Séance 2.1: Analyse données digitales

Analyse de texte

Visseho Adjiwanou, PhD.

SICSS - Montréal

29 November 2021

Plan de présentation

1. Introduction
2. Présentation des données textuelles
3. Comment analyser le texte
   - Traitement (processing)
   - Combien vaut un mot?
4. Approches et applications
5. Évaluation
6. Outils
7. Résumé
8. Ressources

Introduction

Introduction

- Que savons-nous sur les politiques d'immigration au Canada et aux États-Unis?

Introduction

- Que savons-nous sur les politiques d'immigration au Canada et aux États-Unis?
- Que nous renseignent les agendas des garderies sur le bien-être des enfants?

Introduction

- Que savons-nous sur les politiques d'immigration au Canada et aux États-Unis?
- Que nous renseignent les agendas des garderies sur le bien-être des enfants?
- Qu'est-ce qui se dit sur les médias sociaux sur la liberté académique et la lutte contre le racisme?

Introduction

- Que savons-nous sur les politiques d'immigration au Canada et aux États-Unis?
- Que nous renseignent les agendas des garderies sur le bien-être des enfants?
- Qu'est-ce qui se dit sur les médias sociaux sur la liberté académique et la lutte contre le racisme?
- Que pensent les québécois sur la perte du français à Montréal, qu'en est-il des Montréalais ?

Introduction

- De plus en plus recours à l'**analyse quantitative de texte**

Introduction

- De plus en plus recours à l'**analyse quantitative de texte**
- Texte : Tout objet pouvant être « lu »

Introduction

- De plus en plus recours à l'**analyse quantitative de texte**
- Texte : Tout objet pouvant être « lu »
- Quantitative: recours à la quantification, à l'ordinateur

Introduction

- De plus en plus recours à l'**analyse quantitative de texte**
- Texte : Tout objet pouvant être « lu »
- Quantitative: recours à la quantification, à l'ordinateur
- Analyse: Examen systématique de la structure ou des mécanismes de quelque chose

Introduction

- De plus en plus recours à l'**analyse quantitative de texte**
- Texte : Tout objet pouvant être « lu »
- Quantitative: recours à la quantification, à l'ordinateur
- Analyse: Examen systématique de la structure ou des mécanismes de quelque chose
- Définition: Examen systématique assisté par ordinateur de la structure ou mécanismes de contenu lisible

2. Présentation des données textuelles

Évolution de l'analyse quantitative du texte

Années 1600 : l'église catholique suit la proportion de textes imprimés non religieux

1934 : Laswell produit le premier compte de mots-clés

Années 1940: les chercheurs en sciences sociales utilisent des méthodes similaires

1950 : Turin applique l'IA au texte

1952 : Bereleson publie le premier manuel sur l'analyse de contenu

1954 : Première traduction automatique de texte (Georgetown Experiment)

1966 : Stone & Bales utilisent un ordinateur central pour mesurer les propriétés psychométriques du texte

Évolution de l'analyse quantitative du texte

1980 : Apprentissage automatique appliqué à la NLP

1985 : Schrodt introduit le codage automatisé des événements

1986 : Pennebaker développe LIWC

1989 : Franzosi apporte l'analyse narrative quantitative aux sciences sociales

Évolution des techniques

1998 : Premiers modèles thématiques développés

1998 : Mohr effectue la première analyse quantitative des visions du monde

1999: Bearman et al. appliquer les méthodes de réseau aux récits

2001: Blei et al. développer LDA

2003 : Création de MALLET

2005 : Quin et al utilisent l'analyse des discours politiques à l'aide de modèles thématiques

2010: King/Hopkins introduisent les modèles thématiques dans le courant dominant

2014: Margaret Roberts et coll. ont développé des modèles thématiques structurels

3. Comment analyser le texte

Difficulté du langage humain

- Le langage humain est complexe et nuancé

Difficulté du langage humain

- Le langage humain est complexe et nuancé
- humour

Difficulté du langage humain

- Le langage humain est complexe et nuancé
- humour
- double négation

Difficulté du langage humain

- Le langage humain est complexe et nuancé
- humour
- double négation
- Variation des termes selon les différents contexte

Difficulté du langage humain

- Le langage humain est complexe et nuancé
- humour
- double négation
- Variation des termes selon les différents contexte
- *valise* pour désigner le coffre d'une voiture au Québec

Difficulté du langage humain

- Le langage humain est complexe et nuancé
- humour
- double négation
- Variation des termes selon les différents contexte
- *valise* pour désigner le coffre d'une voiture au Québec
- Le but de l'analyse de texte est de réduire cette complexité pour extraire des messages compréhensifs et importants

Type d'analyse

- La réduction de la complexité peut se faire à partir :
  1. *de la catégorisation de texte ou de la classification automatique*
     - *exemple: segmentation des sujets dans un débat politique*
  2. *d'un système de recherche d'information (information retrieval): extraire un message important d'une donnée de texte qui répondra à une requête spéciale*
     - *processus d'analyse d'un texte en entier ou de métadonnées d'un document pour en produire une connaissance données, basée sur la requête*
     - *Exemples:*
     - *analyse de sentiment (sentiment analysis),*
     - *découverte des connaissances (knowledge discovery),*

Type d'analyse

- le choix de l'outil approprié pour répondre à un problème dépend du contexte et de l'application

Type d'analyse

- le choix de l'outil approprié pour répondre à un problème dépend du contexte et de l'application
- Par exemple: les techniques de classification de document peuvent être utilisées pour :

Type d'analyse

- le choix de l'outil approprié pour répondre à un problème dépend du contexte et de l'application
- Par exemple: les techniques de classification de document peuvent être utilisées pour :
- obtenir un aperçu du contenu général d'un grand corpus de documents,

Type d'analyse

- le choix de l'outil approprié pour répondre à un problème dépend du contexte et de l'application
- Par exemple: les techniques de classification de document peuvent être utilisées pour :
- obtenir un aperçu du contenu général d'un grand corpus de documents,
- découvrir un domaine de connaissances particulier, ou

Type d'analyse

- le choix de l'outil approprié pour répondre à un problème dépend du contexte et de l'application
- Par exemple: les techniques de classification de document peuvent être utilisées pour :
- obtenir un aperçu du contenu général d'un grand corpus de documents,
- découvrir un domaine de connaissances particulier, ou
- lier des corpus basés sur des relations sémantiques implicites

Traitement des données

- La première étape est le nettoyage des données (pré-traitement et réduction de la dimensionnalité). Etape essentielle pour une bonne réussite de l'algorithme

Traitement des données

- La première étape est le nettoyage des données (pré-traitement et réduction de la dimensionnalité). Etape essentielle pour une bonne réussite de l'algorithme
- Plus compliqué que dans le cas des données numériques / rectangulaires

Traitement des données

- La première étape est le nettoyage des données (pré-traitement et réduction de la dimensionnalité). Etape essentielle pour une bonne réussite de l'algorithme
- Plus compliqué que dans le cas des données numériques / rectangulaires
- les données de textes sont non structurées

Traitement des données

- La première étape est le nettoyage des données (pré-traitement et réduction de la dimensionnalité). Etape essentielle pour une bonne réussite de l'algorithme
- Plus compliqué que dans le cas des données numériques / rectangulaires
- les données de textes sont non structurées
- Elles sont désordonnées

Traitement des données

- La première étape est le nettoyage des données (pré-traitement et réduction de la dimensionnalité). Etape essentielle pour une bonne réussite de l'algorithme
- Plus compliqué que dans le cas des données numériques / rectangulaires
- les données de textes sont non structurées
- Elles sont désordonnées
- Peut être long et fastidieux, mais se résume à un ensemble de techniques

Traitement des données

- Voici les différentes terminologies et techniques

1. Corpus de texte (text corpora)

Traitement des données

- Voici les différentes terminologies et techniques

1. Corpus de texte (text corpora)
2. Tokenisation (tokenization)

Traitement des données

- Voici les différentes terminologies et techniques

1. Corpus de texte (text corpora)
2. Tokenisation (tokenization)
3. Stop words (mots rares ou vide)

Traitement des données

- Voici les différentes terminologies et techniques

1. Corpus de texte (text corpora)
2. Tokenisation (tokenization)
3. Stop words (mots rares ou vide)
4. N-grams (mots composés)

Traitement des données

- Voici les différentes terminologies et techniques

1. Corpus de texte (text corpora)
2. Tokenisation (tokenization)
3. Stop words (mots rares ou vide)
4. N-grams (mots composés)
5. Stemming (même racine ou radical) and lemmatization (même forme canonique)

Traitement des données

- Voici les différentes terminologies et techniques

1. Corpus de texte (text corpora)
2. Tokenisation (tokenization)
3. Stop words (mots rares ou vide)
4. N-grams (mots composés)
5. Stemming (même racine ou radical) and lemmatization (même forme canonique)
6. Autres aspect important du prétraitement des données.

1. Corpus de texte

- Un ensemble de documents similaires est appelé un corpus

1. Corpus de texte

- Un ensemble de documents similaires est appelé un corpus
- Brown corpus : corpus pour l'anglais américain

1. Corpus de texte

- Un ensemble de documents similaires est appelé un corpus
- Brown corpus : corpus pour l'anglais américain
- Comprends 1 million de mots de texte courant de prose anglaise imprimée aux États-Unis au cours de l'année 1961 (https://www.sketchengine.eu/brown-corpus/)

1. Corpus de texte

- Un ensemble de documents similaires est appelé un corpus
- Brown corpus : corpus pour l'anglais américain
- Comprends 1 million de mots de texte courant de prose anglaise imprimée aux États-Unis au cours de l'année 1961 (https://www.sketchengine.eu/brown-corpus/)
- Nombreux dictionnaires lexicaux de sentiments:

1. Corpus de texte

- Un ensemble de documents similaires est appelé un corpus
- Brown corpus : corpus pour l'anglais américain
- Comprends 1 million de mots de texte courant de prose anglaise imprimée aux États-Unis au cours de l'année 1961 (https://www.sketchengine.eu/brown-corpus/)
- Nombreux dictionnaires lexicaux de sentiments:
- **afinn** qui comprend une liste de mots chargés de sentiments qui sont apparus dans les discussions de Twitter sur le changement climatique;

1. Corpus de texte

- Un ensemble de documents similaires est appelé un corpus
- Brown corpus : corpus pour l'anglais américain
- Comprends 1 million de mots de texte courant de prose anglaise imprimée aux États-Unis au cours de l'année 1961 (https://www.sketchengine.eu/brown-corpus/)
- Nombreux dictionnaires lexicaux de sentiments:
- **afinn** qui comprend une liste de mots chargés de sentiments qui sont apparus dans les discussions de Twitter sur le changement climatique;
- **bing** qui comprend des mots sensibles identifiés sur les forums en ligne; et

1. Corpus de texte

- Un ensemble de documents similaires est appelé un corpus
- Brown corpus : corpus pour l'anglais américain
- Comprends 1 million de mots de texte courant de prose anglaise imprimée aux États-Unis au cours de l'année 1961 (https://www.sketchengine.eu/brown-corpus/)
- Nombreux dictionnaires lexicaux de sentiments:
- **afinn** qui comprend une liste de mots chargés de sentiments qui sont apparus dans les discussions de Twitter sur le changement climatique;
- **bing** qui comprend des mots sensibles identifiés sur les forums en ligne; et
- **nrc** qui est un dictionnaire qui a été créé en demandant aux travailleurs d'Amazon Mechanical Turk de coder la valence émotionnelle d'une longue liste de termes.

1. Corpus de texte

- **Lexicoder Sentiment Dictionary (LSD)** qui est un lexique large noté pour le ton positif et négatif et adapté principalement aux textes politiques (Newspaper,. . .). Il contient plus de 4 500 mots positifs et négatifs utilisés pour transmettre des sentiments. (https://quanteda.io/reference/data_dictionary_LSD2015.html)

1. Corpus de texte

- **Lexicoder Sentiment Dictionary (LSD)** qui est un lexique large noté pour le ton positif et négatif et adapté principalement aux textes politiques (Newspaper,. . .). Il contient plus de 4 500 mots positifs et négatifs utilisés pour transmettre des sentiments. (https://quanteda.io/reference/data_dictionary_LSD2015.html)
- La présentation sur Wikipedia vaut la peine d'être lue : https://fr.wikipedia.org/wiki/Corpus

1. Corpus de texte

- Tous les corpus ne sont pas efficaces pour tous les usages:

1. Corpus de texte

- Tous les corpus ne sont pas efficaces pour tous les usages:
- le nombre et

1. Corpus de texte

- Tous les corpus ne sont pas efficaces pour tous les usages:
- le nombre et
- la portée des documents

1. Corpus de texte

- Tous les corpus ne sont pas efficaces pour tous les usages:
- le nombre et
- la portée des documents
- la portée des documents détermine l'éventail des questions que vous pouvez poser et la qualité des réponses que vous obtiendrez:

1. Corpus de texte

- Tous les corpus ne sont pas efficaces pour tous les usages:
- le nombre et
- la portée des documents
- la portée des documents détermine l'éventail des questions que vous pouvez poser et la qualité des réponses que vous obtiendrez:
- trop peu de documents se traduisent par un manque de couverture.

1. Corpus de texte

- Tous les corpus ne sont pas efficaces pour tous les usages:
- le nombre et
- la portée des documents
- la portée des documents détermine l'éventail des questions que vous pouvez poser et la qualité des réponses que vous obtiendrez:
- trop peu de documents se traduisent par un manque de couverture.
- trop de mauvais types de documents invitent à un bruit confondant

1. Corpus de texte

- Tous les corpus ne sont pas efficaces pour tous les usages:
- le nombre et
- la portée des documents
- la portée des documents détermine l'éventail des questions que vous pouvez poser et la qualité des réponses que vous obtiendrez:
- trop peu de documents se traduisent par un manque de couverture.
- trop de mauvais types de documents invitent à un bruit confondant

- Dans le traitement qu'on fera ici, un corpus est juste l'ensemble du texte qu'on va analyser.

1. Corpus de texte

- La première étape du processing est de décider quels termes et phrases sont significatifs

1. Corpus de texte

- La première étape du processing est de décider quels termes et phrases sont significatifs
- La tokenisation sépare les termes et les phrases les uns des autres

1. Corpus de texte

- La première étape du processing est de décider quels termes et phrases sont significatifs
- La tokenisation sépare les termes et les phrases les uns des autres
- les phrases vont être séparées sur la base des signes de ponctuations

1. Corpus de texte

- La première étape du processing est de décider quels termes et phrases sont significatifs
- La tokenisation sépare les termes et les phrases les uns des autres
- les phrases vont être séparées sur la base des signes de ponctuations
- et ensuite en mots

1. Corpus de texte

- La première étape du processing est de décider quels termes et phrases sont significatifs
- La tokenisation sépare les termes et les phrases les uns des autres
- les phrases vont être séparées sur la base des signes de ponctuations
- et ensuite en mots

- Cela va produire un autre document qui sera analysé

1. Corpus de texte

- La première étape du processing est de décider quels termes et phrases sont significatifs
- La tokenisation sépare les termes et les phrases les uns des autres
- les phrases vont être séparées sur la base des signes de ponctuations
- et ensuite en mots

- Cela va produire un autre document qui sera analysé
- matrice documents-termes

1. Corpus de texte

- La première étape du processing est de décider quels termes et phrases sont significatifs
- La tokenisation sépare les termes et les phrases les uns des autres
- les phrases vont être séparées sur la base des signes de ponctuations
- et ensuite en mots

- Cela va produire un autre document qui sera analysé
- matrice documents-termes
- données tidy (tidy-data)

2. Tokénisation

1. Matrice Documents-termes

- Une façon rapide d'explorer des données textuelles consiste à simplement compter les occurrences de chaque mot ou terme.

2. Tokénisation

1 Matrice Documents-termes

- Une façon rapide d'explorer des données textuelles consiste à simplement compter les occurrences de chaque mot ou terme.
- Le nombre de fois qu'un mot particulier apparaît dans un document donné est appelé terme fréquence (tf).

2. Tokénisation

1 Matrice Documents-termes

- Une façon rapide d'explorer des données textuelles consiste à simplement compter les occurrences de chaque mot ou terme.
- Le nombre de fois qu'un mot particulier apparaît dans un document donné est appelé terme fréquence (tf).
- La statistique tf peut être résumée dans une matrice de termes de document, qui est un tableau rectangulaire avec des lignes représentant des documents et des colonnes représentant des termes uniques.

2. Tokénisation

1 Matrice Documents-termes

- Une façon rapide d'explorer des données textuelles consiste à simplement compter les occurrences de chaque mot ou terme.
- Le nombre de fois qu'un mot particulier apparaît dans un document donné est appelé terme fréquence (tf).
- La statistique tf peut être résumée dans une matrice de termes de document, qui est un tableau rectangulaire avec des lignes représentant des documents et des colonnes représentant des termes uniques.
- L'élément (i, j) de cette matrice donne les décomptes :

2. Tokénisation

1 Matrice Documents-termes

- Une façon rapide d'explorer des données textuelles consiste à simplement compter les occurrences de chaque mot ou terme.
- Le nombre de fois qu'un mot particulier apparaît dans un document donné est appelé terme fréquence (tf).
- La statistique tf peut être résumée dans une matrice de termes de document, qui est un tableau rectangulaire avec des lignes représentant des documents et des colonnes représentant des termes uniques.
- L'élément (i, j) de cette matrice donne les décomptes :
- du jième terme (colonne)

2. Tokénisation

1 Matrice Documents-termes

- Une façon rapide d'explorer des données textuelles consiste à simplement compter les occurrences de chaque mot ou terme.
- Le nombre de fois qu'un mot particulier apparaît dans un document donné est appelé terme fréquence (tf).
- La statistique tf peut être résumée dans une matrice de termes de document, qui est un tableau rectangulaire avec des lignes représentant des documents et des colonnes représentant des termes uniques.
- L'élément (i, j) de cette matrice donne les décomptes :
- du jième terme (colonne)
- dans le ième document (ligne).

2. Tokénisation

1 Matrice Documents-termes

- Nous pouvons également inverser les lignes et les colonnes et convertir une matrice documents-termes en une matrice termes-documents où les lignes et les colonnes représentent respectivement les termes et les documents.

Document Term Matrix

	intelligent	applications	creates	business	processes	bots	are	i	do	intelligence
Doc 1	2	1	1	1	1	0	0	0	0	0
Doc 2	1	1	0	0	0	1	1	0	0	0
Doc 3	0	0	0	1	0	0	0	1	1	1

2. Tokénisation

1 Matrice Documents-termes

```
library(tidyverse)

# Document à analyser

texte <- c("texte1", "texte2")
phrase <- c("je suis malade", "je vais à l'hôpital")

document <- data.frame(texte, phrase)
class(document$phrase)
```

```
## [1] "factor"
```

```
document <- document %>%
  mutate(phrase = as.character(phrase))
```

2. Tokénisation

1. Matrice Documents-termes

```
# package tm


library(tm)


document_corpus <- Corpus(VectorSource(as.vector(document$p
document_DTM <- DocumentTermMatrix(document_corpus, control
```

2. Tokénisation

1. Matrice Documents-termes

```
document_DTM
```

```
## <<DocumentTermMatrix (documents: 2, terms: 6)>>
## Non-/sparse entries: 7/5
## Sparsity           : 42%
## Maximal term length: 9
## Weighting          : term frequency (tf)
```

2. Tokénisation

1 Matrice Documents-termes

```
inspect(document_DTM)
```

```
## <<DocumentTermMatrix (documents: 2, terms: 6)>>
## Non-/sparse entries: 7/5
## Sparsity           : 42%
## Maximal term length: 9
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs à je l'hôpital malade suis vais
##    1 0  1          0      1    1    0
##    2 1  1          1      0    0    1
```

2. Tokénisation

2 Tidy-data

- L'utilisation de principes de données bien rangées est un moyen puissant de rendre la gestion des données plus facile et plus efficace, et cela n'est pas moins vrai lorsqu'il s'agit de traiter du texte.

2. Tokénisation

2 Tidy-data

- L'utilisation de principes de données bien rangées est un moyen puissant de rendre la gestion des données plus facile et plus efficace, et cela n'est pas moins vrai lorsqu'il s'agit de traiter du texte.
- Comme décrit par Hadley Wickham (Wickham 2014), les données bien rangées ont une structure spécifique:

2. Tokénisation

2 Tidy-data

- L'utilisation de principes de données bien rangées est un moyen puissant de rendre la gestion des données plus facile et plus efficace, et cela n'est pas moins vrai lorsqu'il s'agit de traiter du texte.
- Comme décrit par Hadley Wickham (Wickham 2014), les données bien rangées ont une structure spécifique:
- Chaque variable est une colonne

2. Tokénisation

2 Tidy-data

- L'utilisation de principes de données bien rangées est un moyen puissant de rendre la gestion des données plus facile et plus efficace, et cela n'est pas moins vrai lorsqu'il s'agit de traiter du texte.
- Comme décrit par Hadley Wickham (Wickham 2014), les données bien rangées ont une structure spécifique:

- Chaque variable est une colonne
- Chaque observation est une rangée

2. Tokénisation

2 Tidy-data

- L'utilisation de principes de données bien rangées est un moyen puissant de rendre la gestion des données plus facile et plus efficace, et cela n'est pas moins vrai lorsqu'il s'agit de traiter du texte.
- Comme décrit par Hadley Wickham (Wickham 2014), les données bien rangées ont une structure spécifique:

- Chaque variable est une colonne
- Chaque observation est une rangée
- Chaque type d'unité d'observation est un tableau

2. Tokénisation

2 Tidy-data

- L'utilisation de principes de données bien rangées est un moyen puissant de rendre la gestion des données plus facile et plus efficace, et cela n'est pas moins vrai lorsqu'il s'agit de traiter du texte.
- Comme décrit par Hadley Wickham (Wickham 2014), les données bien rangées ont une structure spécifique:

- Chaque variable est une colonne
- Chaque observation est une rangée
- Chaque type d'unité d'observation est un tableau

- Nous définissons donc le format de texte bien rangé (tidy-text) comme étant une table avec un jeton (token) par ligne.

2. Tokénisation

2 Tidy-data

```
library(tidytext)

tidy_texte <-
  document %>%
  unnest_tokens("mot", phrase)
```

2. Tokénisation

2 Tidy-data

```
tidy_texte
```

```
##       texte        mot
## 1    texte1         je
## 1.1  texte1       suis
## 1.2  texte1    malade
## 2    texte2         je
## 2.1  texte2       vais
## 2.2  texte2          à
## 2.3  texte2 l'hôpital
```

3. Stop words

- Une fois que les jetons (mots) sont clairement séparés, il est possible d'effectuer un traitement de texte supplémentaire à un niveau de jeton plus granulaire.

3. Stop words

- Une fois que les jetons (mots) sont clairement séparés, il est possible d'effectuer un traitement de texte supplémentaire à un niveau de jeton plus granulaire.
- Les mots vides (stop word) sont une catégorie de mots qui ont une signification sémantique limitée quel que soit le contenu du document:

3. Stop words

- Une fois que les jetons (mots) sont clairement séparés, il est possible d'effectuer un traitement de texte supplémentaire à un niveau de jeton plus granulaire.
- Les mots vides (stop word) sont une catégorie de mots qui ont une signification sémantique limitée quel que soit le contenu du document:
- Prépositions,

3. Stop words

- Une fois que les jetons (mots) sont clairement séparés, il est possible d'effectuer un traitement de texte supplémentaire à un niveau de jeton plus granulaire.
- Les mots vides (stop word) sont une catégorie de mots qui ont une signification sémantique limitée quel que soit le contenu du document:
- Prépositions,
- Articles,

3. Stop words

- Une fois que les jetons (mots) sont clairement séparés, il est possible d'effectuer un traitement de texte supplémentaire à un niveau de jeton plus granulaire.
- Les mots vides (stop word) sont une catégorie de mots qui ont une signification sémantique limitée quel que soit le contenu du document:
- Prépositions,
- Articles,
- Noms communs, etc.

3. Stop words

- **Hapax legomena** sont des mots qui sont utilisés une seule fois ou très rarement dans tout un corpus.

3. Stop words

- **Hapax legomena** sont des mots qui sont utilisés une seule fois ou très rarement dans tout un corpus.
- Ces mots (noms, fautes d'orthographe ou termes techniques rares) sont également peu susceptibles d'avoir une signification contextuelle significative.

3. Stop words

- **Hapax legomena** sont des mots qui sont utilisés une seule fois ou très rarement dans tout un corpus.
- Ces mots (noms, fautes d'orthographe ou termes techniques rares) sont également peu susceptibles d'avoir une signification contextuelle significative.
- Semblables aux mots vides, ces jetons sont souvent ignorés dans la modélisation ultérieure, soit par la conception des méthodes, soit par la suppression manuelle du corpus avant l'analyse proprement dite.

4. N-grams

- Les mots individuels ne sont parfois pas la bonne unité d'analyse.

4. N-grams

- Les mots individuels ne sont parfois pas la bonne unité d'analyse.
- Supprimer aveuglement des mots vides peut masquer des phrases importantes.

4. N-grams

- Les mots individuels ne sont parfois pas la bonne unité d'analyse.
- Supprimer aveuglement des mots vides peut masquer des phrases importantes.
- Exemple: faire la queue, faire la grasse matinée, "systems of innovation" en anglais

4. N-grams

- Les mots individuels ne sont parfois pas la bonne unité d'analyse.
- Supprimer aveuglement des mots vides peut masquer des phrases importantes.
- Exemple: faire la queue, faire la grasse matinée, "systems of innovation" en anglais
- L'identification de ces N-grammes nécessite la recherche de modèles statistiques pour découvrir des phrases qui apparaissent souvent ensemble dans des modèles fixes.

4. N-grams

- Les mots individuels ne sont parfois pas la bonne unité d'analyse.
- Supprimer aveuglement des mots vides peut masquer des phrases importantes.
- Exemple: faire la queue, faire la grasse matinée, "systems of innovation" en anglais

- L'identification de ces N-grammes nécessite la recherche de modèles statistiques pour découvrir des phrases qui apparaissent souvent ensemble dans des modèles fixes.
- Ces combinaisons de phrases sont souvent appelées **collocations**, car leur signification globale est plus que la somme de leurs parties

5. Stemming and lemmatization

- La normalisation du texte est un autre aspect important du prétraitement des données textuelles.

5. Stemming and lemmatization

- La normalisation du texte est un autre aspect important du prétraitement des données textuelles.
- Compte tenu de la complexité du langage naturel, les mots peuvent prendre plusieurs formes en fonction de la structure syntaxique avec un changement limité de leur signification originale.

5. Stemming and lemmatization

- La normalisation du texte est un autre aspect important du prétraitement des données textuelles.
- Compte tenu de la complexité du langage naturel, les mots peuvent prendre plusieurs formes en fonction de la structure syntaxique avec un changement limité de leur signification originale.
- Par exemple, le mot «système» a morphologiquement un pluriel «systèmes» ou un adjectif «systématique».

5. Stemming and lemmatization

- La normalisation du texte est un autre aspect important du prétraitement des données textuelles.
- Compte tenu de la complexité du langage naturel, les mots peuvent prendre plusieurs formes en fonction de la structure syntaxique avec un changement limité de leur signification originale.
- Par exemple, le mot «système» a morphologiquement un pluriel «systèmes» ou un adjectif «systématique».
- Tous ces mots sont sémantiquement similaires et - pour de nombreuses tâches - doivent être traités de la même manière.

5. Stemming and lemmatization

- Par exemple, si un document contient le mot «système» trois fois, «systèmes» une fois et «systématique» deux fois, on peut supposer que le mot «système» avec une signification et une structure morphologique similaires peut couvrir toutes les instances et que la variance être réduit à «système» avec six instances.

5. Stemming and lemmatization (dérivation et lemmatisation)

- Le processus de normalisation de texte est souvent mis en œuvre à l'aide d'algorithmes de lemmatisation et de dérivation établis.

5. Stemming and lemmatization (dérivation et lemmatisation)

- Le processus de normalisation de texte est souvent mis en œuvre à l'aide d'algorithmes de lemmatisation et de dérivation établis.
- Un **lemme** est la forme originale du dictionnaire d'un mot.

5. Stemming and lemmatization (dérivation et lemmatisation)

- Le processus de normalisation de texte est souvent mis en œuvre à l'aide d'algorithmes de lemmatisation et de dérivation établis.
- Un **lemme** est la forme originale du dictionnaire d'un mot.
- Exemple, «allé», «aller» et «va» auront tous le lemme «aller».

5. Stemming and lemmatization (dérivation et lemmatisation)

- Le processus de normalisation de texte est souvent mis en œuvre à l'aide d'algorithmes de lemmatisation et de dérivation établis.
- Un **lemme** est la forme originale du dictionnaire d'un mot.
- Exemple, «allé», «aller» et «va» auront tous le lemme «aller».
- Autre exemple: bien, meilleur, mieux

5. Stemming and lemmatization (dérivation et lemmatisation)

- Le processus de normalisation de texte est souvent mis en œuvre à l'aide d'algorithmes de lemmatisation et de dérivation établis.
- Un **lemme** est la forme originale du dictionnaire d'un mot.
- Exemple, «allé», «aller» et «va» auront tous le lemme «aller».
- Autre exemple: bien, meilleur, mieux
- Le **radical (stem)** est une partie centrale d'un mot donné portant sa signification sémantique primaire et unissant un groupe d'unités lexicales similaires.

5. Stemming and lemmatization (dérivation et lemmatisation)

- Le processus de normalisation de texte est souvent mis en œuvre à l'aide d'algorithmes de lemmatisation et de dérivation établis.
- Un **lemme** est la forme originale du dictionnaire d'un mot.
- Exemple, «allé», «aller» et «va» auront tous le lemme «aller».
- Autre exemple: bien, meilleur, mieux
- Le **radical (stem)** est une partie centrale d'un mot donné portant sa signification sémantique primaire et unissant un groupe d'unités lexicales similaires.
- Exemple, les mots «ordre» et «ordorné» auront le même radical «ord».

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.
- Filtrage sur les fréquences des mots afin d'éviter l'influence de certains mots qui apparaissent très souvent, d'autres très rarement.

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.
- Filtrage sur les fréquences des mots afin d'éviter l'influence de certains mots qui apparaissent très souvent, d'autres très rarement.
- Toutes les étapes de traitement de texte sont essentielles à une analyse réussie.

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.
- Filtrage sur les fréquences des mots afin d'éviter l'influence de certains mots qui apparaissent très souvent, d'autres très rarement.

- Toutes les étapes de traitement de texte sont essentielles à une analyse réussie.
- Certains d'entre eux ont plus d'importance que d'autres, en fonction de l'application spécifique, des questions de recherche et des propriétés du corpus.

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.
- Filtrage sur les fréquences des mots afin d'éviter l'influence de certains mots qui apparaissent très souvent, d'autres très rarement.

- Toutes les étapes de traitement de texte sont essentielles à une analyse réussie.
- Certains d'entre eux ont plus d'importance que d'autres, en fonction de l'application spécifique, des questions de recherche et des propriétés du corpus.
- Il est impératif de disposer de tous ces outils pour produire une entrée propre pour la modélisation et l'analyse ultérieures.

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.
- Filtrage sur les fréquences des mots afin d'éviter l'influence de certains mots qui apparaissent très souvent, d'autres très rarement.

- Toutes les étapes de traitement de texte sont essentielles à une analyse réussie.
- Certains d'entre eux ont plus d'importance que d'autres, en fonction de l'application spécifique, des questions de recherche et des propriétés du corpus.
- Il est impératif de disposer de tous ces outils pour produire une entrée propre pour la modélisation et l'analyse ultérieures.
- Certaines règles simples doivent être suivies pour éviter les erreurs typiques.

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.
- Filtrage sur les fréquences des mots afin d'éviter l'influence de certains mots qui apparaissent très souvent, d'autres très rarement.

- Toutes les étapes de traitement de texte sont essentielles à une analyse réussie.
- Certains d'entre eux ont plus d'importance que d'autres, en fonction de l'application spécifique, des questions de recherche et des propriétés du corpus.
- Il est impératif de disposer de tous ces outils pour produire une entrée propre pour la modélisation et l'analyse ultérieures.
- Certaines règles simples doivent être suivies pour éviter les erreurs typiques.

6. Autres aspect important du prétraitement des données textuelles.

- Retrait des chiffres, de la ponctuation, des URLs, espaces, séparatuers, symboles et d'autres mots spécifiques.
- Filtrage sur les fréquences des mots afin d'éviter l'influence de certains mots qui apparaissent très souvent, d'autres très rarement.

- Toutes les étapes de traitement de texte sont essentielles à une analyse réussie.
- Certains d'entre eux ont plus d'importance que d'autres, en fonction de l'application spécifique, des questions de recherche et des propriétés du corpus.
- Il est impératif de disposer de tous ces outils pour produire une entrée propre pour la modélisation et l'analyse ultérieures.
- Certaines règles simples doivent être suivies pour éviter les erreurs typiques.

Conclusion

- Un radical ne doit pas être utilisé lorsque les données sont complexes et nécessitent la prise en compte de toutes les formes et significations possibles des mots.

Conclusion

- Un radical ne doit pas être utilisé lorsque les données sont complexes et nécessitent la prise en compte de toutes les formes et significations possibles des mots.
- L'examen des résultats intermédiaires à chaque étape du processus peut être utile.

3.2 Combien vaut un mot?

- Tous les mots ne valent pas la même chose; dans un article sur l'électronique, «condensateur» est plus important que «aspect».

3.2 Combien vaut un mot?

- Tous les mots ne valent pas la même chose; dans un article sur l'électronique, «condensateur» est plus important que «aspect».
- La pondération et le calibrage appropriés des mots sont importants pour les consommateurs humains et machines de données textuelles:

3.2 Combien vaut un mot?

- Tous les mots ne valent pas la même chose; dans un article sur l'électronique, «condensateur» est plus important que «aspect».
- La pondération et le calibrage appropriés des mots sont importants pour les consommateurs humains et machines de données textuelles:
- les humains ne veulent pas voir «le» comme le mot le plus fréquent de chaque document dans les résumés,

3.2 Combien vaut un mot?

- Tous les mots ne valent pas la même chose; dans un article sur l'électronique, «condensateur» est plus important que «aspect».
- La pondération et le calibrage appropriés des mots sont importants pour les consommateurs humains et machines de données textuelles:
- les humains ne veulent pas voir «le» comme le mot le plus fréquent de chaque document dans les résumés,
- les algorithmes de classification bénéficient de la connaissance des fonctionnalités réellement importantes pour la création une décision.

3.2 Combien vaut un mot?

- La pondération des mots nécessite d'équilibrer la fréquence d'apparition d'un mot dans un contexte local (tel qu'un document) avec son apparition globale dans la collection de documents.

3.2 Combien vaut un mot?

- La pondération des mots nécessite d'équilibrer la fréquence d'apparition d'un mot dans un contexte local (tel qu'un document) avec son apparition globale dans la collection de documents.
- La fréquence inverse des documents (term frequency–inverse document frequency - TFIDF) est un schéma de pondération pour équilibrer explicitement ces facteurs et hiérarchiser les mots les plus significatifs.

3.2 Combien vaut un mot?

- La pondération des mots nécessite d'équilibrer la fréquence d'apparition d'un mot dans un contexte local (tel qu'un document) avec son apparition globale dans la collection de documents.
- La fréquence inverse des documents (term frequency–inverse document frequency - TFIDF) est un schéma de pondération pour équilibrer explicitement ces facteurs et hiérarchiser les mots les plus significatifs.
- Le modèle TFIDF prend en compte à la fois le terme fréquence d'un jeton et sa fréquence dans le document de sorte que si un mot très fréquent apparaît également dans presque tous les documents, sa signification pour le contexte spécifique du corpus est négligeable.

3.2 Combien vaut un mot?

- Les mots vides sont un bon exemple lorsque les mots très fréquents ont également une signification limitée puisqu'ils apparaissent dans pratiquement tous les documents d'un corpus donné

3.2 Combien vaut un mot?

- Pour chaque token t et chaque document d du corpus D, TFIDF est calculé comme :

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

- où la fréquence des termes est soit un simple décompte :

$$tf(t, d) = f(t, d)$$

- et la fréquence inverse du document est

$$idf(t, D) = log(\frac{N}{df(t)})$$

- avec N = le nombre total de document, df(t) = fréquence du document, ou le nombre de documents qui contient le terme t

8. Ressources

Où se trouvent les données

TWITTER - Barbera (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. Political Analysis. Munger (2017). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. Political Behavior. - Tan, Lee, & Pang (2014). The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. arXiv.org.

REDDIT - Chandrasekhara et al. (2017). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. ACMHCI.

FACEBOOK - Bail, Brown, Mann (2017). Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation. ASR.

Où se trouvent les données

KICKSTARTER - Mitra & Gilbert (2014). The Language That Gets People to Give: Phrases That Predict Success on Kickstarter. CSCW.

AIRBNB - Ma et al. (2017). Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. CSCW.

OTHER - King, Pan, & Roberts (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. American Political Science Review.

Où se trouvent les données

OPEN-ENDED SURVEYS - Roberts et al. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science.

HISTORICAL ARCHIVES - Bearman & Stovel (2000). Becoming a Nazi: A model for narrative networks. Poetics.

- Miller (2013). Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach. Poetics.

ENRON EMAILS - Prabhakaran & Rambow (2017). Dialog Structure Through the Lens of Gender, Gender Environment, and Power. Dialogue & Discourse.

Où se trouvent les données

POLITICAL DOCUMENTS - Rule, Cointet, Bearman (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. PNAS.

- Mohr, Wagner-Pacifici, Breiger, & Bogdanov (2013). Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics. Poetics.

NEWSPAPERS - DiMaggio, Nag, Blei (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics.

- Andrews & Caren (2010). Making the News: Movement Organizations, Media Attention, and the Public Agenda. ASR.