Mini-projet
Été 2021

MTI815 - Systèmes de communication vocale

1. Présentation

1.1 Objectif du mini-projet

L'objectif de ce mini-projet est de comparer deux méthodes de réduction de dimension, LSA et NMF, sur un corpus de textes qui doit être préparé. Le corpus est en format texte et est rédigé en anglais.

1.2 Description détaillée

Vous devez comparer la factorisation matricielle non négative (NMF) et l'analyse sémantique latente (LSA) en utilisant deux (2) composants. Pour ce faire, le texte doit d'abord être préparé. Cette préparation doit comprendre:

- l'uniformisation les textes en minuscules,
- l'élimination des nombres,
- l'élimination de la ponctuation,
- l'élimination des "stop words" de l'anglais (voir la section sur le contenu du rapport).

Une fois les factorisations sont calculées, la matrice est construite en utilisant chaque factorisation. L'erreur est ensuite calculée. Les visualisations 2D de l'espace des documents correspondant à chaque factorisation sont présentées sous forme de graphiques en nuage de points. La structure du rapport, présentée à la section 3, reprend chacun de ces éléments et les associe à chacune des composantes de l'évaluation (voir la section suivante).

2. Évaluation

2.1 Objectif d'apprentissage

En réalisant ce mini-projet, vous développerez les compétences suivantes:

1. nettoyer et préparer un corpus de textes pour une application de TALN;
2. utiliser les librairies nécessaires et utilisez la paramétrisation adéquate;
3. comparer quantitativement et qualitativement deux méthodes visant le même objectif (soit ici, réduire la dimension des données);
4. visualiser les résultats;
5. discuter des limites de chaque approche;
6. synthétiser votre expérience dans un bref rapport d'analyse.

2.2 Critères d'évaluation

Les critères d'évaluation reflètent les objectifs d'apprentissage énoncés plus haut. Le projet est corrigé sur 50.

1. **Nettoyage du corpus:** Tous les éléments demandés et un ajout pertinent que vous jugez nécessaire d'incorporer. Le ou les ajouts doivent être justifié(s). (7 pts)
2. **Utilisation des librairies:** Aucune erreur de paramétrisation n'est présente. Notamment, les choix de lignes ou de colonnes dans les matrices factorisées doivent être corrects. (10 pts)
3. **Comparaison quantitative des méthodes**: La comparaison quantitative fait apparaître l'erreur de reconstruction des matrices grâce à chaque factorisation. Le calcul doit être correct. Vous donnez ici une opinion éclairée de ce qui est obtenu. Il est permis d'utiliser une graphique si cela permet de mettre en relief les différences (9 pts)
4. **Comparaison qualitative des méthodes**: Vous devez présenter au moins les deux graphiques demandés (espaces 2D des documents). Les ajouts sont permis, mais ils doivent être justifiés. N'ajoutez pas de graphiques pour simplement remplir des pages. Comparez ensuite les résultats de manière qualitative en utilisant les visualisations. (8 pts = 4 pts pour les graphiques + 4 pts pour les commentaires sur ceux-ci)
5. **Discussion des résultats**: C'est votre conclusion. Qu'est-ce que vous avez pu observer? Quelles sont les limites? Les avantages? Quels documents se retrouvent en retrait? Peut-on faire des observations intéressantes sur les mots? (12 pts)
6. **Synthèse de votre expérience**: Le rapport doit respecter les consignes. Le respect du format est essentiel. Attention également à l'orthographe et la grammaire. Nous supposons ici que vous faites ce très court rapport dans le cadre d'une exploration réalisée en milieu professionnel. Vous envisagez ce rapport comme la version finale. C'est donc dire que nous ne voulons pas un historique. Nous voulons les faits! (4 pts)

3. Rapport de projet

Votre rapport doit être bref et concis et respecter les consignes présentés à la section 3.1. Un rapport fictif est présenté à la section 3.2.

3.1 Sections du rapport

Votre bref rapport contient les 5 premières sections suivantes. Vous pouvez utiliser les cases ci-dessous pour vous assurer que vous avez tout couvert. NOTE: "très brièvement" = une phrase ou deux.

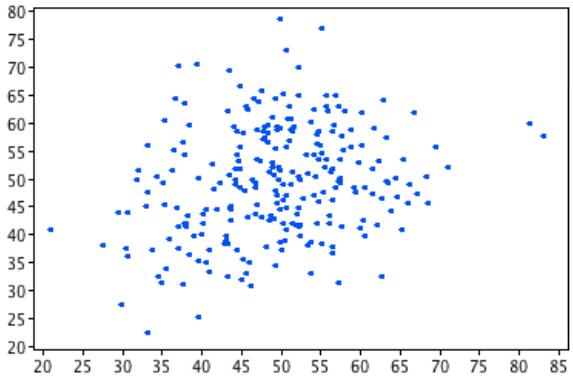
1. **Corpus et prétraitement** (lié au critère d'évaluation 1)
   - ⬚ Décrivez le corpus: Définir ce qu'est un "document"; Nombre de documents; brève description de son contenu.
   - ⬚ Mettre en minuscule: Copiez la commande utilisée ou décrivez TRÈS brièvement comment vous avez procédé.
   - ⬚ Enlevez la ponctuation: quelle ponctuation a été enlevée? Copiez la commande utilisée ou décrivez TRÈS brièvement comment vous avez procédé.
   - ⬚ Éliminez les "stop words": Quelle liste avez-vous utilisée? Mettre liste en annexe ou fournir un lien. Combien de "stop words" contient-elle?
   - ⬚ Décrivez brièvement tout ajout effectué et expliquez très brièvement pourquoi.

2. **Utilisation de l'environnement** (lié au critère d'évaluation 2)
   - ⬚ Indiquez le procédé (commande ou ligne de code) pour créer une matrice TF.
   - ⬚ Indiquez le procédé (commande ou ligne de code) pour faire une factorisation NMF réduite.
   - ⬚ Pour NMF, indiquez où (lignes et colonnes de quelles matrices?) sont les vecteurs correspondant aux documents et les vecteurs correspondant aux topics.
   - ⬚ Indiquez le procédé (commande ou ligne de code) pour faire une factorisation LSA réduite
   - ⬚ Pour LSA, indiquez où (lignes et colonnes de quelles matrices?) sont les vecteurs correspondant aux documents et les vecteurs correspondant aux topics.
3. **Comparaison quantitative des méthodes** (lié au critère d'évaluation 3)
   - ⬚ Indiquez la formule utilisée pour reconstruire la matrice à partir de la factorisation NMF.
   - ⬚ Calculez et présentez l'erreur de reconstruction avec NMF
   - ⬚ Indiquez la formule utilisée pour reconstruire la matrice à partir de la factorisation LSA
   - ⬚ Calculez et présentez l'erreur de reconstruction avec LSA
4. **Comparaison qualitative des méthodes** (ié au critère d'évaluation 4)
   - ⬚ Présentez et commentez brièvement le graphique 2D de l'espace des documents obtenu à partir de la factorisation NMF.
   - ⬚ Présentez et commentez brièvement le graphique 2D de l'espace des documents obtenu à partir de la factorisation LSA.
5. **Discussion** (lié au critère d'évaluation 5)
   - ⬚ Discutez des avantages et des inconvénients que vous avez observés (pas ceux tirés de la littérature!)
   - ⬚ Quels documents se retrouvent en retrait?
   - ⬚ Peut-on faire des observations intéressantes sur les mots?
   - ⬚ Quel serait l'impact de modifier la définition de "document"?
6. **Le rapport en soi** (lié au critère d'évaluation 6)
   - ⬚ Qualité du français
   - ⬚ Respect de ce gabarit

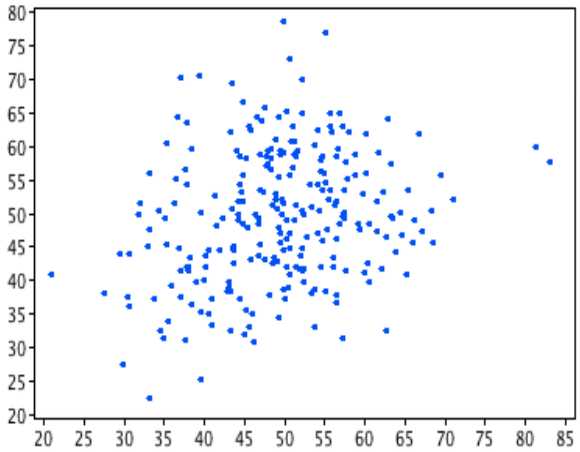
3.2 Rapport fictif

1. **Corpus et prétraitement**
   - Nous interprétons un document comme étant *Lorem ipsum dolor sit amet, graece*. Conséquemment, le corpus contient XXX documents. Chaque document décrit *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos.*
   - minuscule : la commande est *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - ponctuation : Nous avons enlevé NN ponctuations :liste. La commande est *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - stop word : la liste contient XX stop words. elle est en annexe. La commande est : *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - ajout : nous avons ajouté XYZ parce que *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea cu partem necessitatibus. In mea omnis affert lucilius, pri ut solet pericula, eros latine praesent eu his. Vix liber discere contentiones at, ea mei case habeo intellegebat, malis quodsi pri cu.*

2. **Utilisation de l'environnement**
    - Création de la matrice TF. *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
    - Factorisation NMF réduite. *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
    - NMF: les vecteurs des documents se trouvent dans les lignes de la matrice XXX
    - Factorisation LSA réduite. *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
    - LSA: les vecteurs des documents se trouvent dans les lignes de la matrice XXX
3. **Comparaison quantitative des méthodes**
    - X = formule pour reconstruction avec NMF
    - Formule erreur = valeur, valeur.
    - X = formule pour reconstruction avec LSA
    - Formule erreur = valeur, valeur
4. **Comparaison qualitative des méthodes**
    - Graphique pour NMF, espace de document



    - Commentaire NMF: *orem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del orem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del*
    - Graphique pour LSA, espace de document



    - Commentaire LSA: *orem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del orem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del*.

5. **Discussion**
    - Avantages et inconvénients: *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos.*
    - Quels documents se retrouvent en retrait? *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos.*
    - Peut-on faire des observations intéressantes sur les mots? *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos.*

Mini-project
Summer 2021

MTI815 - Voice communication systems

1. Presentation

1.1 Objective of the mini-project

The objective of this mini-project is to compare two dimension reduction methods, LSA and NMF, on a body of texts, which must be prepare. The corpus is in text format and is written in English.

1.2 Detailed Description

You must compare the Non-Negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA) using two (2) components. To do this, the text must first be prepared. This preparation must include:

- standardization of lowercase texts,
- elimination of numbers,
- elimination of punctuation,
- elimination of “stop words” from English (see the section on the content of the report).

Once the factorizations are calculated, the matrix is constructed using each factorization. The error is then calculated. The 2D visualizations of the document space corresponding to each factorization are presented in the form of point cloud graphics. The structure of the report, presented in section 3, takes each of these elements and links them to each of the evaluation components (see next section).

2. Assessment

2.1 Learning objective

By carrying out this mini-project, you will develop the following skills:

1. cleaning and preparing a body of texts for a NLP application;
2. use the necessary libraries and use the appropriate parameterization;
3. to compare quantitatively and qualitatively two methods aiming at the same objective (ie here, to reduce the dimension of the data);
4. visualize the results;
5. discuss the limitations of each approach;
6. synthesize your experience in a brief analysis report.

2.2 Assessment criteria

The assessment criteria reflect the learning objectives stated above. The project is corrected on 50.

1. **Cleaning of the corpus:** All the requested elements and a relevant addition that you deem necessary to incorporate. The addition (s) must be justified. (7 pts)

2. **Use of libraries:** No parameterization error is present. In particular, the choices of rows or columns in the factorized matrices must be correct. (10 pts)
3. **Quantitative comparison of the methods**: The quantitative comparison reveals the error of reconstruction of the matrices thanks to each factorization. The calculation must be correct. Here you are giving an informed opinion of what is achieved. It is allowed to use a graph if it allows to highlight the differences (9 pts)
4. **Qualitative comparison of the methods**: You must present at least the two requested graphs (2D spaces of the documents). Additions are allowed, but they must be justified. Don't add graphics just to fill pages. Then compare the results qualitatively using the visualizations. (8 pts = 4 pts for the graphs + 4 pts for the comments on them)
5. **Discussion of the results**: This is your conclusion. What were you able to observe? What are the limits? Benefits? Which documents are indented? Can we make some interesting observations about words? (12 pts)
6. **Summary of your experience**: The report must respect the instructions. Respect for the format is essential. Also pay attention to spelling and grammar. We are assuming here that you are doing this very short report as part of an exploration carried out in a professional environment. You see this report as the final version. This means that we do not want a history. We want the facts! (4 pts)

3. Project report

Your report must be short and concise and follow the instructions presented in section 3.1. A mock report is presented in section 3.2.

3.1 Report Sections

Your brief report contains the following first 5 sections. You can use the boxes below to make sure you've got everything covered. NOTE: “very briefly” = a sentence or two.

1. **Corpus and preprocessing** (linked to evaluation criterion 1)
   - ⬚ Describe the corpus: Define what a “document” is; Number of documents; brief description of its contents.
   - ⬚ Lowercase: Copy the command used or describe VERY briefly how you did it.
   - ⬚ Remove punctuation: what punctuation has been removed? Copy the command you used or describe VERY briefly how you did it.
   - ⬚ Eliminate the “stop words”: Which list did you use? Attach list or provide a link. How many “stop words” does it contain?
   - ⬚ Briefly describe any additions made and explain very briefly why.
2. **Use of the environment** (linked to evaluation criterion 2)
   - ⬚ Indicate the process (command or line of code) to create a TF matrix.
   - ⬚ Indicate the process (command or line of code) to do a reduced NMF factorization.
   - ⬚ For NMF, indicate where (rows and columns of which matrices?) Are the vectors corresponding to the documents and the vectors corresponding to the topics.
   - ⬚ Indicate the process (command or line of code) to perform a reduced LSA factorization
   - ⬚ For LSA, indicate where (rows and columns of which matrices?) Are the vectors corresponding to the documents and the vectors corresponding to the topics.

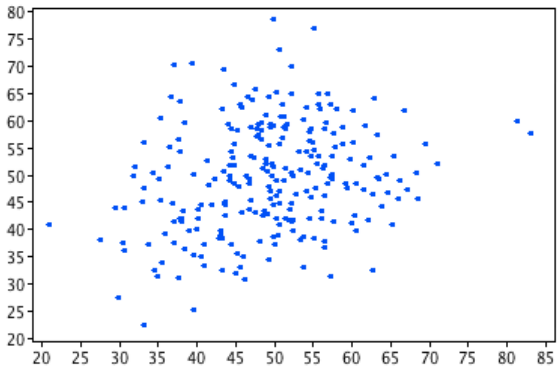
3. **Quantitative comparison of methods** (linked to evaluation criterion 3)
   - ⬚ Indicate the formula used to reconstruct the matrix from the NMF fadtorisation.
   - ⬚ Calculate and present the reconstruction error with NMF
   - ⬚ Indicate the formula used to reconstruct the matrix from the LSA fadtorisation
   - ⬚ Calculate and present the reconstruction error with LSA
4. **Qualitative comparison of the methods** (related to evaluation criterion 4)
   - ⬚ Present and comment briefly the 2D graph of the document space obtained from the NMF factorization.
   - ⬚ Present and briefly comment on the 2D graph of the document space obtained from the LSA factorization.
5. **Discussion** (linked to evaluation criterion 5)
   - ⬚ Discuss the advantages and disadvantages that you observed (not those in the literature!)
   - ⬚ Which documents are indented?
   - ⬚ Can we make some interesting observations about words?
   - ⬚ What would be the impact of changing the definition of "document"?
6. **The report itself** (linked to evaluation criterion 6)
   - ⬚ Quality of English
   - ⬚ Compliance with this template

3.2 Fictitious report

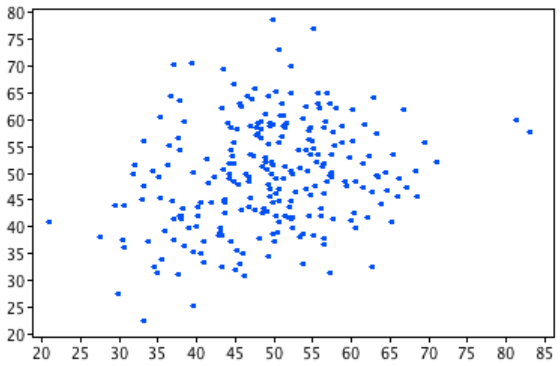
1. **Corpus and preprocessing**
   - We interpret a document as being *Lorem ipsum dolor sit amet, graece*. Consequently, the corpus contains XXX documents. Each document describes *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos.*
   - lowercase: the command is *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - punctuation: We removed NN punctuation: list. The command is *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - stop word: the list contains XX stop words. it is in the appendix. The command is: *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. My*
   - addition: we added XYZ because *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea cu partem necessitatibus. In mea omnis affert lucilius, pri ut solet pericula, Latin eros praesent eu his. Vix liber discere contentiones at, ea mei case habeo intellegebat, malis quodsi pri cu. Vis ut diam ponderum, mei homero eripuit consulatu and.*
2. **Use of the environment**
   - Creation of the TF matrix. *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - Reduced NMF factorization. *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - NMF: the vectors of the documents are in the rows of the matrix XXX
   - Reduced LSA factorization. *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Mea*
   - LSA: the vectors of the documents are in the rows of the matrix XXX

3. **Quantitative comparison of the methods**
    - X = formula for reconstruction with NMF
    - Formula error = value, value.
    - X = formula for reconstruction with LSA
    - Formula error = value, value

4. **Qualitative comparison of methods**
    - Graph for NMF, document space



    - NMF: *commentorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del orem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del*
    - Graphic for LSA, document space



    - LSA: *commentorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del orem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum del*.

5. **Discussion**
    - Advantages and disadvantages: *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio..*
    - Which documents are indented? *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio.*
    - Can we make some interesting observations about words? *Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos. Lorem ipsum dolor sit amet, graece albucius voluptatum mel ut, te ius inermis quaestio. Cu his malorum delenit tractatos.*