and let $\bar{L}(w)$ be the list of remaining elements $\mathcal{W} \setminus L(w)$.
  Define for each $x \in \mathcal{W}$ the score of a $\perp$ element as $f_{\perp}(w) = -\gamma + 2\ln(|\bar{L}(w)|)/\varepsilon$
end for
Selection:
Given input $x \in \mathcal{W}$, for every element in $L(w) \cup \perp$ add noise from a Gumbel distribution with mean 0 and scale $2/\varepsilon$ to each score $-d_{\mathcal{W}}(\cdot, x)$
Set y as the element with maximum noisy score
if $y = \perp$ **then**
  Return random sample of $\bar{L}(w)$
else
  Return y
end if

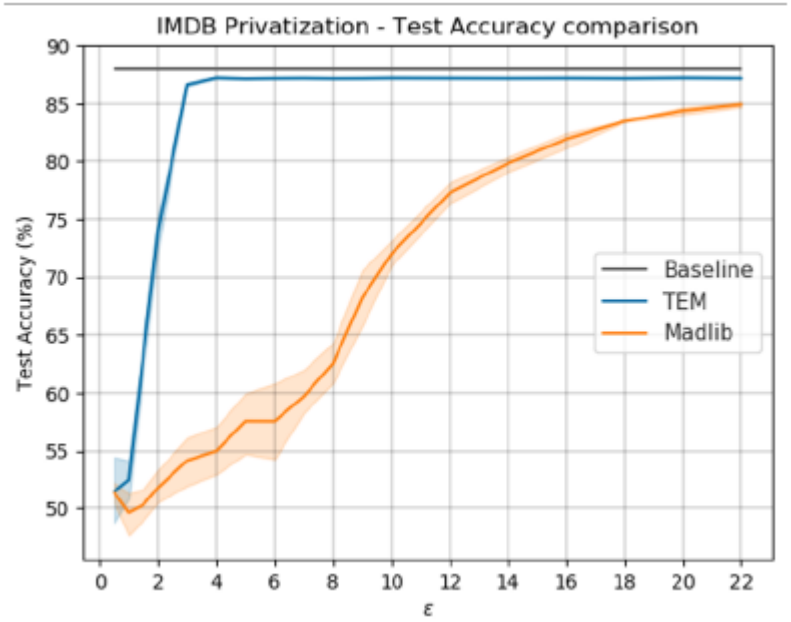
算法原理 实验数据

- 数据集
  - IMDB数据集，训练集和测试集各包含25000条电影评论文本，标签为正和负，分别对应评论者好评和差评
- 多维对比
  - 可用性评估：文档在脱敏之后保留的可用性
  - 隐私性评估：针对成员推理攻击的隐私保护效果
- 可用性评估
  - 实验过程
    - 使用该方法和Madlib方法脱敏训练集，用脱敏的训练集训练情感二分类模型，并使用模型对原始的测试集进行预测，计算模型预测的准确率

算法原理 实验结果

- 可用性评估
  - 评估指标
    - 准确率：模型正确分类的文本样本/总文本样本
  - 实验结果
    - 与Madlib方法相比，TEM方法保留的数据可用性更高，并且在$\varepsilon > 4$时与未脱敏文本的可用性相差不大

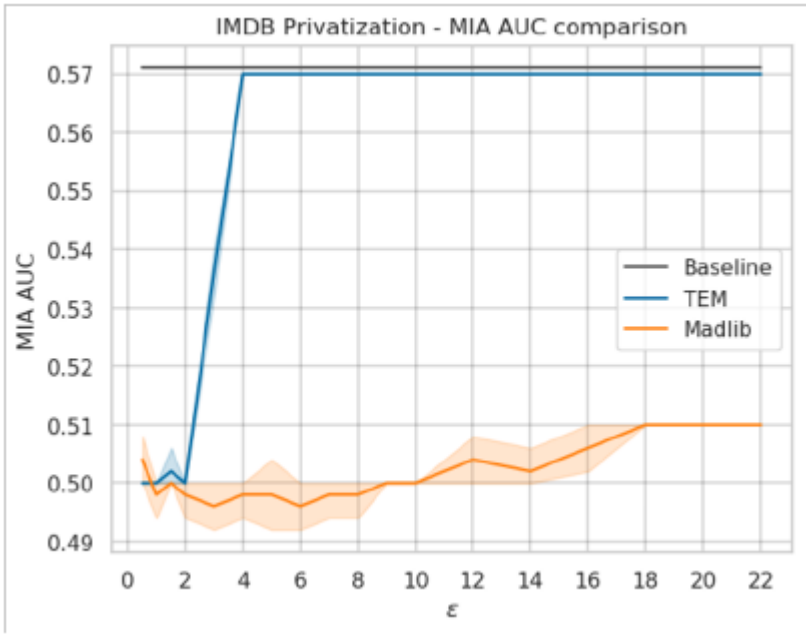
值越高，可用性越好



算法原理 实验结果

- 隐私性评估
  - 实验过程
    - 使用该方法和Madlib方法脱敏训练集，用脱敏的训练集训练情感二分类模型，用成员推理攻击分类器攻击模型，计算攻击分类器的AUC值
  - 评估指标
    - AUC值：ROC曲线下与坐标轴围成的面积，通常用来评价模型分类的效果
  - 实验结果
    - Madlib方法给文本添加了更多的噪声，隐私保护效果更好，而TEM方法在$\varepsilon < 4$时才能有效防御攻击

值越低，防御效果越好



- 优势：
  - 单词替换时添加了较小的噪声，增加了数据可用性
  - 选取了对单词词向量距离加噪，绕过了高维诅咒
  - 通过ε控制隐私和可用性的权衡，量化隐私保护效果
- 局限性：
  - $\varepsilon > 4$时不能防御成员推理攻击

应用总结

应用总结

- 与表格数据隐私保护的比较
  - 不能彻底防御重识别攻击、成员推理攻击等攻击手段，因为输出单词与原单词存在关联
  - 表格数据可以生成假数据，文本数据不能生成假“单词”
- 拓展方向
  - 防御成员推理攻击、重识别攻击等多种攻击手段
  - 发布和共享社交媒体文本、医疗文本和法律文本等文本数据集

参考文献

[1] Hassan F, Sanchez D, Domingo-Ferrer J. Utility-Preserving Privacy Protection of Textual Documents via Word Embeddings[J]. IEEE Transactions on Knowledge and Data Engineering, 2021.

[2] Feyisetan O, Balle B, Drake T, et al. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. 2020: 178-186.

[3] Carvalho R S, Vasiloudis T, Feyisetan O. TEM: High Utility Metric Differential Privacy on Text[J]. arXiv preprint arXiv:2107.07928, 2021.

[4] https://zhuanlan.zhihu.com/p/474642680

[5] https://playground.tensorflow.org

大成若缺，其用不弊。
大盈若冲，其用不穷。
大直若屈。大巧若拙。
大辩若讷。静胜躁，寒
胜热。清静为天下正。

谢谢！