Faculté des arts et des sciences
Département d'informatique et de recherche opérationnelle

Université de Montréal

Evaluating LSA Sensibility to Disclosure in Learners' Interactions

Thursday, March-17-16

OUTLINE

Université de Montréal

- Introduction
- Related Work
- LSA-based method for privacy preserving interactions
- Testing and Validation
- Conclusions and Future Work

OUTLINE

Université de Montréal

- **Introduction**
- Related Work
- LSA-based method for privacy preserving interactions
- Testing and Validation
- Conclusions and Future Work

INTRODUCTION

Université de Montréal

- Today's distant learning environment
  - problem: lack of face-to-face interactions
  - solution: collaborative strategies and social interactions tools (Puustinen et *al.*, 2015)

- Example of collaborative strategies: *Peer feedback*

  soliciting co-learners to provide feedback in response to learner request (Pridmore and Overocker, 2014)

- decrease learners' social isolation
- encourage motivation and engagement in learning activities

INTRODUCTION

Université de Montréal

- Peers feedback in educational context (Zhao et *al.*, 2012)
  - in form of corrections, opinions, suggestions, ideas, etc.
  - different types, mainly two:
    - cognitive
      - context independent
      - targeting work content
      - example: peer assessment in writing
    - affective
      - context dependent
      - using affective language (praising, criticism)
      - targeting individual performance

INTRODUCTION

Université de Montréal

- Example of peer affective feedback posted on English forum discussion:

INTRODUCTION

Université de Montréal

- Example of peer affective feedback posted on English forum discussion:

INTRODUCTION

Université de Montréal

- Example of peer affective feedback posted on English forum discussion:

INTRODUCTION

Université de Montréal

- Example of peer affective feedback posted on English forum discussion:

INTRODUCTION

Université de Montréal

- Learners involved in interaction process (Zhao et *al.*, 2012)

- express themselves freely
- share personal experiences
- disclose information about themselves to others (sometimes *unwillingly*)

- Example of a learner request posted on an English forum discussion:

> *"I am 22 years old, engineering student from India and my family cannot speak English and I am feeling bad to speak with my American girlfriend because I could be wrong…"*

INTRODUCTION

Université de Montréal

- Personal information disclosed

  - age: 22 years
  - origin: Indian
  - education qualification: student engineer
  - relationship status: American girlfriend

- Privacy threats, if personal data is exposed, or misused by abusive users (Lee et *al.*, 2013)
  - *psychological damage* (*cyber-bullying: origin, race, religion, etc.)*
  - *social and financial damage (identity theft, or impersonation, etc.)*

INTRODUCTION

Université de Montréal

- Consequences of personal information disclosure in *educational context* (Puustinen et *al.*, 2015)
  - *unsafe* e-learning environment
  - *abandon of learning*

- What is needed in this context?
  - *scrutinize* learners' interactions
  - *detect* and *minimize* disclosure of personal data
  - create *favorable* learning environment
  - *protect* users from privacy risks

INTRODUCTION

Université de Montréal

- Existing work
  - lack of research on affective feedback in educational contexts
  - no solution for self-disclosure risks in educational contexts
  - no solution for disclosure in natural language interactions
    - Natural language tasks in educational context:
      - Southavilay et *al.*, 2013: analysis of collaborative writing processes evolution
      - Nye et *al.*, 2014: evaluation of students' answers in Intelligent Tutoring System
      - Selmi et *al.*, 2014: semantic analysis for privacy preserving peer feedback

OUTLINE

Université de Montréal

- Introduction
- **Related Work**
- LSA-based method for privacy preserving interactions
- Testing and Validation
- Conclusions and Future Work

Related work

Université de Montréal

A. Semantic Analysis tasks and techniques

- support collaborative activities
  - examine discussed topics between students in discussion forum (Despotakis et *al.*, 2013)
  - visualize and analyse collaborative writing process by extracting semantic topics associated to its evolution (Southavilay et *al.*, 2013)
- automatic evaluation of students' responses
  - compare responses to predefined model by examining the differences between semantic vectors of responses and model (Nye et *al.*, 2014)
- protect students' privacy
  - explore semantics of students' interactions
  - discard negative feedback using Latent Semantic Analysis (Selmi et *al.*, 2014)

Related work

Université de Montréal

B. LSA: solution for self-disclosure

- recognizing personal self disclosed data
- hiding or modifying disclosed data
- stripping sentences revealing personal information

- Latent Semantic Analysis (LSA)
  - technique of vectorial semantics
  - patented in 1988 by Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum and Lynn Streeter (Deerwester et *al.*, 1990)
  - called Latent Semantic Indexing (LSI) in information retrieval context

LATENT SEMANTIC ANALYSIS

Université de Montréal

- Technique for extracting and representing the contextual-usage meaning of words based on statistical computations applied to a large corpus of text
- LSA is used in:
  - data clustering and text classification
  - cross language retrieval
  - text summary
  - questions answering systems
- LSA is based on 3 steps :
  - *occurrence matrix construction*
  - *matrix decomposition*
  - *dimensionality reduction* (low-rank approximation)

LATENT SEMANTIC ANALYSIS

Université de Montréal

- **First step in LSA: occurrence matrix construction**
  - input matrix A representing text of peers' provided feedback
    - columns: sentences of feedback
    - rows: terms appearing in feedback

$$A = \begin{matrix} d_j \rightarrow \\ \\ t_i^T \rightarrow \end{matrix} \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix}$$

LATENT SEMANTIC ANALYSIS

Université de Montréal

- **Example:**

Request : « *Hello. I am from Georgia and I want speak English. Who can help me to practice my English?? Can you for me some advice?* »

- **Feedback 1 :** « *I would like to practice English with you. Please add me on skype. My skype id is ***** »
- **Feedback 2 :** « *No pain... no gain* »
- **Feedback 3 :** « *I will study English with u every day* »

LATENT SEMANTIC ANALYSIS

Université de Montréal

- **Example:** Occurrence matrix construction

	Request	Feedback1	Feedback 2	Feedback 3
Georgia	1	0	0	0
want	1	0	0	0
speak	1	0	0	0
English	1	1	0	1
help	1	0	0	0
practice	1	1	0	0
like	0	1	0	0
add	0	1	0	0
Skype	0	1	0	0
pain	0	0	1	0
gain	0	0	1	0
study	0	0	0	1

LATENT SEMANTIC ANALYSIS

Université de Montréal

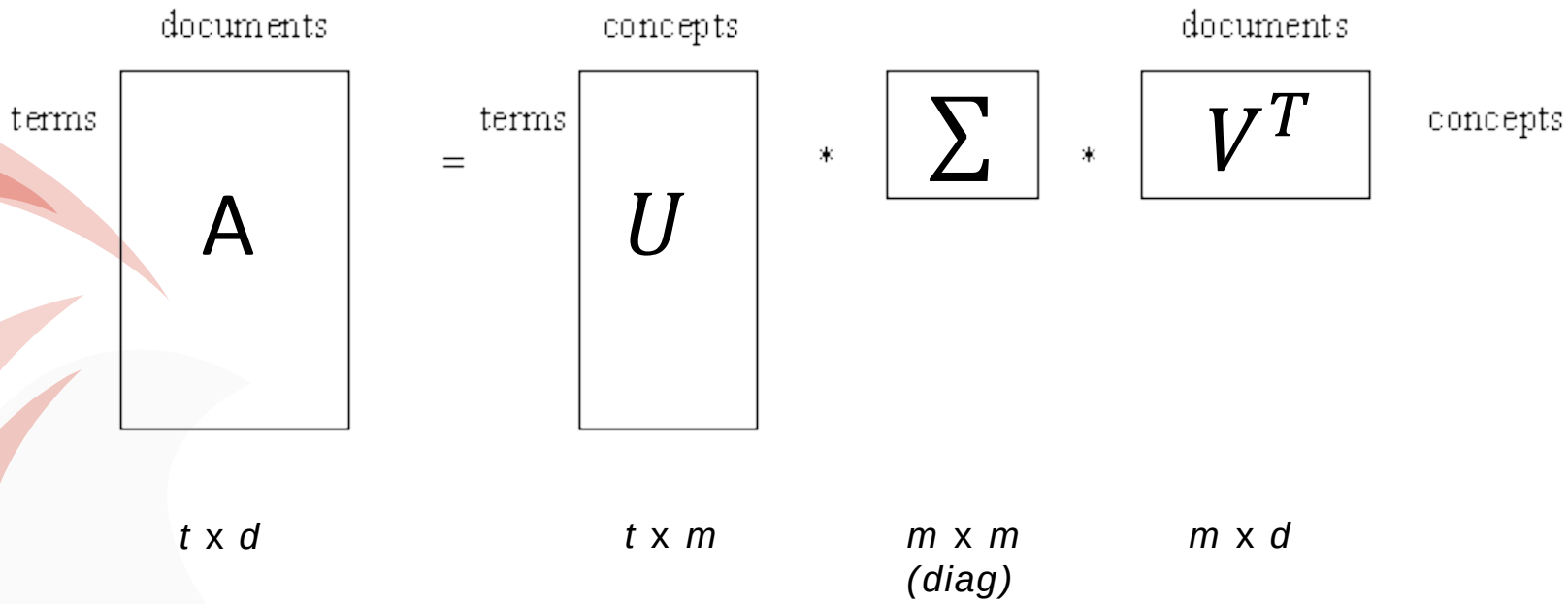
- **Second step in LSA: matrix decomposition**
  - applying a factorization method called Singular Value Decomposition (SVD) to derive latent semantic structure (Deerwester *et al.*, 1990)
  - decomposing matrix A into 3 matrices

$$A = U \sum V^T$$

A: input matrix with dimensions *t* x *d*
U: t x *m* matrix of extracted topics or concepts (columns)
Σ: m x *m* diagonal matrix containing scaling values sorted in descending order
V: m x *d* matrix of extracted concepts from the provided feedback (rows)

LATENT SEMANTIC ANALYSIS

Université de Montréal



LATENT SEMANTIC ANALYSIS

Université de Montréal

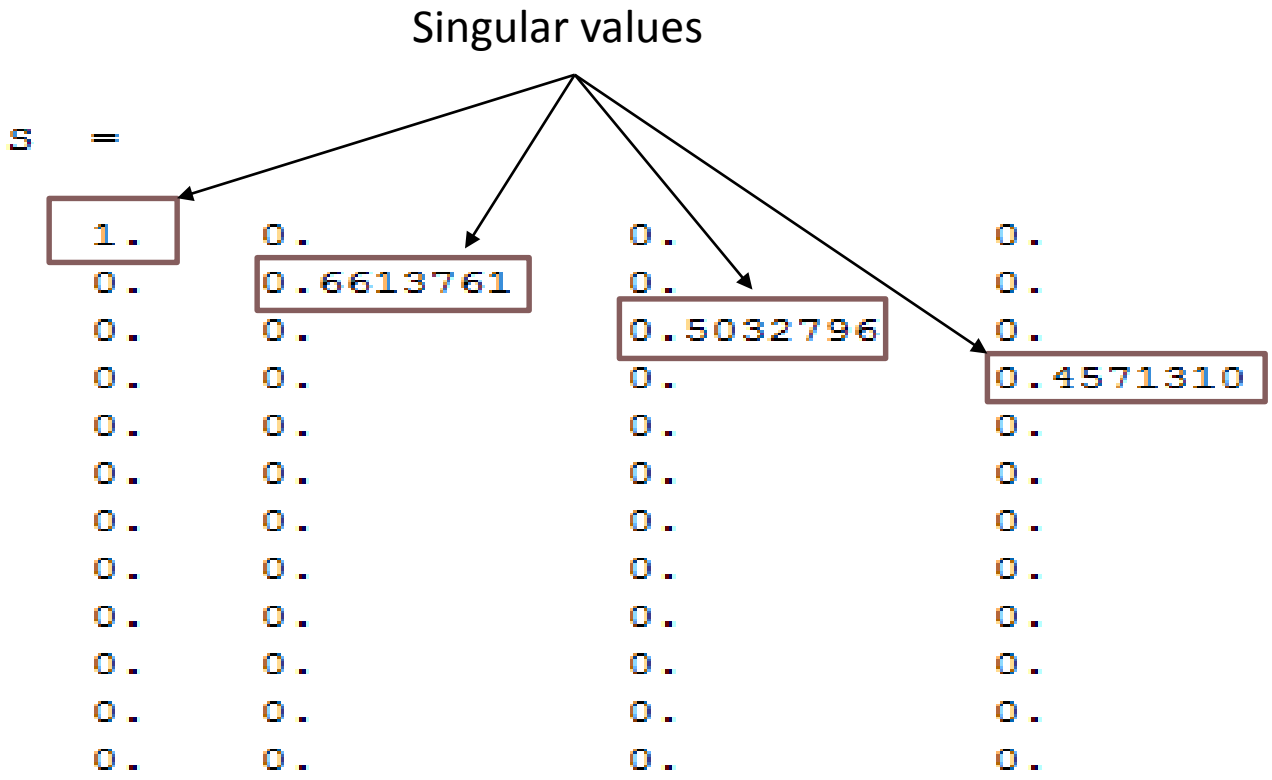
- **Example:** Occurrence matrix A

	Request	Feedback1	Feedback 2	Feedback 3
Georgia	1	0	0	0
want	1	0	0	0
speak	1	0	0	0
English	1	1	0	1
help	1	0	0	0
practice	1	1	0	0
like	0	1	0	0
add	0	1	0	0
Skype	0	1	0	0
pain	0	0	1	0
gain	0	0	1	0
study	0	0	0	1

LATENT SEMANTIC ANALYSIS

Université de Montréal

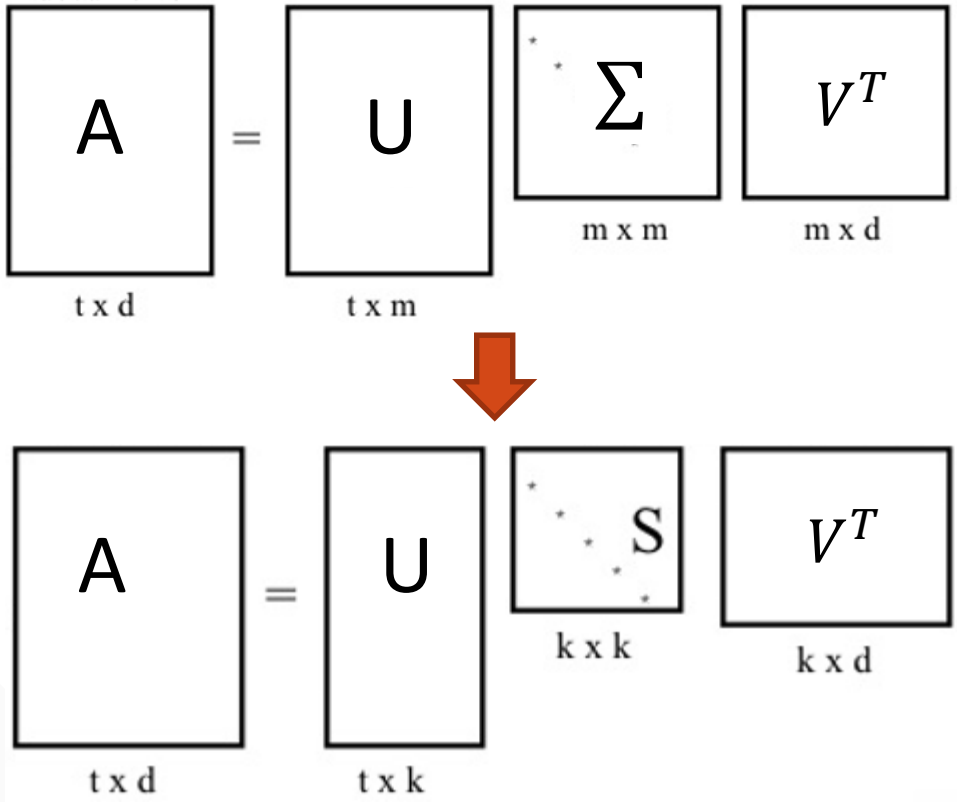
- **Example:** Decomposition of matrix A into 3 matrices U, S and V



LATENT SEMANTIC ANALYSIS

Université de Montréal

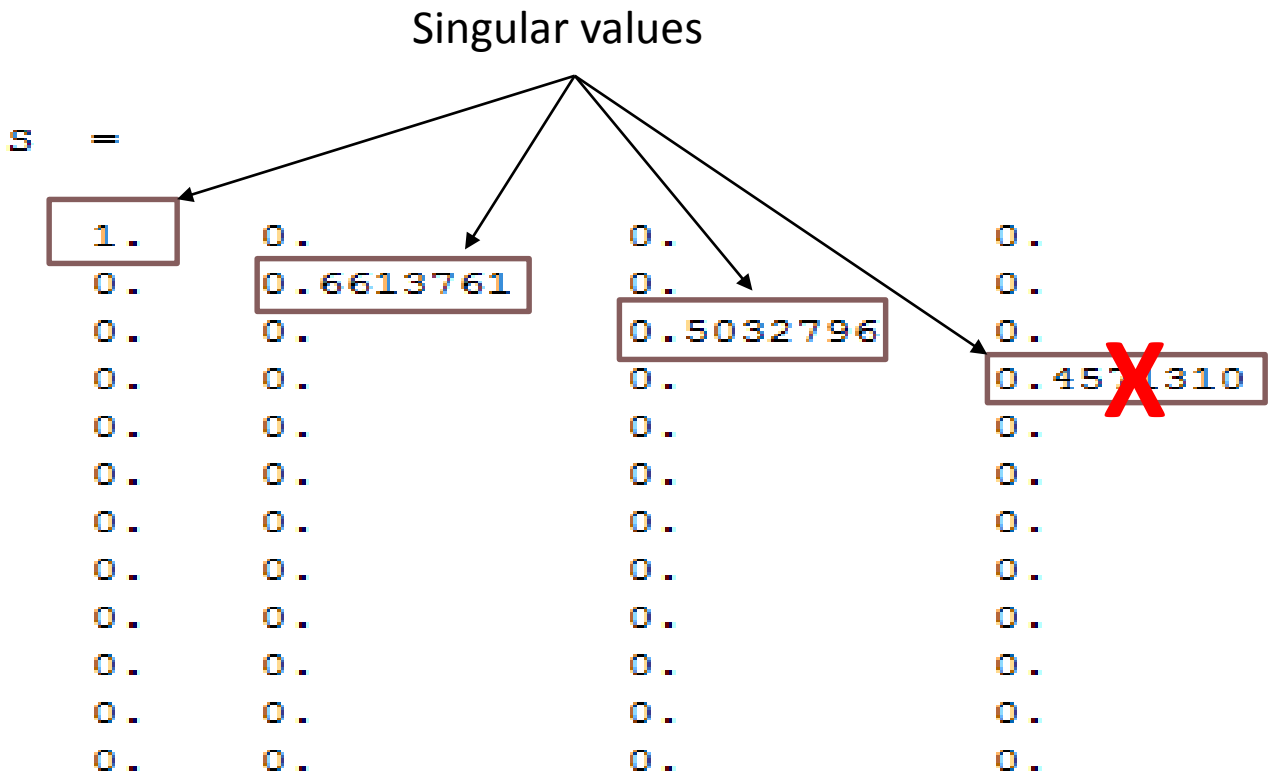
- **Third step in LSA: dimensionality reduction**
  - selelct *k* greatest singular values to construct an approximation A' of matrix A



LATENT SEMANTIC ANALYSIS

Université de Montréal

- **Example:** dimensionality reduction of A with *k*=3



LATENT SEMANTIC ANALYSIS

Université de Montréal

- **Third step in LSA:** dimensionality reduction and construction of the approximation

documents

terms

```
ans  =

    0.3504278   - 0.0094777     0.              0.0426393
    0.3504278   - 0.0094777     0.              0.0426393
    0.3504278   - 0.0094777     0.              0.0426393
    0.3835894     0.4041210     0.              0.1388072
    0.3504278   - 0.0094777     0.              0.0426393
    0.3409500     0.3298962     0.              0.1168641
  - 0.0094777     0.3393739     0.              0.0742248
  - 0.0094777     0.3393739     0.              0.0742248
  - 0.0094777     0.3393739     0.              0.0742248
    4.492D-19     3.612D-17     0.3558724     - 7.045D-17
    4.492D-19     3.612D-17     0.3558724     - 7.045D-17
    0.0426393     0.0742248     0.              0.0219431
```

OUTLINE

Université de Montréal

- Introduction
- Related Work
- **LSA-based method for privacy preserving interactions**
  - Mining Step
  - Composition Step
- Test and Validation
- Conclusions and Future Work

Our Approach

Université de Montréal

- Goals of this work
  - *scrutinize* learners' interactions
  - *detect* and *minimize* disclosure of personal data
  - create *favorable* learning environment
  - *protect* users from privacy risks
- 2-step proposed approach
  - *mining step*
    - discarding negative feedback messages that negatively affect learning
  - *composition step*
    - eliminating any self-disclosing sentences from mined feedback
    - reconstructing new feedback

ARCHITECTURE

Université de Montréal

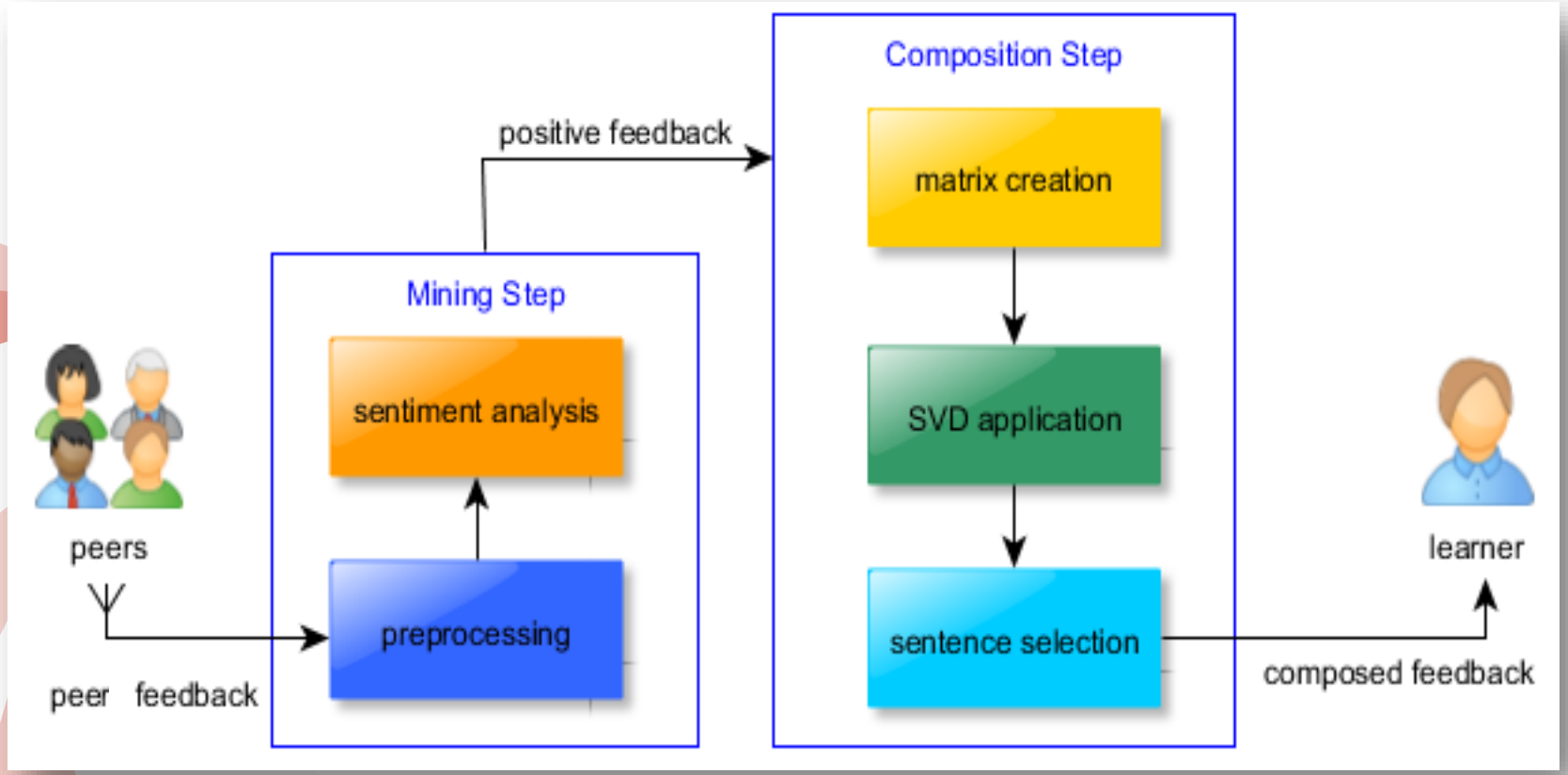


Fig. 1. Architecture of our approach

Our Approach

Université de Montréal

- Goals of this work
  - *scrutinize* learners' interactions
  - *detect* and *minimize* disclosure of personal data
  - create *favorable* learning environment
  - *protect* users from privacy risks

- 2-step proposed approach
  - *mining step*
    - discarding negative feedback messages that negatively affect learning
  - *composition step*
    - eliminating any self-disclosing sentences from mined feedback
    - reconstructing new feedback

ARCHITECTURE

Université de Montréal

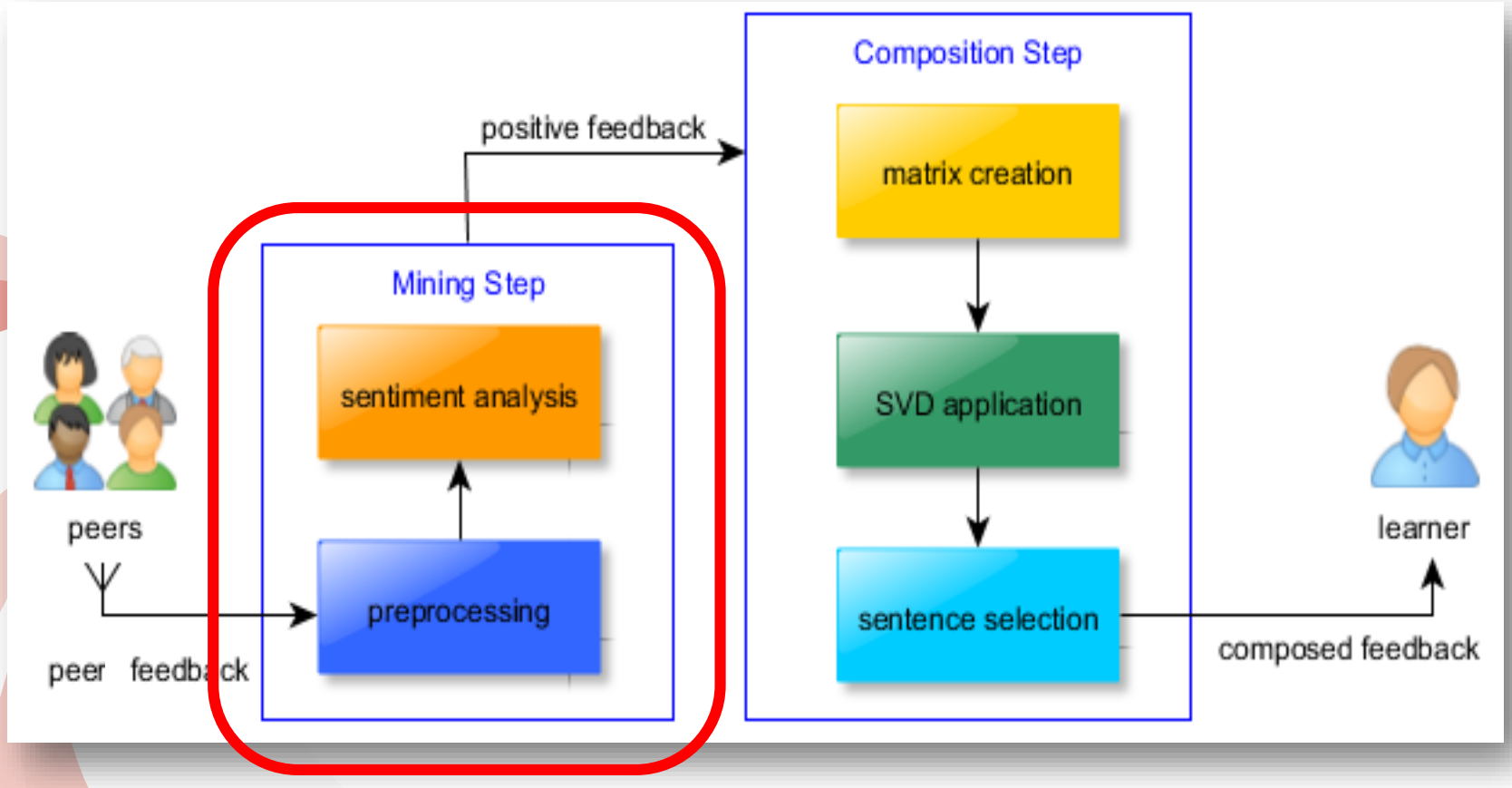


Fig. 1. Architecture of our approach

Our Approach: mining step

Université de Montréal

- Role: discarding negative feedback affecting learning process

  negative feedback: bullying, demeaning, or other negative comments

- Preprocessing: recognizing negative feedback
  - short text document
  - vector of text attribute values (frequency computing)
  - natural language processing techniques
    - *bag of words* as linguistic model
    - no stop words and non-content bearing words
    - no grammatical structural characteristics or positional information

Our Approach: mining step

Université de Montréal

- **Example of feedback message**

  "Whenever i speak to native (English speaker), I feel very frustrated and i'll start to stammer. The phrasing, sentence structure & grammar of my sentences become all in a mess"

- **Preprocessing steps**

  - *tokenization*

  <Whenever, I, speak, to, native, English, speaker, I, feel, very, frustrated, and, i',ll, start, to, stammer, The, phrasing, sentence, structure, &, grammar, of, my, sentences, become, all, in, a, mess>

  - *stop words removal*

  <Speak, native, English, speaker, feel, frustrated, start, stammer, phrasing, sentence, structure, grammar, sentences, become, mess>

Our Approach: mining step

Université de Montréal

- **Example of feedback message**

  "Whenever i speak to native (English speaker), I feel very frustrated and i'll start to stammer. The phrasing, sentence structure & grammar of my sentences become all in a mess"

- **Preprocessing steps**

  - *stemming*: converting token to its morphological format

    <speak, speaker, speaking> → *speak*

  - *frequency computing*: result of preprocessing steps

  <Speak, native, English, speaker, feel, frustrated, start, stammer, phrasing, sentence (2), structure, grammar, become, mess>

Our Approach: mining step

Université de Montréal

- *Sentiment analysis:* classifying feedback as positive and negative
  - Naive Bayes classifier
    - probabilistic classifier based on Bayes' theorem
    - independence assumptions on words' position in text (Pang et *al.*, 2002)
    - for a given set of classes, probability of a class:

$$P(c|d) = P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \qquad (1)$$

c: target class
d: current document
t: current term
n: number of terms in current document

Our Approach: mining step

Université de Montréal

- *Sentiment analysis:* classifying feedback as positive and negative
  - Naive Bayes classifier
    - result: class with the highest probability given the feedback
    - estimation of log probability, given by:

$$\arg\max_{c} \log(\hat{P}(c)) \sum_{1 \le k \le n_d} \log(\hat{P}(t_k|c)) \qquad (2)$$

c: target class
t: current term
n: number of terms in current document

ARCHITECTURE

Université de Montréal

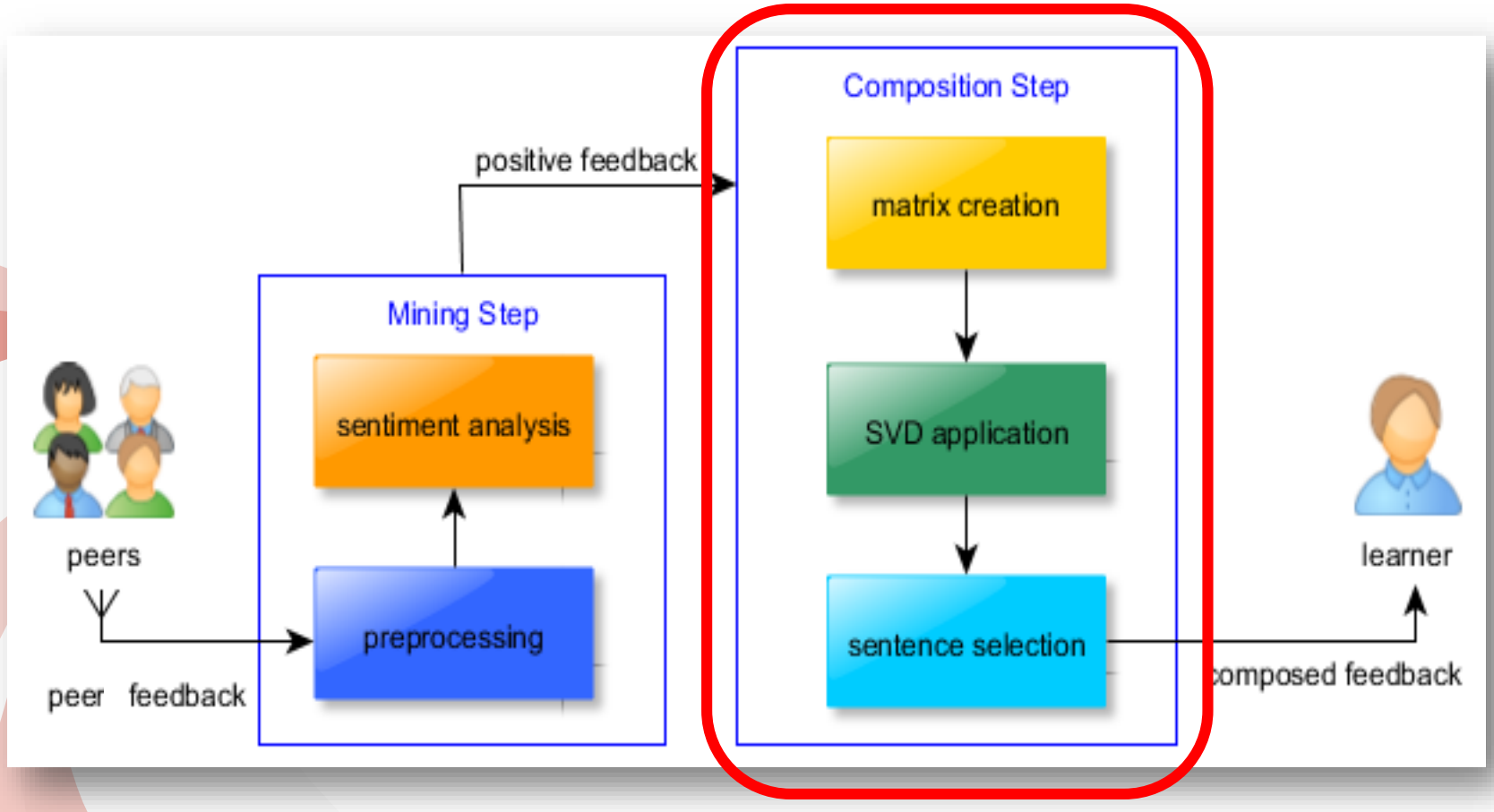


Fig. 1. Architecture of our approach

Our Approach: Composition step

Université de Montréal

- Role: removing any self-disclosure of personal data in interactions
  - detect self disclosing messages
  - preventing message *observer* from gaining knowledge of disclosed personal information

- LSA for privacy protection from self-disclosure in interactions:
  - learner message: set of sentences and concepts
  - sentence: set of terms
  - most representative sentences of current concept: set of terms best representing that concept

Our Approach: Composition step

Université de Montréal

- **LSA in composition step:** we propose to consider
  - three main steps:
    1. *matrix creation*
    2. *SVD application*
    3. *sentences selection*
  - three parameters:
    - *weighting schemes*
    - *approximation rank or number of dimensions*
    - *similarity measure*

ARCHITECTURE

Université de Montréal

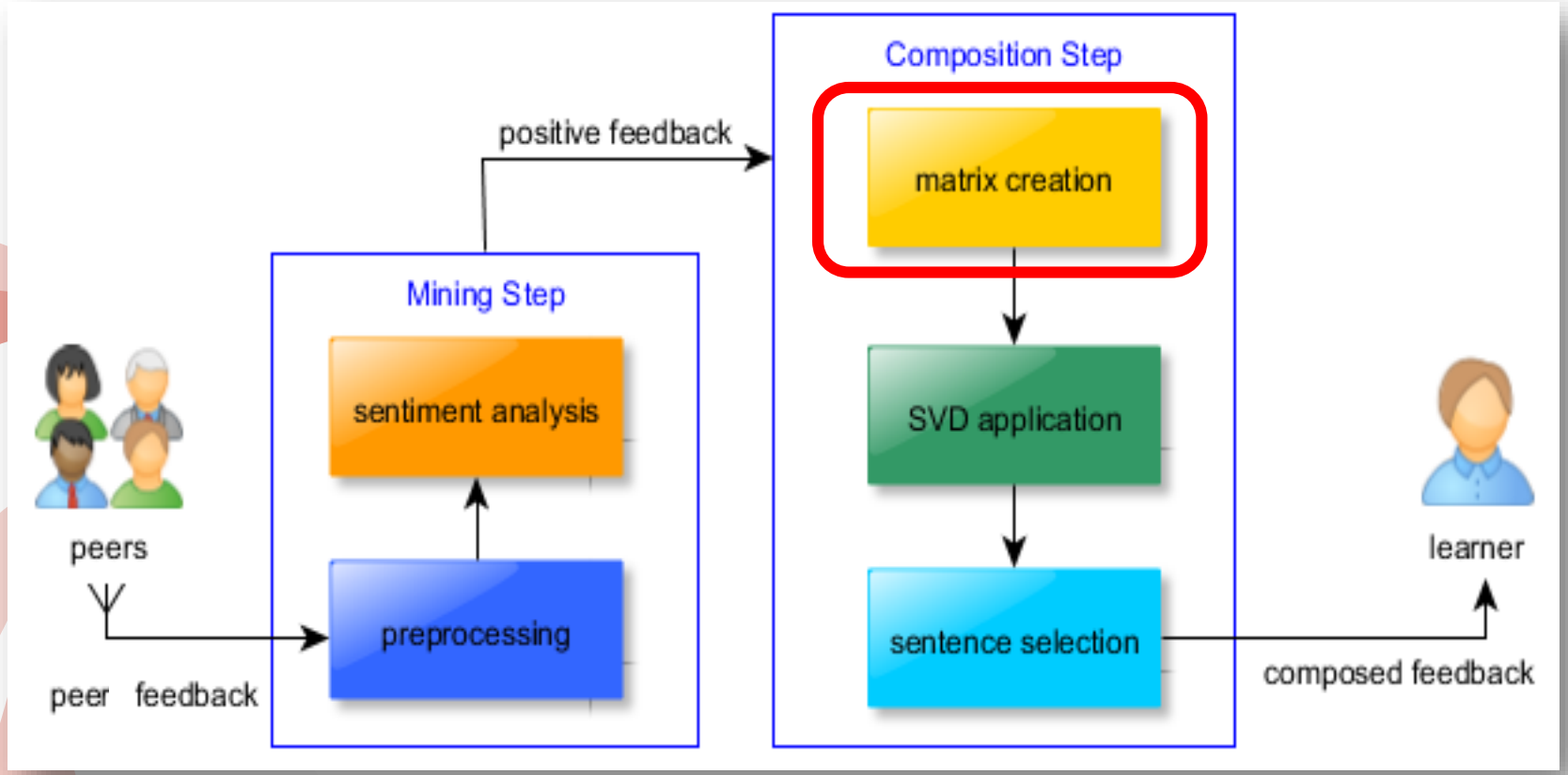


Fig. 1. Architecture of our approach

Our Approach: Composition step

Université de Montréal

- First step in LSA
  - input matrix representing input text message
    - columns: sentences of input text
    - rows: terms appearing in message, extracted using preprocessing tools
  - weighting schemes parameter
    - affecting differently LSA performance depending on data size and nature
    - three weighting categories: local (Term Frequency), global (Inverse Document frequency) and hybrid (Log Entropy, TF-IDF)

Hypothesis 1:

Local weighting schemes are the most appropriate to discard disclosing terms and sentences

Our Approach: Composition step

Université de Montréal

- **Example:** Occurrence matrix A

	Request	Feedback1	Feedback 2	Feedback 3
Georgia	1	0	0	0
want	1	0	0	0
speak	1	0	0	0
English	1	1	0	1
help	1	0	0	0
practice	1	1	0	0
like	0	1	0	0
add	0	1	0	0
Skype	0	1	0	0
pain	0	0	1	0
gain	0	0	1	0
study	0	0	0	1

Our Approach: Composition step

Université de Montréal

- **Example:** TF-IDF weighting computing

  t_1=*English,* R (Request), F1 (Feedback 1), F2(Feedback 2) and F3 (Feedback 3)

  - **Term Frequency (TF)**
    - TF (t_1, R) = occurrences of t_1 / terms number in R
    - TF (t_1, R) = 2/24
  - **Inverse Document Frequency (IDF)**
    - IDF (t_1)= log (number of docs in corpus / number of documents containing t_1)
    - IDF (t_1)= $\log(4/3)$

Our Approach: Composition step

Université de Montréal

- **Example:** TF-IDF weighting computing

  t_1=*English,* R (Request), F1 (Feedback 1), F2(Feedback 2) and F3 (Feedback 3)

  - **TF-IDF computing**
    - TF-IDF (t_1, R) = $\frac{2}{24} * \log\frac{4}{3}$= 0.04
    - TF-IDF (t_1, F1) = $\frac{1}{18} * \log\frac{4}{3}$= 0.04
    - TF-IDF (t_1, F2) = $\frac{0}{4} * \log\frac{4}{3}$= 0
    - TF-IDF (t_1, F3) = $\frac{1}{8} * \log\frac{4}{3}$= 0.06

ARCHITECTURE

Université de Montréal

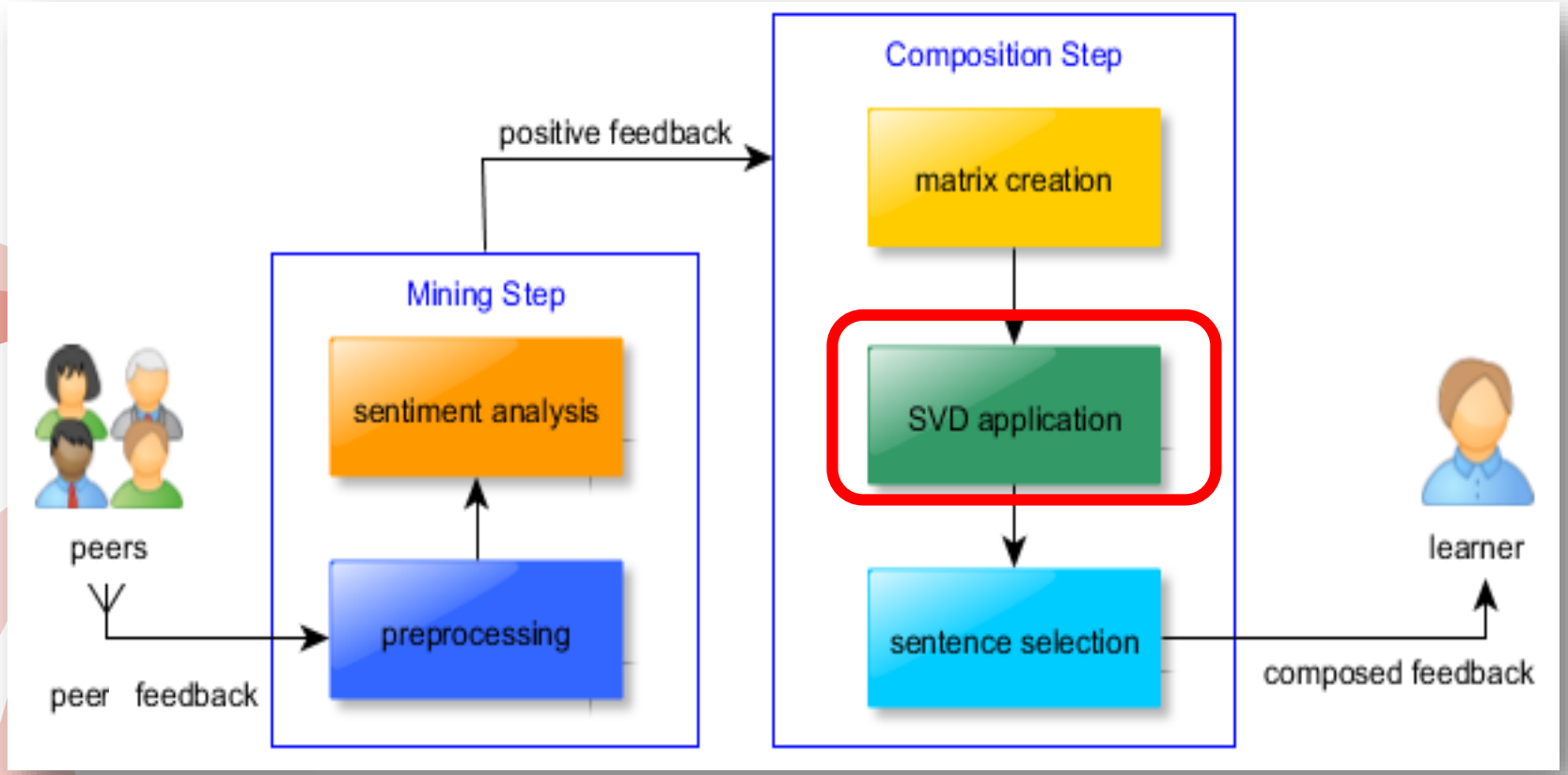


Fig. 1. Architecture of our approach

Our Approach: Composition step

Université de Montréal

- Second step in LSA
  - SVD on matrix A to derive latent semantic structure
  - decomposing A into 3 matrices to extract significant terms and sentences

$$A = U \sum V^T$$

A: input matrix with dimensions *m* x *n*
U: *m* x *n* matrix of extracted topics or concepts (columns)
Σ: *n* x *n* diagonal matrix containing scaling values sorted in descending order
V: *m* x *n* matrix of extracted concepts from the provided feedback (rows)

Our Approach: Composition step

Université de Montréal

- Second step in LSA
  - SVD on matrix A to derive latent semantic structure
  - decomposing A into 3 matrices to extract significant terms and sentences
  - dimensionality parameter
    - reducing dimensions to enhance relationships between two terms or documents
    - different impacts on LSA performance
    - no consensus regarding optimal reduction value (generally used k=300)

Hypothesis 2:

Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences

ARCHITECTURE

Université de Montréal

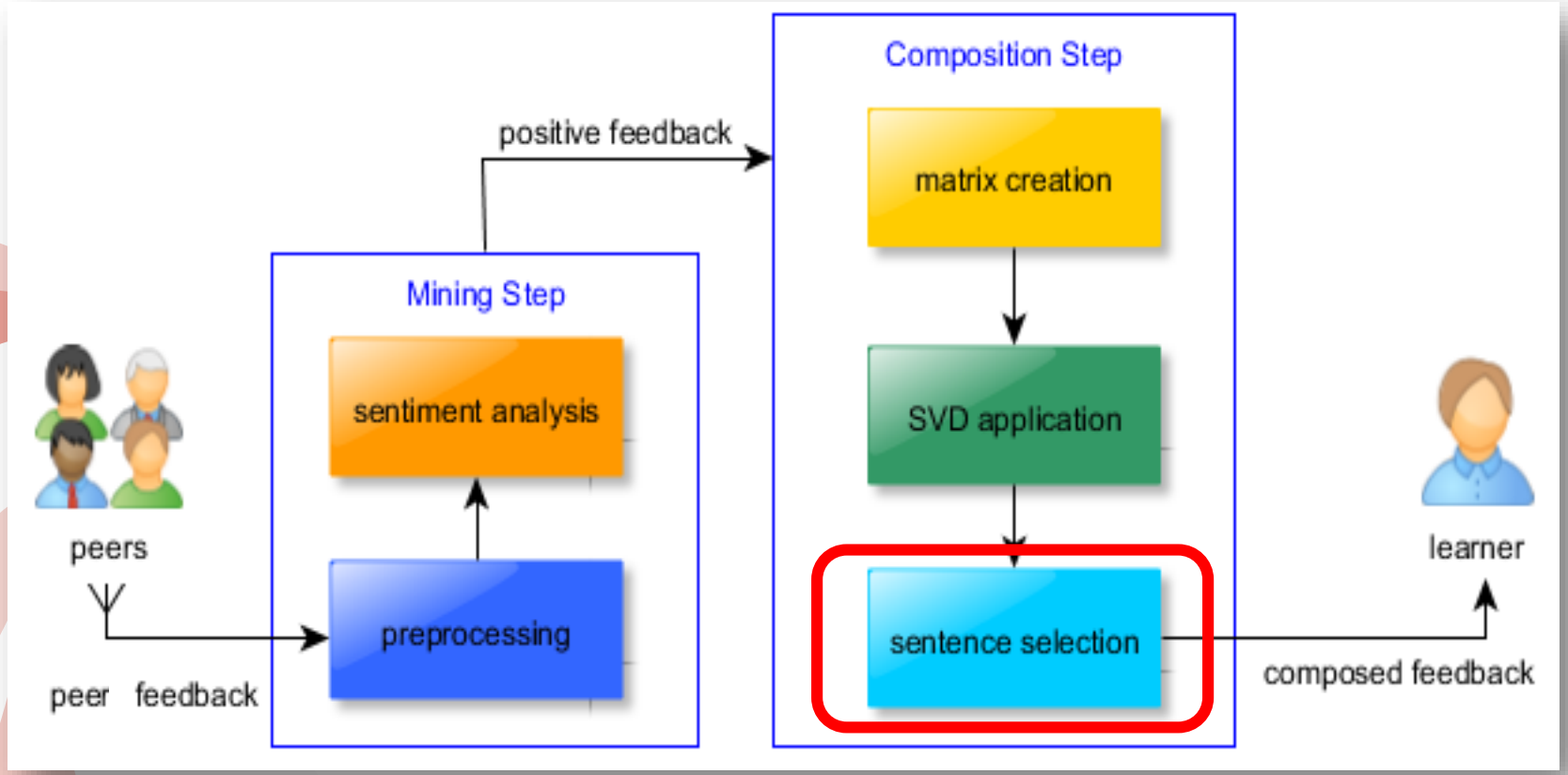


Fig. 1. Architecture of our approach

Our Approach: Composition step

Université de Montréal

- Third step in LSA
  - selecting important sentences in learners messages
    - request initiating interactions
    - all learners messages to the same request
  - similarity metrics parameter
    - computing similarity between two semantic vectors
    - different metrics : distance (Euclidean distance and Jaccard) and similarity (cosine)
    - affecting differently LSA outputs depending on data size and nature

Hypothesis 3:

Distance metrics are more appropriate to discard disclosing terms and sentences

Our Approach: Composition step

Université de Montréal

- **Example:**

Request : « *Hello. I am from Georgia and I want speak English. Who can help me to practice my English?? Can you for me some advice?* »

- **Feedback 1 :** « *I would like to practice English with you. Please add me on skype. My skype id is ***** »
- **Feedback 2 :** « *No pain... no gain* »
- **Feedback 3 :** « *I will study English with u every day* »

Our Approach: Composition step

Université de Montréal

- **Example:** similarity computing using Cosine and Jaccard between Request and Feedback 1

  - **Cosine**

$$\cos(R, F1) = \frac{R.F1}{\|R\|\|F1\|} = 0.36$$

  - **Jaccard**

$$\mathrm{d}(R, F1) = \frac{|R \cap F1|}{|R \cup F1|} = 0.25$$

OUTLINE

Université de Montréal

- Introduction
- Related Work
- LSA-based method for privacy preserving interactions
- **Testing and Validation**
- Conclusions and Future Work

TESTING AND VALIDATION

Université de Montréal

- Corpus
  - over 1000 comments threads to build semantic space
  - forum discussion for English Second Language (ESL)
  - challenge: noisy and informal data

- Data Preprocessing
  - data preprocessed to reduce dimensionality
    - average of 5 peer feedback received for each request
    - short feedback (with less than 3 words) excluded
  - cleansing
    - noise removed: “Hhhhhhhhh gr8” (for great)
    - morphological format to reduce noise and misspelling: “scheduale”, “schedual”, “skedul” →schedul

TESTING AND VALIDATION

Université de Montréal

- Correlations between human judges' scores
  - 4 point Likert scale regarding request: 1 = "very bad" ; 4 = "very good"
  - 2 independent human judges: inter-rater agreement using Cohen kappa (κ=0.68)
    - Correlations between human judges and LSA scores
      - N_words: number of words
      - Feed_Req: LSA cosine(feedback, request)
      - Avg_Feed: average LSA cosine(feedback, all peers' feedback to the request)

Table 1. Correlation between human judges and LSA measures

Measure	Correlation
N_words	0.233
Feed_Req	0.458
Avg_Feed	0.464

TESTING AND VALIDATION

Université de Montréal

- Optimal dimensionality using 2 methods
  - energy: 90% of the information
  - cumulative variance: 60% to 80% of the information
  - tested dimensionality values: 25%, 50%, 70%, 80% and 100%

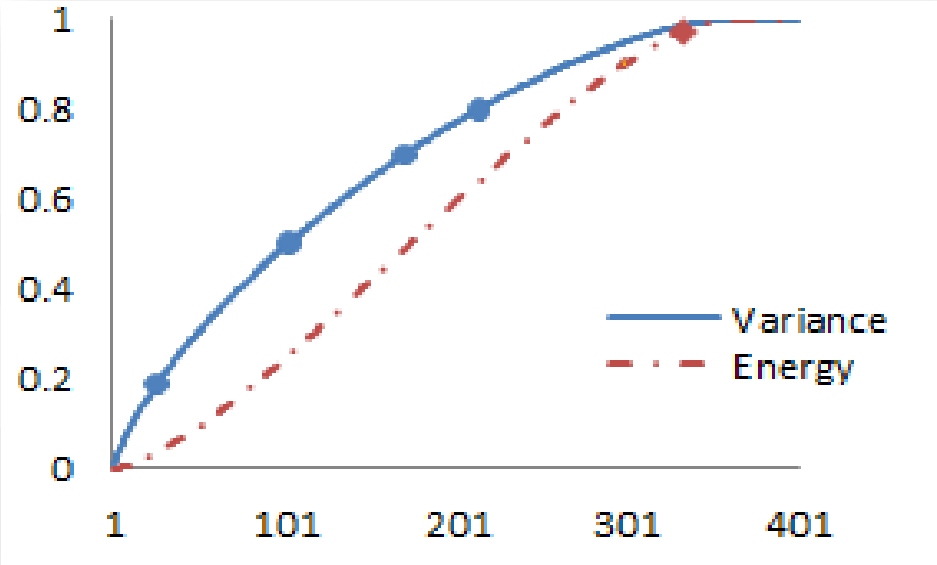


Fig. 2. Energy and variance methods

TESTING AND VALIDATION

Université de Montréal

Hypothesis 1: Weighting

Local weighting schemes are the most appropriate to discard disclosing terms and sentences

Hypothesis 2: Dimensionality

Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences

Hypothesis 3: Similarity metrics

Distance metrics are more appropriate to discard disclosing terms and sentences

TESTING AND VALIDATION

Université de Montréal

Hypothesis 1: Weighting

Local weighting schemes are the most appropriate to discard disclosing terms and sentences

Hypothesis 2: Dimensionality

Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences

Hypothesis 3: Similarity metrics

Distance metrics are more appropriate to discard disclosing terms and sentences

TESTING AND VALIDATION

Université de Montréal

- Impact of weighting on dimensionality
  - local scheme: Binary Term Frequency
  - hybrid schemes: Term Frequency-Inverse Document Frequency (TF-IDF) and Log Entropy
  - best performance: no-weight (less variance)

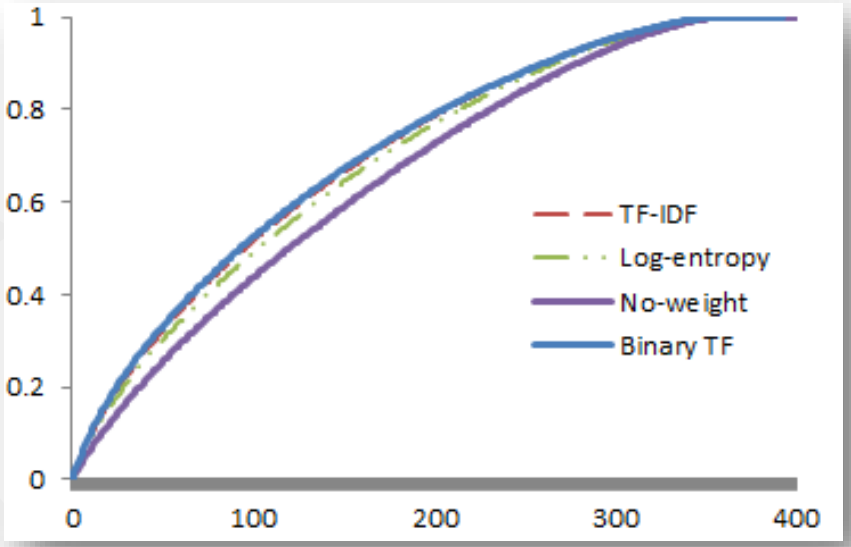


Fig. 3. Impact of weighting on dimensionality

TESTING AND VALIDATION

Université de Montréal

- Impact of dimensionality and weighting on LSA
  - multiple regression analysis
    - variables: independent (Feed_Req, Avg_Feed), dependent (average human judges score)
    - similarity measure: cosine
    - best performance with Log Entropy and no dimensionality reduction

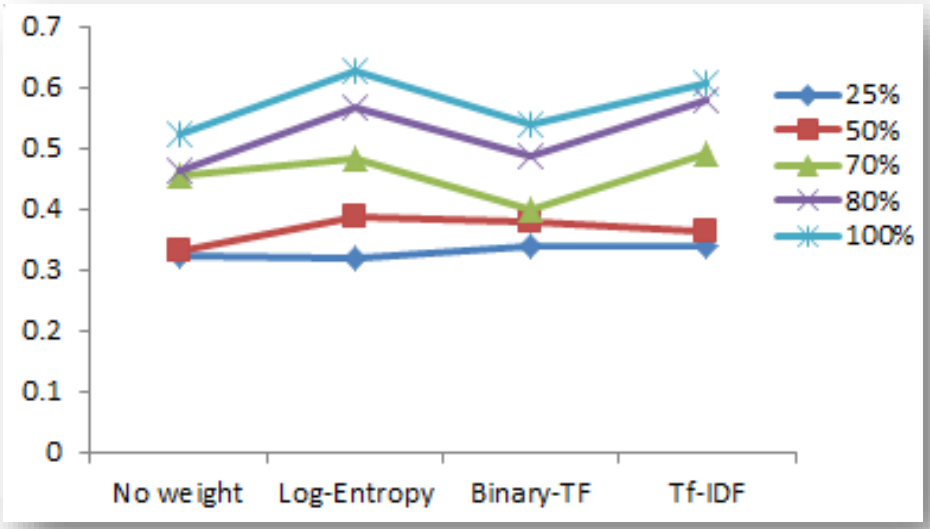


Fig. 4. Correlations with human graders: Interaction between dimensionality and weighting

TESTING AND VALIDATION

Université de Montréal

Hypothesis 1: Weighting

Local weighting schemes are the most appropriate to discard disclosing terms and sentences

Hypothesis 2: Dimensionality

Dimensionality reduction positively affect LSA ability to discard disclosing terms and sentences

Hypothesis 3: Similarity metrics

Distance metrics are more appropriate to discard disclosing terms and sentences

TESTING AND VALIDATION

Université de Montréal

- Impact of dimensionality and similarity metrics on LSA
  - correlation with human graders
    - $r = 0.64$, $p < 0.001$
    - similarity metrics: Euclidean distance, Jaccard and Cosine
    - best performance model: no reduction, Euclidean distance and TF-IDF

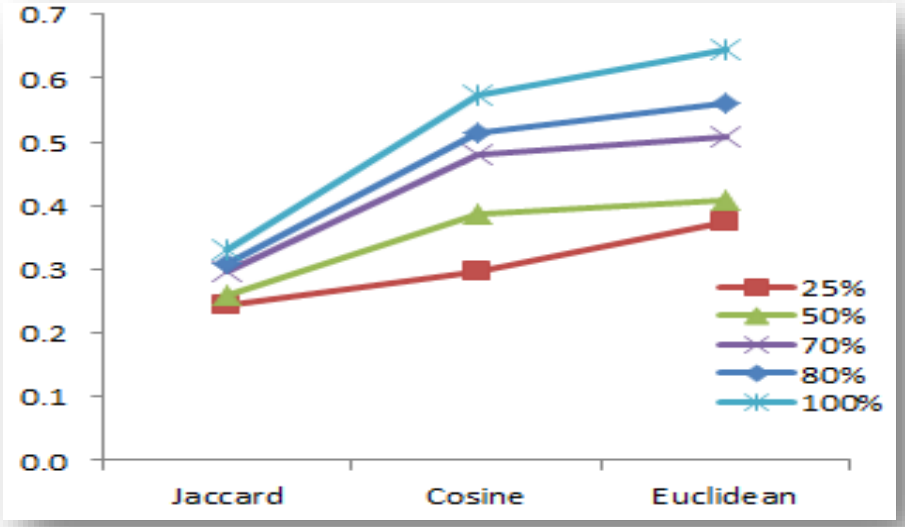


Fig. 5. Correlations with human graders: Interaction between dimensionality and similarity metrics

TESTING AND VALIDATION

Université de Montréal

- Testing on entire corpus
  - implementation of two models
    - model 1: TF-IDF, no dimensionality reduction, Euclidean distance
    - model 2: TF-IDF, no dimensionality reduction, Cosine similarity
  - testing with binary approach
    - conversion of average human scores and LSA scores
    - score < 3 : self-disclosing message (removed)
    - score > 3 : relevant message
  - percentage of right graded messages
    - model 1: 41% (best performance model)
    - model 2: 39 %

OUTLINE

Université de Montréal

- Introduction
- Related Work
- LSA-based method for privacy preserving interactions
- Testing and Validation
- **Conclusions and Future Work**

CONCLUSIONS

Université de Montréal

- Interactions in distant learning environment
  - necessary to complete learning
  - privacy threats and challenges associated to self-disclosure (cyber-bullying)

create favorable learning environment and protect learners' privacy

- Two-step based approach
  - mining step: eliminating negative messages causing psychological harm
  - composition step: detecting and removing self-disclosure of personal data

CONCLUSIONS

Université de Montréal

- Two-step based approach
  - composition step: based on Latent Semantic Analysis (LSA)
    - highly parameterized
    - tested parameters: dimensionality reduction, weighting schemes, similarity metrics
  - best performance model
    - human judges correlation $r = 0.64$, $p < 0.001$
    - no reduction, TF-IDF, Euclidean distance
- **Challenges**
  - data preprocessing: peers' interactions informal, many mistakes, and symbols
  - sensitivity of human raters to self-disclosure

FUTURE WORK

Université de Montréal

- **Future work**
  - inclusion of advanced natural language processing techniques
    - enhancing coherence of composed feedback
    - resolving problem of loss of information
  - inclusion of regular expressions in our approach
    - specifying terms associated to self-disclosure
    - decrease sensitivity of human raters
  - larger set of experiments to demonstrate feasibility of proposed approach in real educational scenarios with large volumes of data

Université de Montréal

THANK YOU

QUESTIONS?

AUTHORS

Université de Montréal

- **Mouna Selmi**
  Ph.D Student
  selmimou@iro.umontreal.ca

- **Hicham Hage**
  Assistant Professor
  hhage@ndu.edu.lb

- **Esma Aïmeur**
  Full Professor
  aimeur@iro.umontreal.ca

REFERENCES

Université de Montréal

- D. Despotakis, V. Dimitrova, L. Lau, D. Thakker, A. Ascolese, and L. Pannese, "ViewS in user generated content for enriching learning environments: A semantic sensing approach", In Artificial Intelligence in Education. Springe, 2013.
- M. Puustinen, J. Bernicot, O. Volckaert-Legrier, and M. Baker, "Naturally occurring help-seeking exchanges on homework help forum". *Computers & Education*, 81: p. 89-101, 2015.
- V. Southavilay, K. Yacef, P. Reimann, R.A Calvo, "Analysis of collaborative writing processes using revision maps and probabilistic topic models". *In Proceedings of the Third International Conference on Learning Analytics and Knowledge.* ACM, 2013.
- Selmi, M., Hage, H., and Aïmeur, E. "Opinion Mining For Predicting Peer Effective Feedback Helpfulness". In Proceedings of the 6th International Conference on Knowledge Management and Information Sharing (KMIS), pp. 419-425, October 2014.
- Squicciarini, A., Karumanchi, S., Lin, D., DeSisto, N.: Identifying hidden social circles for advanced privacy configuration. Computers & Security, vol. 41(0), pp. 40-51 (2014).
- Stutzman, F., Vitak, J., Ellison, N. B., Gray, R., Lampe, C.: Privacy in Interaction: Exploring Disclosure and Social Capital in Facebook. in ICWSM (2012).
- Berendt, B., Günther, O., Spiekermann, S.: Privacy in e-commerce: stated preferences vs. actual behavior, vol. 48(4), pp. 101-106. Communications of the ACM (2005).
- Buckel, T., Thiesse, F.: Predicting The Disclosure of Personal Information on Social Networks: An Empirical Investigation. In Wirtschaftsinformatik, pp.101 (2013).
- Zhao, C., Hinds, P., Gao, G.: How and to whom people share: The role of culture in self-disclosure in online communities. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp.67-76. ACM (2012).
- Li, N., Li, T., Venkatasubramanian, M.: t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, IEEE 23rd International Conference on, vol. 7, pp.106-115. IEEE (2007).
- D. Kiela, and S. Clark, "A Systematic Study of Semantic Vector Space Model Parameters". *In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality*, 2014.
- T. Vijayakumar, R. Priya, and C. Palanisamy, "Effective Pattern Discovery and Dimensionality Reduction for Text Under Text Mining", *in Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*, p. 615-623. Springer, 2015.
- S. Sorour, K. Goda, and T. Mine, "Correlation of Topic Model and Student Grades Using Comment Data Mining". In Proceedings of the 46th ACM Technical Symposium on Computer Science Education, pp. 441-446, ACM, February, 2015.

PEARSON CORRELATION

Université de Montréal

- Pearson correlation coefficient
  - measure of the linear correlation between two variables X and Y

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - interpretation of Pearson's correlation coefficient

	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

WEIGHTING SCHEMES (1/2)

Université de Montréal

- Term frequency (TF)
  - *raw frequency*: number of occurences of term *t* in document *d*
  - local approach
- Inverse Document Frequency (IDF)
  - global approach

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

With
- N: total number of documents in the corpus
- $|\{d \in D : t \in d\}|$: number of documents where the term t appears

- Term frequency-Inverse Document frequency (TF-IDF)
  - hybrid approach

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

WEIGHTING SCHEMES (2/2)

Université de Montréal

- Term frequency (TF)
  - *raw frequency*: number of occurences of term *t* in document *d*
  - local approach
- Inverse Document Frequency (IDF)
  - global approach

With
- N: total number of documents in the corpus
- $|\{d \in D : t \in d\}|$: number of documents where the term t appears

- Term frequency-Inverse Document frequency (TF-IDF)
  - hybrid approach
  - assigning low weights to frequent terms

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

WEIGHTING SCHEMES: EXAMPLE

Université de Montréal

Example : number of documents in corpus (5)

- Term frequency
  - high weights: English, speak, feel (2 occurences)
- Inverse Document Frequency:
  - high weights: non frequent terms (patient, hardworking, engineering, etc.)
  - idf(India, Corpus)= log 5/1= 0.69
- Term frequency-Inverse Document Frequency:
  - tf-idf (India)= TF(India, d2) x IDF (India, Corpus)

Corpus	
d1 →	{It is necessary to be determined, patient and hardworking when learning English}
d2 →	{I am 22 years old, engineering student from India}
d3 →	{my family cannot speak English}
d4 →	{I feel bad to speak with my American girlfriend because I think I could be wrong}
d5 →	{I think that we fail in learning and conversing efficiently because we are just feel shy to speak to someone}

SIMILARITY/DISTANCE METRICS

Université de Montréal

- Cosine
  - measure of similarity between two vectors of attributes A and B

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

- Jaccard Coefficient
  - measure of similarity between finite sample sets
  - defined as the size of the intersection divided by the size of the union of the sample sets

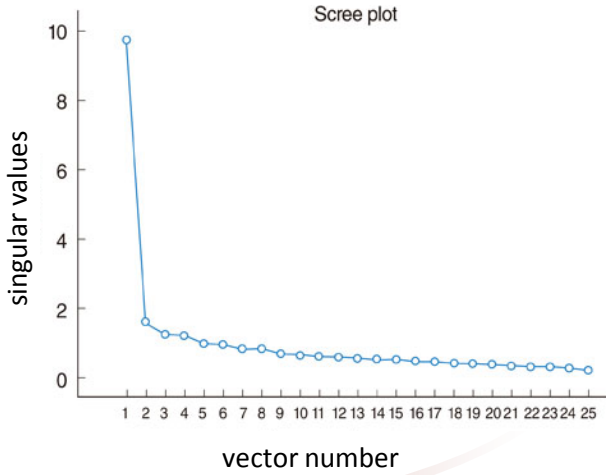
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

ENERGY AND VARIANCE

Université de Montréal

- Energy method
  - retaining enough singular values to make up 90% of the energy in initial matrix
  - sum of the squares of the retained singular values should be at least 90% of the sum of the squares of all the singular values

- Variance method
  - measuring probability distribution
  - plotting singular values in a scree plot
  - retaining dimensionality associated with the knee of the curve

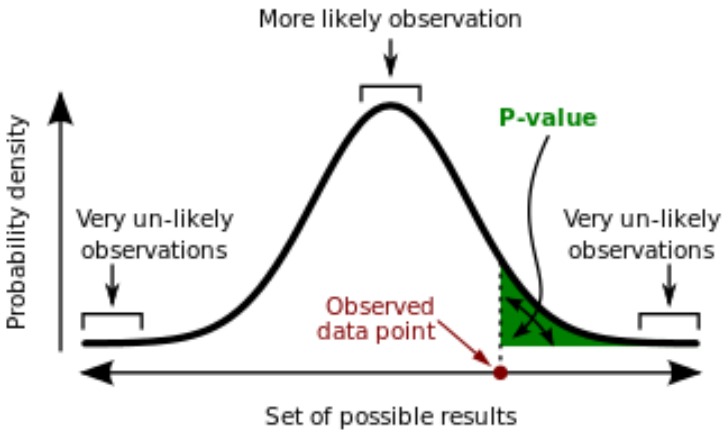


P-VALUE

Université de Montréal

- Definition
  - function of observed sample results used for testing a statistical hypothesis
  - significance level of the test, traditionally 5% or 1%

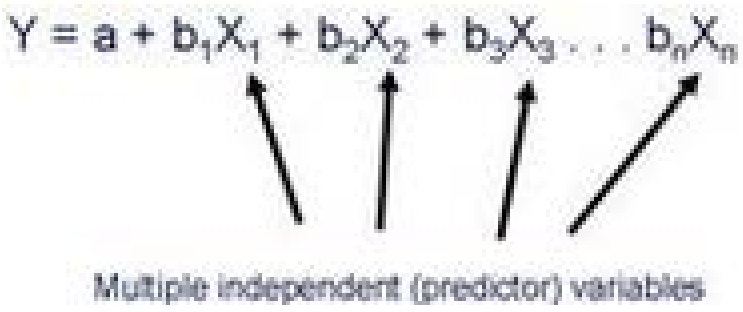
- Example of a p-value computation
  - *p*-value is the area under the curve past the observed data point



REGRESSION ANALYSIS

Université de Montréal

- Regression analysis
  - Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors.
  - more than one independent variable, regression line cannot be visualized in the two dimensional space
  - predict the value of Y for given values of X_1, X_2, …, X_k



Regression Analysis

Université de Montréal

- Regression analysis
  - statistical relationship between one or more independent variables and a dependent variable
  - mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant
  - more than one independent variable, regression line cannot be visualized in the two dimensional space
- Coefficient interpretation
  - testing p-value for null hypothesis
  - low p-value (< 0.05): reject null hypothesis
  - larger (insignificant) p-value: no association between dependent and independent variables