

Investigating the Application of Differential Privacy to Mitigate Privacy Issues in Natural Language Processing

Stephen Meisenbacher, 22.03.21, Guided Research Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
www.matthes.in.tum.de

Outline

Introduction (recap)

Research Questions

Methodology

- Literature Review
- Interviews
- Paper

Findings

- Privacy Vulnerabilities
- Differential Privacy for NLP
- Differential Privacy $\rightarrow$ d_x -privacy
- Benefits and Limitations

Future Work

DP in NLP Learning Nugget / Web App Examples

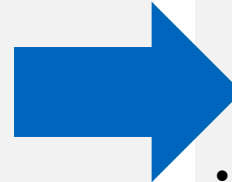
Conclusion

Background

Main driving points:

1. People's viewpoint on privacy is becoming increasingly transparent
 - Growing awareness of risks → skepticism
 - + perceived lack of control
2. "Big data" on the rise
 - Private sector: booming market cap
 - Academic research follows accordingly
3. Data breaches, their ensuing consequences
 - General upward trend in occurrences – costly!
 - Result: GDPR, CCPA, what's next?
4. Privacy as a topic of interest
 - Privacy research = hot topic
 - "Data Privacy Will Be The Most Important Issue In The Next Decade" *

*<https://www.forbes.com/sites/marymeehan/2019/11/26/data-privacy-will-be-the-most-important-issue-in-the-next-decade/>



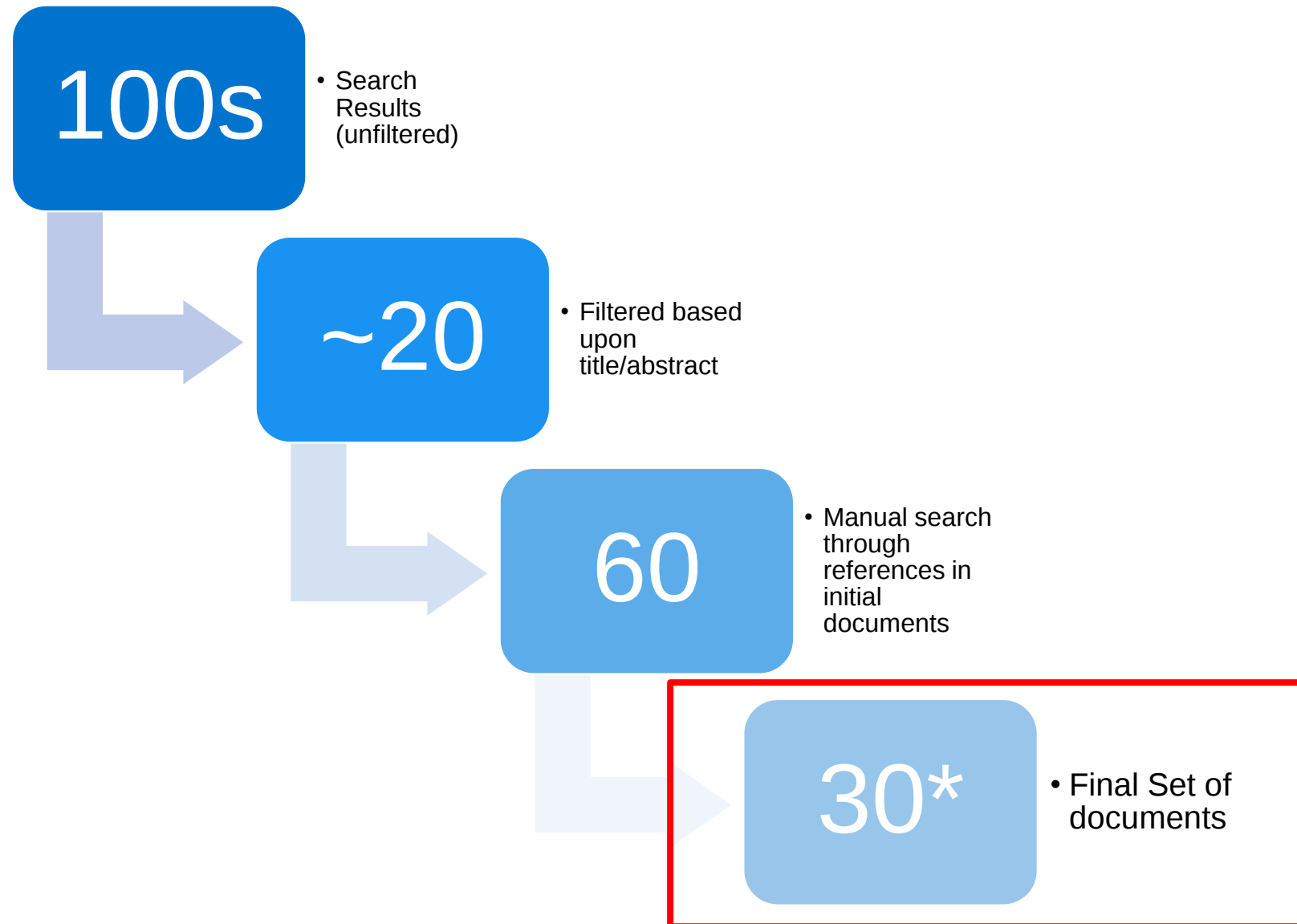
Motivation + Goals

"As a result":

- Protection of privacy is as important as ever
 - Need for privacy-enhancing technologies (PETs)
- Great candidate: Differential Privacy
 - Mathematical foundations, privacy guarantee, useful properties
- Scope: Natural Language Processing
 - Dealing with large-scale, unstructured text data
 - DP + NLP feasible?
- Goal: gain an overview of DP in NLP
 - Privacy vulnerabilities
 - Technical applications and use cases
 - Pros, cons
 - Overall: current work + potential

1. What vulnerabilities to current NLP techniques is Differential Privacy capable of preventing?
2. What are the foundations of Differential Privacy, and how can it be applied to NLP tasks?
3. What are the distinct benefits and limitations of applying Differential Privacy to NLP tasks?





- **Semi-structured interviews** with 4 people
 - Experts in the field working with privacy + NLP
 - e.g. Researcher at Amazon, PhD candidate
- **30-60 minute** video conference interview
- Generally following a pre-defined list of questions:
 - **General:** current work, background with privacy, thoughts on privacy + NLP
 - **RQ1:** NLP privacy vulnerabilities, attack types, preventative work so far
 - **RQ2:** DP foundations, application to NLP, use cases, technical implementations
 - **RQ3:** major advantages, current limitations, future improvement, thoughts on future of private NLP
 - Total: 18 questions + sub-questions + 1-2 tailored specifically to interviewee
- After the interview: transcribe
 - ~3 hours → ~25 pages
 - Helpful for writing the paper

Methodology – Paper

- Overall goal: summary report of research findings
- Main sections:
 1. Introduction of problem
 2. Overview of current/related work
 3. Privacy vulnerabilities to NLP
 4. Foundations of DP + early applications to NLP
 5. Foundations of d_x -privacy + applications
 6. In-depth discussion of applicability, benefits, limitations
 7. Prognosis for future work
- End result:
 - 12 page Guided Research report
 - Separate journal submission (PETS 2021)



Word Embeddings

- Representations of words in vector space
- Main goal: capture semantic (or contextual) associations between words
 - “Distributional” semantics
- Heavily used in modern NLP
- Models trained on billions of sentences → possibly sensitive
- Issue: when private information is encoded into embeddings

Language Leakage

- Text organized into some representation
 - e.g. syntactical, lexical features
- Inherently a “stylometric profile”
- Good for features
- But: conveniently expose implicit information about authorship style and identity
- “Reverse-engineering” pieces of text is relatively easy

Neural NLP

- Much of NLP today is done using a neural component
- Main goal: learn patterns to make accurate predictions, classifications, etc.
- Problem: almost too powerful
- Memorization of sensitive parts of the input text
- *Exposure* is tied to learning → not ideal

Data Release

Setup: data is released in some form to a third party

Goal: utilization in research, creation of a product, etc.

Problem: a malicious user gains access to this data

There are many cases where such (text) data is shared:

- Medical context (doctors' notes, hospital records)
- Online reviews
- Social media / blog posts
- Government records
- Genome sequences

Even problematic in embedding form!

Model Abuse

Setup: centralized or decentralized learning

- Centralized: central model does computations
- Decentralized: local computations, central server (i.e. in cloud)
- Good example: IoT applications

Goal: any shared learning task

Problem: presence of a malicious user

Two ways for malicious user to infer data:

- Black-box access: query outputs
- White-box access: black-box + access to (sample of) original data

In the wrong hands, query outputs can be exploited!

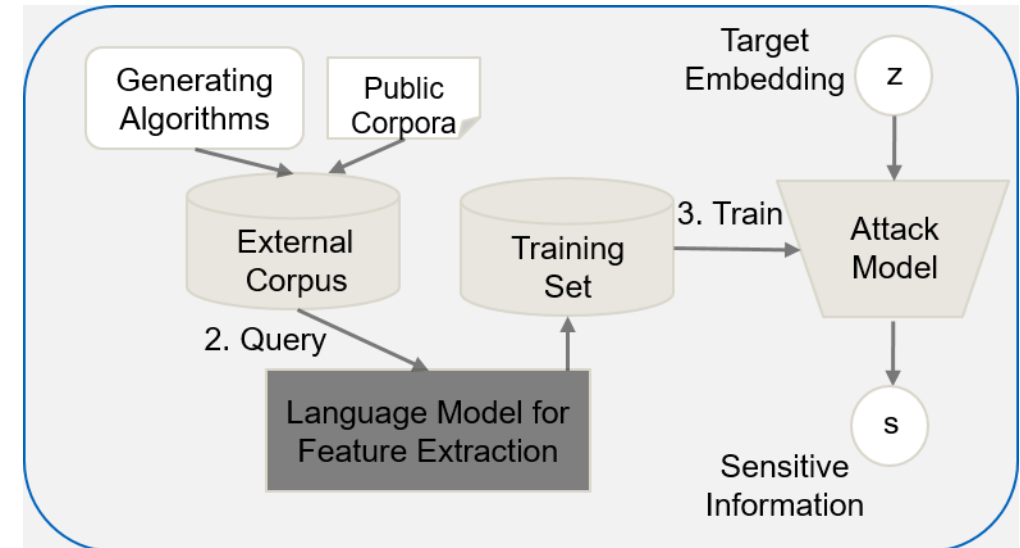
- Focus here: Inference Attacks
 - Infer information from obtained data
 - Text: “learn what’s inside”
- Two classes (relevant to DP + NLP)

Pattern Reconstruction

- Target: sensitive information with fixed format
 - e.g. SSNs, zip codes, birth dates, phone #s
- Attacker’s goal: reconstruct these fixed-format strings given some text representation

Keyword Inference

- Target: keywords given some domain knowledge
- Little to no prior knowledge is really necessary



A general attack pipeline that can be used to leverage the mentioned attacks

Based on: X. Pan, M. Zhang, S. Ji and M. Yang, "Privacy Risks of General-Purpose Language Models"

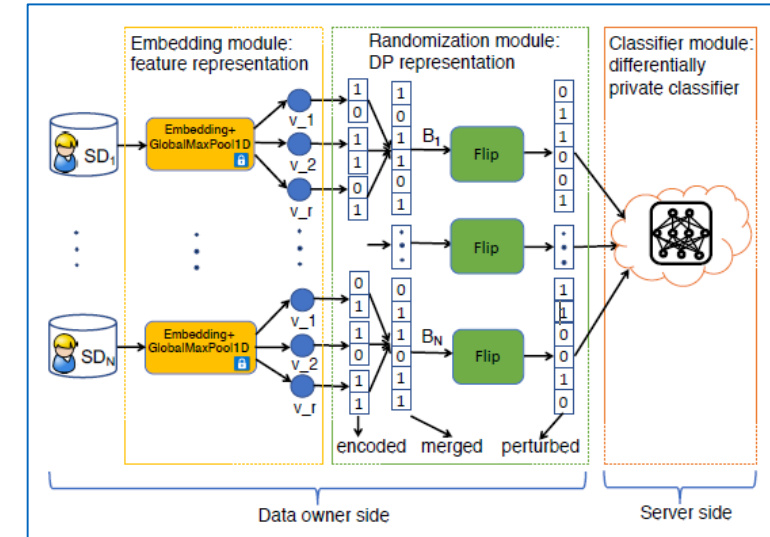
Differential Privacy for NLP – Foundations

- DP key idea: privacy protection for the *individual*
 - Originally for (structured) database setup
 - Randomized response → “plausible deniability”
 - ϵ parameter to quantify privacy guarantee
- Why DP for NLP?
 - Not protection against inferences themselves
 - Protection of the *individual* against information gained through inference
 - NLP: protect contributor’s of original text from attacker knowledge gain
 - Plausible deniability w.r.t. whether inferred information is true
- Immediate Challenge with NLP
 - Mainly unstructured data!
 - Who/what is the “individual” in a dataset now?
 - A document, a word?
 - What “databases” are adjacent/neighboring?
 - Standard DP: two databases differing by one entry
 - DP with NLP: unclear

$$\frac{\Pr[\mathcal{M}(x) \in S]}{\Pr[\mathcal{M}(x') \in S]} \leq e^\epsilon$$

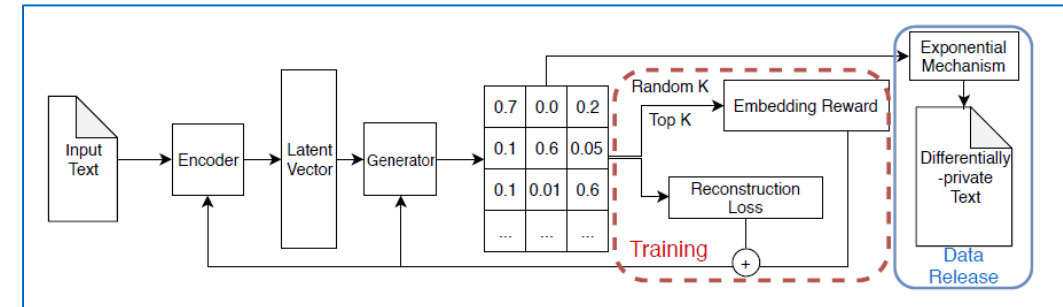
Differential Privacy for NLP

- How to transfer concepts?
 - Two “databases”, each with one document
 - Documents differ by exactly one word (token)
 - Adjacent = can achieve the other document by make a single edit
- Extrapolate: arbitrary number of edits
 - Compositionality of DP important!
 - More edits needed = more distinguishable / less private, and vice versa
- Key concept: *indistinguishability* between documents achieved via composition of edits
- Application examples:
 - OME**: perturbed binary text vector representations
 - SYN-TF**: synthetic term frequency vectors via the Exponential Mechanism
 - ER-AE**: differentially private text generation
 - DP-SGD**
- (Numerical) text representation required!



Optimized Multiple Encoding (OME)

Lyu, Lingjuan et al. "Towards Differentially Private Text Representations"



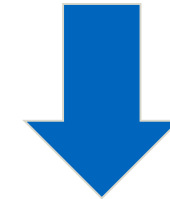
Embedding Reward Autoencoder (ER-AE)

Bo, Haohan et al. "ER-AE: Differentially-private Text Generation for Authorship Anonymization"

From Differential Privacy to d_x -privacy

- Problems with (standard) DP with NLP
 - Made to fit the original inequality
 - Too strict – any two documents are neighboring
 - Result: high ϵ values required
 - No consideration to semantics of language
 - Good start: Exponential Mechanism
- Need something more tailored to NLP!
- Solution: d_x -privacy
 - a.k.a. metric DP, “generalized DP”
- Key concept: *metric spaces*
 - Data represented as points (think embedding space)
 - (Distance) metric d_x = adjacency measure
- Incorporate into new, modified inequality
- Intuition: required indistinguishability depends on similarity of two documents
 - Smaller distance (greater similarity) = more indistinguishable output must be
 - = “scaling” the noise required

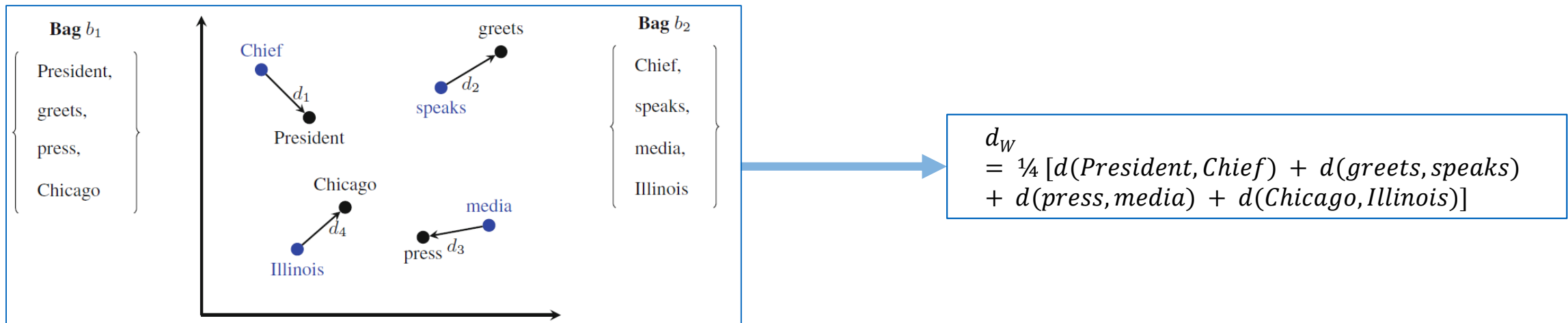
$$K(x)(z) \leq e^\epsilon K(x')(z)$$



$$K(x)(z) \leq e^{\epsilon d_x(x, x')} K(x')(z)$$

d_x -privacy for NLP

- Example metric: *Word Mover's Distance (WMD)*
 - i.e. how we “move” one document to another using single word (vector) edits
 - Using the minimal WMD, the DP inequality works
- In the word embedding space, use cosine distance for single words
- Then, required indistinguishability is scaled by semantic closeness
- Key takeaway
 - Standard DP: emphasis is on databases differing by an individual entry
 - Metric DP: emphasis on **how** these individuals (i.e. documents) differ



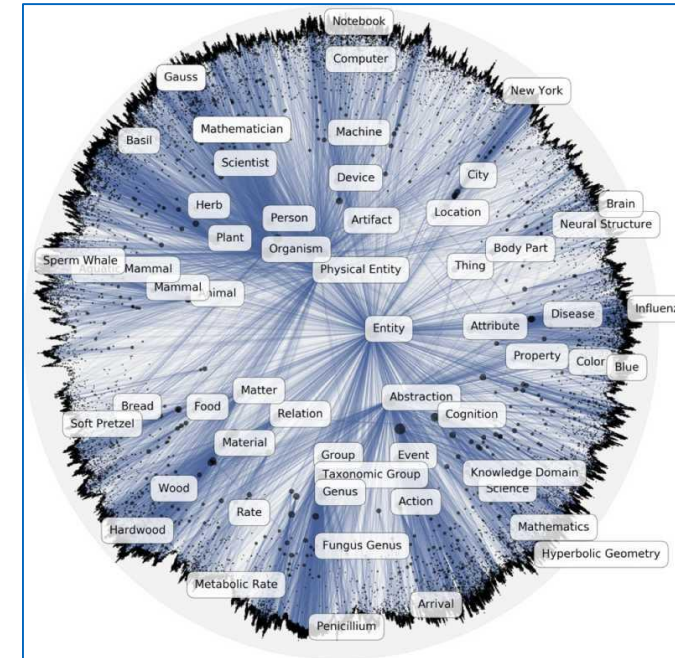
Example from Fernandes, Natasha et al. “Author Obfuscation Using Generalised Differential Privacy.”

d_x -privacy for NLP (cont.)

Some applications:

- Original approach: Euclidean space
 - Calibrated noise via Laplace Mechanism
 - Projection to nearest neighbor
- Hyperbolic space
 - Key idea: model hierarchical relationships in language
 - Makes sense, especially with hyperbolic geometry
- Mahalanobis distance
 - Tackle problem of sparseness
 - Take into account the shape of a particular (sub-)space

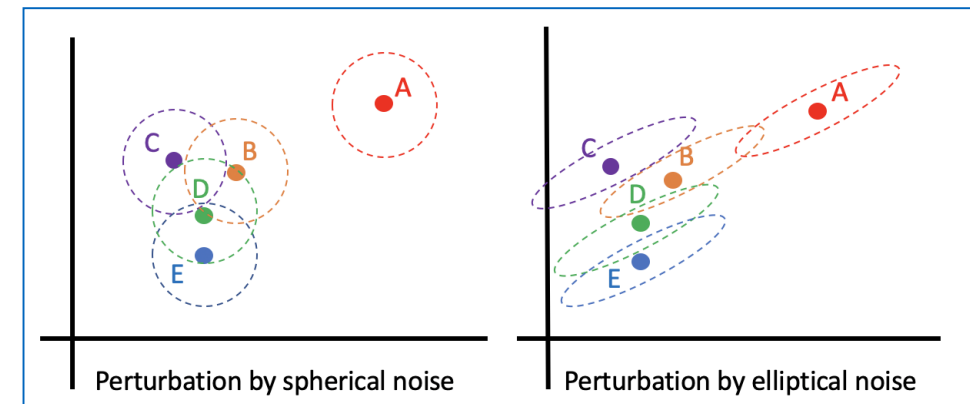
Important to note: change of metric / metric space quite seamless with metric DP!



Hyperbolic Embeddings

example, source:

<https://www.techleer.com/articles/478-insight-into-hierarchical-representations-through-poincare-embedding/>



Xu, Zekun et al. "A Differentially Private Text Perturbation Method Using a Regularized Mahalanobis Metric."

DP in NLP: Benefits

1. Reasoning about privacy in NLP

- Traditionally hard with NLP / text
- DP provides the foundation

2. Flexibility

- Today: many different NLP architectures with underlying space, geometries, metrics, etc.
- DP provides flexibility to work with these
- “Future-proof”?

3. Scalability

- Much of the unstructured data in question is at scale → big data
- PET of choice must be able to handle this
- DP: “schematic” that allows for optimization and tailoring to the problem at hand

4. Promising Privacy Preservation

- Great initial results in *privacy experiments*
- Best way to exhibit: decreasing performance of attack models
- Also: privacy statistics, e.g. N_w and S_w

Utility

- Not necessarily negative
- Often though: clear utility hit
 - Lower $\epsilon \rightarrow$ lower utility
- But: “no free lunch”
 - More privacy must mean less “something else”
- Quantifying and evaluating this tradeoff is crucial
- Ultimately: design choice
 - What do I **value** more?
 - ϵ : a great indicator for balancing privacy and utility

Structural Limitations

- DP imposes structure to inherently unstructured text
 - i.e. reasoning about text as documents
 - “Transfer” from standard DP
- d_x -privacy addresses this
 - Criticism: how much are we willing to deviate?
- Another limitation: DP assumes static nature
 - But much of NLP data today is streaming / dynamic
 - Time value?

Downstream vs. Natural Language

- Related to utility
- Basically with DP + NLP: add noise to text representations
- Downstream task is end goal:
 - Effects of noise are less important
- However:
 - Projecting back to natural language is non-trivial
 - So far: grammatically shoddy, quite unnatural language
- Bottleneck: choice of embedding model

Explainability

- Key issue
- Essentially: how to explain what is going on with DP + NLP
- Some questions:
 - How does DP fit with NLP?
 - What changes does text undergo?
 - When is text *truly* private?
- DP gives a good start
 - ϵ useful but not complete
 - Still somewhat cryptic
- Ultimately, greater transparency is needed
 - (Hopefully) obtained through time with further research

Not a Fix-all

- Perhaps self-evident
- Good: quantifiable guarantee
- But:
 - At the cost of structural limitations / design requirements
 - Only when using data to make broader generalizations (“learning tasks”)
- Mapping to text representation not always desired or possible
- Needed: better study of DP vs. other PETs w.r.t. NLP

1. Further exploration of the **privacy-utility tradeoff** when applying DP to NLP
 - Maximum privacy + minimum utility loss = ultimate goal
 - Important for better explainability
2. Integration of DP into **modern NLP architectures**
 - So far: only rather “simple” NLP pipelines
 - Would be nice to see with advanced sequence models (e.g. transformers), adversarial networks, etc.
3. Compatibility with **more recent text representation models**
 - For example, with contextual word embeddings (e.g. BERT) – not easy!
4. The role of DP in **non-static data settings**
 - DP’s role, applicability, effectiveness
 - Investigating DP with streaming datasets, the “time value” of ϵ
5. Other **generalizations of standard Differential Privacy**
 - d_x -privacy is a good start, others possible?
6. Explaining DP in NLP
 - Privacy concerns the individual user
 - → maintain transparency with this person!

Results (Deliverables)

List of deliverables from this Guided Research:

1. Guided Research Report
2. Journal Submission to the PET Symposium
3. (8) Learning Nuggets for new “DP in NLP” learning path:
 - Privacy Vulnerabilities in Natural Language Processing
 - Differential Privacy in Natural Language Processing
 - Differential Privacy: Applications to NLP
 - d_x -privacy: Generalized Differential Privacy
 - d_x -privacy: Applications to NLP
 - The Exponential Mechanism
 - Benefits of Differential Privacy in NLP
 - Limitations of Differential Privacy in NLP
4. Web App for corresponding Learning Nuggets (developed by SEBA Lab Course)

DP in NLP Learning Nugget / Web App Examples

Introduction

- Differential Privacy is a relatively novel concept that boasts a mathematical foundation in quantifying privacy preservation.
- The topic has been largely researched in the context of protection of individuals within a structured dataset.
- When considering Natural Language Processing and its goals, it is necessary to change our view of Differential Privacy.
- This can be accomplished in two ways: one will be discussed here, another way can be found in [d_x-privacy: Generalized Differential Privacy](#).
- It is also useful to view Differential Privacy juxtaposed to other privacy-preserving techniques, especially when coming from the standpoint of privacy in NLP.
- Also important to note is that the study of Differential Privacy's applicability to NLP is an ongoing research topic, and the current state certainly encompasses the infancy stages in this study.

Learning Objectives

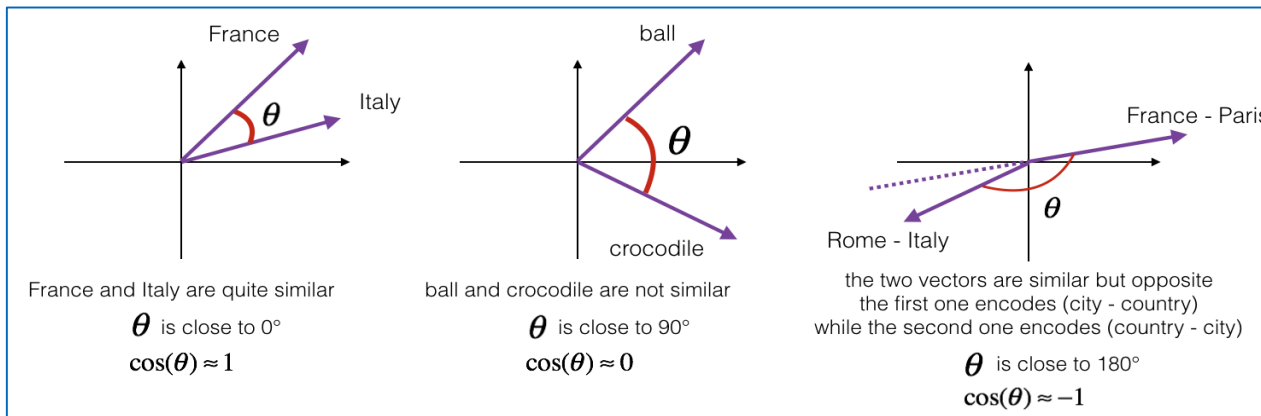
- You will (re-)learn what Differential Privacy is, and what its core concepts and guarantees are
- You will encounter some other privacy-preserving techniques, and realize how they might not be the best option when dealing with unstructured text data
- You will see how this notion of Differential Privacy can be transferred to Natural Language Processing

DP in NLP Learning Nugget Examples (2)

Word Embeddings and their Risks

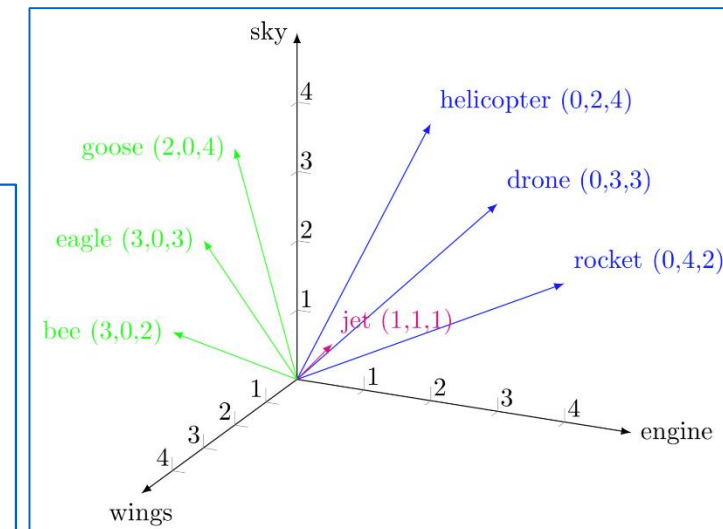


- The growing success of word embeddings for use as *general purpose language models* has increased their attention and utilization.
- To recap, word embeddings:
 - Are a representation of words in vector space, have a set dimension, and are real-valued
 - Provide a convenient basis for computation due to its numerical nature
 - Can be easily used to define semantic similarity (i.e. cosine similarity between vectors)
 - Can be generalized even further to represent sentences or documents
- Recently, word embeddings have been used in high-profile application such as Google's search engine.
- To achieve an accurate language model, these embedding models are usually pre-trained on **billions** of sentences.
 - Naturally, one can imagine that these sentence might contain private or sensitive information.
 - To further exacerbate the issue, embeddings are designed to capture information and relations within text.
- As a result, language models based upon these word embeddings can be at risk.
 - This can especially be true when considering NLP tasks in which sensitive data is used.
 - The real-valued representation of embeddings might be misconstrued as safe, but:
 - We will see in a few slides how these vulnerabilities might be exploited.



Right: a simplified example of word embedding vectors (dim=3),
<https://corpling.hypotheses.org/495>

Left: an explanation of the cosine similarity measure used with word embeddings,
https://datascience-enthusiast.com/DL/Operations_on_word_vectors.html

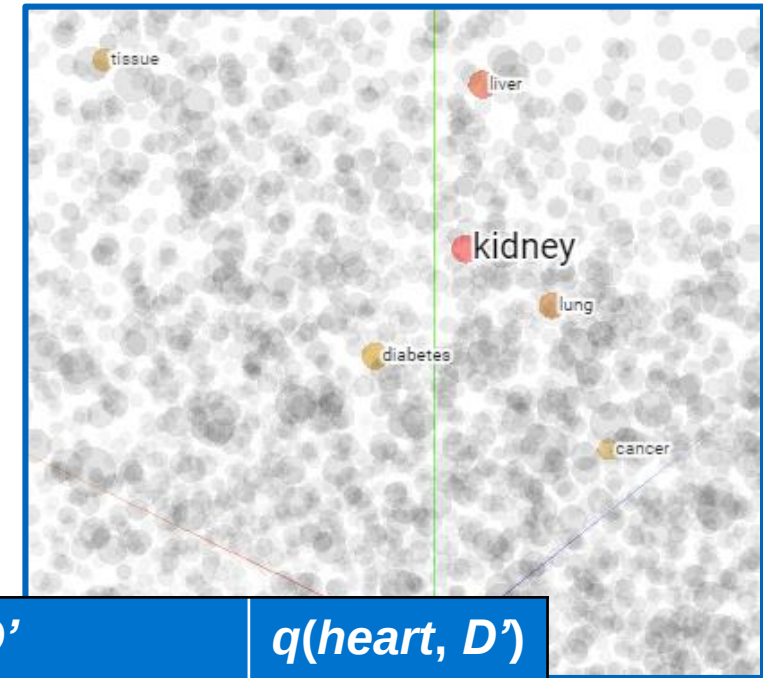


DP in NLP Learning Nugget Examples (3)

The Exponential Mechanism for NLP: an Example



- Let's see a simplified example showing how the Exponential Mechanism can be useful to NLP applications
 - Imagine the following simplified embedding space* (see right):
 - heart* and its 5 nearest neighbors
 - Following the notation:
 - $D = \text{heart}$
 - $\mathcal{R} = \{\text{liver}, \text{lung}, \text{tissue}, \text{diabetes}, \text{cancer}\}$
 - The corresponding scores are given in the table (1 – cosine distance)
 - If we apply the Exponential Mechanism, we obtain the following probabilities:
 - Here, ε is chosen to be 2
 - Sensitivity would be $0.59 - 0.38 = 0.21$
 - Probabilities are obtained by normalization



D'	$\exp(\frac{\varepsilon q(D, r)}{2\Delta q})$	$\text{Pr}[D']$
<i>liver</i>	16.602	0.367
<i>lung</i>	9.647	0.213
<i>tissue</i>	6.560	0.145
<i>diabetes</i>	6.315	0.140
<i>cancer</i>	6.108	0.135

D'	$q(\text{heart}, D')$
<i>liver</i>	0.590
<i>lung</i>	0.476
<i>tissue</i>	0.395
<i>diabetes</i>	0.387
<i>cancer</i>	0.380

*Created from <https://projector.tensorflow.org/>

- The application of Differential Privacy to NLP does not come without its limitations.
- Chief among these limitations are an often-observed utility hit, a general challenge for better explainability about privacy in the textual domain, and the fact that Differential Privacy is only a good answer in specific use cases.
- An important takeaway here is that much like the privacy-utility tradeoff, the need for improved privacy cannot come without sacrifices in other areas – “there is no free lunch”.
- While it is true that some of these limitations can cause concern, one must not forget that it comes as the price of improved privacy preservation in an age where this is becoming more and more crucial*.
- Furthermore, these limitations serve as a great basis for future work in pursuit of better privacy-preserving techniques (for NLP).

Outlook

- If not done already, check out [Benefits of Differential Privacy in NLP](#) for an overview of some of the best reasons for using Differential Privacy in conjunction with NLP techniques
- The limitations discussed here take root in many of the ideas discussed within this unit, particularly [Privacy Vulnerabilities in Natural Language Processing](#), [Differential Privacy in Natural Language Processing](#), [d-privacy: Generalized Differential Privacy](#), [Differential Privacy: Applications to NLP](#) and [d-privacy: Applications to NLP](#)

*<https://www.forbes.com/sites/marymeehan/2019/11/26/data-privacy-will-be-the-most-important-issue-in-the-next-decade/>

DP in NLP Web App Examples (1)



Dashboard

Learning Nuggets

Learning Paths

Bookmarks

Glossary

Lab

Create Nugget

Create Learning Path

Bookmarks

Search

HS Hannes Simon

Apply Filters

Time to Complete

0 Minutes35 Minutes

Categories

☐ Motivation

☐ Foundations

☐ Application

☐ Outlook

Time to Complete

Search for keywords

Privacy Vulnerabilities in Natural Language Processing

The proliferation of Machine Learning techniques and their many success stories has seen an unprecedented amount of data being generated and later used in a variety of tasks. This is without a doubt also true for Natural Language Processing (NLP). At the heart of much of the success in state-of-the-art NLP is the concept of word embeddings, which provide a mathematical basis for the processing of textual data in downstream tasks. With these recent advances also comes privacy issues and vulnerabilities, which have been brought to light by recent research. In this unit, some of these vulnerabilities, including formal attacks that have been shared in the literature. In addition, some potential situations in which these vulnerabilities may occur will be portrayed. Ultimately, these privacy vulnerabilities will demonstrate the need to consider applying privacy-preserving methods to NLP techniques.

Completed

15min

Differential Privacy in Natural Language Processing

Differential Privacy is a relatively novel concept that boasts a mathematical foundation in quantifying privacy preservation. The topic has been largely researched in the context of protection of individuals within a structured dataset. When considering Natural Language Processing and its goals, it is necessary to change our view of Differential Privacy. This can be accomplished in two ways: one will be discussed here, another way can be found in dx-privacy: Generalized Differential Privacy. It is also useful to view Differential Privacy juxtaposed to other privacy-preserving techniques, especially when coming from the standpoint of privacy in NLP. Also important to note is that the study of Differential Privacy's applicability to NLP is an ongoing research topic, and the current state certainly encompasses the infancy stages in this study.

13min

d-privacy: Generalized Differential Privacy

As seen in Differential Privacy in Natural Language Processing, the idea of Differential Privacy can be very useful for the domain of textual data as well, yet in order to achieve this notion of privacy, one must reason a bit further than standard Differential Privacy allows. One initial way of doing this knowledge transfer involves adapting what it means for different pieces of text to be adjacent, thus allowing for a proper fitting to the definition of Differential Privacy. From this arises a new formulation of Differential Privacy called dx-privacy (also simply d-privacy), which maintains the original mathematical basis, but makes the necessary changes for applications existing in metric spaces. As will be discussed, this generalized form of Differential Privacy may be the key to better and more explainable reasoning about Differential Privacy in Natural Language Processing, that is when dealing with text-based data.

20min

Differential Privacy: Applications to NLP

In Differential Privacy in Natural Language Processing, the idea that Differential Privacy is in some way transferable to Natural Language Processing was introduced. In particular, one alternate approach (way of thinking) to Differential Privacy was provided, in which documents become the "databases" in the traditional sense of Differential Privacy. This way of viewing Differential Privacy maintains the standard definition, while altering a bit the way about which textual data is reasoned. To make this thought process a bit clearer, the next few slides will provide concrete applications that have surfaced in research literature, which all use Differential Privacy as a basis to achieve privacy in text processing, i.e. in NLP tasks.

20min

d-privacy: Applications to NLP

In dx-privacy: Generalized Differential Privacy, an more general form of Differential Privacy was introduced, which was created to adapt the privacy guarantee established by standard Differential Privacy, but adapt it to fit in metric spaces. This newer notion provides the opportunity for new (possibly better) reasoning about privacy in the realm of textual data, which lies at the center of Natural Language Processing. As in Differential Privacy: Applications to NLP, the next few slides will provide concrete examples of some important implementations of d-privacy in the area of NLP. Each of these applications comes with some unique aspects, especially in how metrics and their underlying spaces are utilized. All in all, these examples will illustrate how d-privacy enables a new outlook on privacy preservation in NLP.

25min

The Exponential Mechanism for NLP

Previously, the Laplace Mechanism was introduced as a natural way to design differentially private algorithms, by way of adding noise to the results of some particular query. This method is very intuitive and provides a straightforward way to achieve privacy in many applications of standard Differential Privacy. In other cases, though, the application or implementation we are talking about does not exactly produce a simple numeric answer to a query on a database. This is especially true in NLP, which operates primarily on textual data. In Differential Privacy: Applications to NLP, a few specific ways of applying the Exponential Mechanism to achieve the same end goal – add noise in a careful way to achieve a differentially private algorithm – the slightly nuanced mechanism turns out to be a quite useful tool to Differential Privacy in NLP, and its basic function will be the topic of the next few slides.

Completed

20min

Benefits of Differential Privacy in NLP

So far, Differential Privacy in the realm of NLP has been recognized as a viable and potentially useful concept to approach the protection of privacy when dealing with textual data. A couple of similar notions were introduced, namely standard Differential Privacy (Differential Privacy in Natural Language Processing) and d-privacy (dx-privacy: Generalized Differential Privacy). Using these concepts as a foundations, multiple implementations from recent research literature were introduced. From these applications to NLP, we can extract some observed benefits that come as a result of incorporating Differential Privacy (in some way) when tackling NLP tasks. Many benefits root themselves in the advantages of Differential Privacy itself, while others pertain specifically to the text domain.

30min

Limitations of Differential Privacy in NLP

The study of the usage of Differential Privacy would certainly not be complete without an analysis of its limitation, which also includes some of the major challenges. Maybe it already has become clear from Differential Privacy: Applications to NLP and d-privacy: Applications to NLP that current solutions to applying Differential Privacy in conjunction with NLP are definitely not without some shortcomings. Some of these limitations will be briefly introduced here, and they should be taken into consideration along with the Benefits of Differential Privacy in NLP.

15min

210322 Meisenbacher Guided Research Final Presentation

© sebis 27

Dashboard

Learning Nuggets

Learning Paths

Bookmarks

Glossary

Lab

Create Nugget

Create Learning Path

Learning Nuggets

Benefits of Differential Privacy in NLP

So far, Differential Privacy in the realm of NLP has been recognized as a viable and potentially useful concept to approach the protection of privacy when dealing with textual data. A couple of similar notions were introduced, namely standard Differential Privacy (Differential Privacy in Natural Language Processing) and d-privacy (dx-privacy: Generalized Differential Privacy). Using these concepts as a foundations, multiple implementations from recent research literature were introduced. From these applications to NLP, we can extract some observed benefits that come as a result of incorporating Differential Privacy (in some way) when tackling NLP tasks. Many benefits root themselves in the advantages of Differential Privacy itself, while others pertain specifically to the text domain.

Keywords:

Promising Privacy Preservation

- In terms of preserving privacy, Differential Privacy has reported some very promising results, as reported in the corresponding literature.
- Here, a few selected sets of results will be discussed, but for a broader overview of the experimental results, it is highly recommended to look over the original papers – (see the Sources sections from this unit).
- The privacy achieved by a model is evaluated in a variety of ways, through privacy experiments.
- Implementation: SynTF [5]
- Compared privacy preservation vs. classic scrubbing methods (via libraries)
- Using an attack scenario as a baseline, SynTF lowers the attacker's F1 from 90% to 66% (vs. scrubbing).
- Implementation: ER-AE [2]
- Compared against the accuracy of state-of-the-art authorship identification neural network
- ER-AE lowers the network's accuracy to roughly 6%, thus effectively rendering it useless.
- Implementation: DPTxt [1]
- Compared F1 of (1) Sentiment Analysis and (2) POS Tagging
- Also observed attacker F1 score on age, location, and gender attributes
- DPTxt caused the lowest F1 scores, by a significant amount, on all private attributes.
- Implementation: text perturbation with Mahalanobis metric [6] – compared vs. Laplace (Euclidean) mechanism
- Nw = probability of word not getting redacted (lower the better)
- Sw = number of distinct words with probability greater than ϵ of being outputted (higher the better)

Nw 5th percentile

Nw 50th percentile

Nw 95th percentile

Sw 5th percentile

Sw 50th percentile

Sw 95th percentile

Laplace ($\lambda=0$)

$\lambda=0.25$

$\lambda=0.5$

$\lambda=0.75$

Mahalanobis ($\lambda=1$)

<

1

2

3

4

5

6

>

DP in NLP Web App Examples (3)



Dashboard

Learning Nuggets

Learning Paths

Bookmarks

Glossary

Lab

Create Nugget

Create Learning Path

Learning Nuggets

Differential Privacy in Natural Language Processing

Differential Privacy is a relatively novel concept that boasts a mathematical foundation in quantifying privacy preservation. The topic has been largely researched in the context of protection of individuals within a structured dataset. When considering Natural Language Processing and its goals, it is necessary to change our view of Differential Privacy. This can be accomplished in two ways: one will be discussed here, another way can be found in dx-privacy: Generalized Differential Privacy. It is also useful to view Differential Privacy juxtaposed to other privacy-preserving techniques, especially when coming from the standpoint of privacy in NLP. Also important to note is that the study of Differential Privacy's applicability to NLP is an ongoing research topic, and the current state certainly encompasses the infancy stages in this study.

Keywords:

Key Takeaways

- Many privacy-preserving methods are not readily equipped to handle unstructured data like text.
- Differential Privacy arose as a way to provide a mathematically founded basis for privacy-preservation in dataset, particularly in the analysis of this data.
- Even so – Differential Privacy in the original sense is also not instantly transferable to the domain of textual data.
- With some slight modifications in the main concept behind Differential Privacy, one can begin to reason about quantifying privacy in NLP tasks, while gaining the guarantees and advantages that Differential Privacy offers.
- The idea of differentially private NLP requires some extra work beyond the standard approach, and exactly how it can be achieved lies in the implementation details (see below).

Outlook

- Here, one way of adapting standard Differential Privacy to NLP was introduced. For another, more involved development to this point, refer to dx-privacy (Generalized Differential Privacy) to learn more
- To find out some of the reasons why privacy-preserving NLP might be needed in the first place, consider looking at Privacy Vulnerabilities in Natural Language Processing
- The intuition of adapting Differential Privacy to NLP was discussed here, but for specific implementations, check out Differential Privacy: Applications to NLP
- Of course, making this shift to applying Differential Privacy to NLP does not come without some implications: for the pros and cons, refer to Benefits of Differential Privacy in NLP and Limitations of Differential Privacy in NLP

<

1

2

3

4

>

FINISH

- With the amount and nature of textual data nowadays, there are bound to be some risks in NLP
- DP + NLP is initially a somewhat challenging match, but:
 - When reasoned about, serves a great candidate for privacy preservation
 - In some respects, the “best” to date
- Several very interesting aspects:
 - How to represent text
 - How to define certain concepts, i.e. the “individual”, adjacency
 - In what ways can noise be efficiently added (to the pipeline)
- Many possible future directions
 - Still a young, budding, somewhat theoretical field
- Will DP + NLP become commonplace?
 - Experts: hopefully!
 - My opinion: only a matter of time
- In the end: much learned, new appreciation for the field (+ new take on NLP in general)



Stephen Meisenbacher

stephen.meisenbacher@tum.de

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132

Fax +49.89.289.17136

matthes@in.tum.de
wwwmatthes.in.tum.de

