Introduction à l'analyse de données textuelles

Cours-bloc *Statistique textuelle et topic models, 8-9.11.17*

Aris Xanthos

Section des sciences du langage et de l'information (SLI), Unil

Matières

1. Bases conceptuelles
    1. Analyse de données, modélisation, tableau de données
    2. Questionnements uni-, bi- et multivariés
    3. Matrice documents-termes, segmentations, annotations
2. Annotations
    1. Définition, exemples
    2. Bases du format XML
3. Travaux pratiques
    1. Import XML et création de matrices documents-termes
    2. Expressions régulières pour l'extraction de données

1. Bases conceptuelles

Introduction à l'analyse de données textuelles

Le monde mis en données

- L'analyse de données implique de construire une représentation d'un phénomène:
  - Quelle est l'étendue du phénomène représenté?
  - Quels sont les objets (individus, observations, ...) représentés?
  - Par quelles propriétés (variables, caractéristiques, attributs, propriétés, métadonnées, …) sont-ils représentés?
- Données = objets + propriétés
- Généralement organisées sous forme de tableau (lignes = objets et colonnes = propriétés)

Exemples

- Bande dessinée

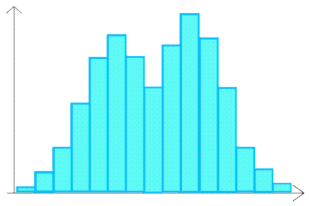
planche	titre	annee	nb_cases	haut_moy_cases	larg_moy_cases	…
1	*BD1*	1978	8	5.3	7.2	…
2	*BD1*	1978	5	5.5	9.2	…
…	…	…	…	…	…	…

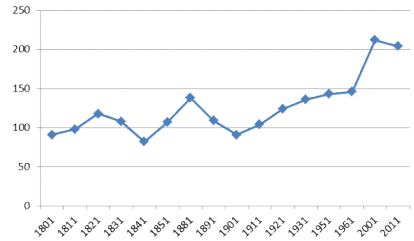
- Théâtre

acte	piece	auteur	long_moy_phrases	prop_adv	…
1	*L'Avare*	*Molière*	12.3	5.6	…
2	*L'Avare*	*Molière*	11.6	3.9	…
…	…	…	…	…	…

Questionnements uni- et bivariés

- Comment résumer l'information contenue dans une colonne de la table?
  - tendance centrale (moyenne, médiane, ...)
  - dispersion (variance, écart-type, ...)
- Quelle relation existe entre deux colonnes?
- Comment visualiser ces informations?
  - histogrammes
  - séries temporelles
  - etc.





Questionnements multivariés

- Dans quelle mesure les profils des objets (i.e. les lignes du tableau) sont-ils similaires?
- Peut-on former des catégories d'objets sur la base de leur similarité?
- Comment peut-on visualiser cette structure de similarité?

Matrice documents-termes (MDT)

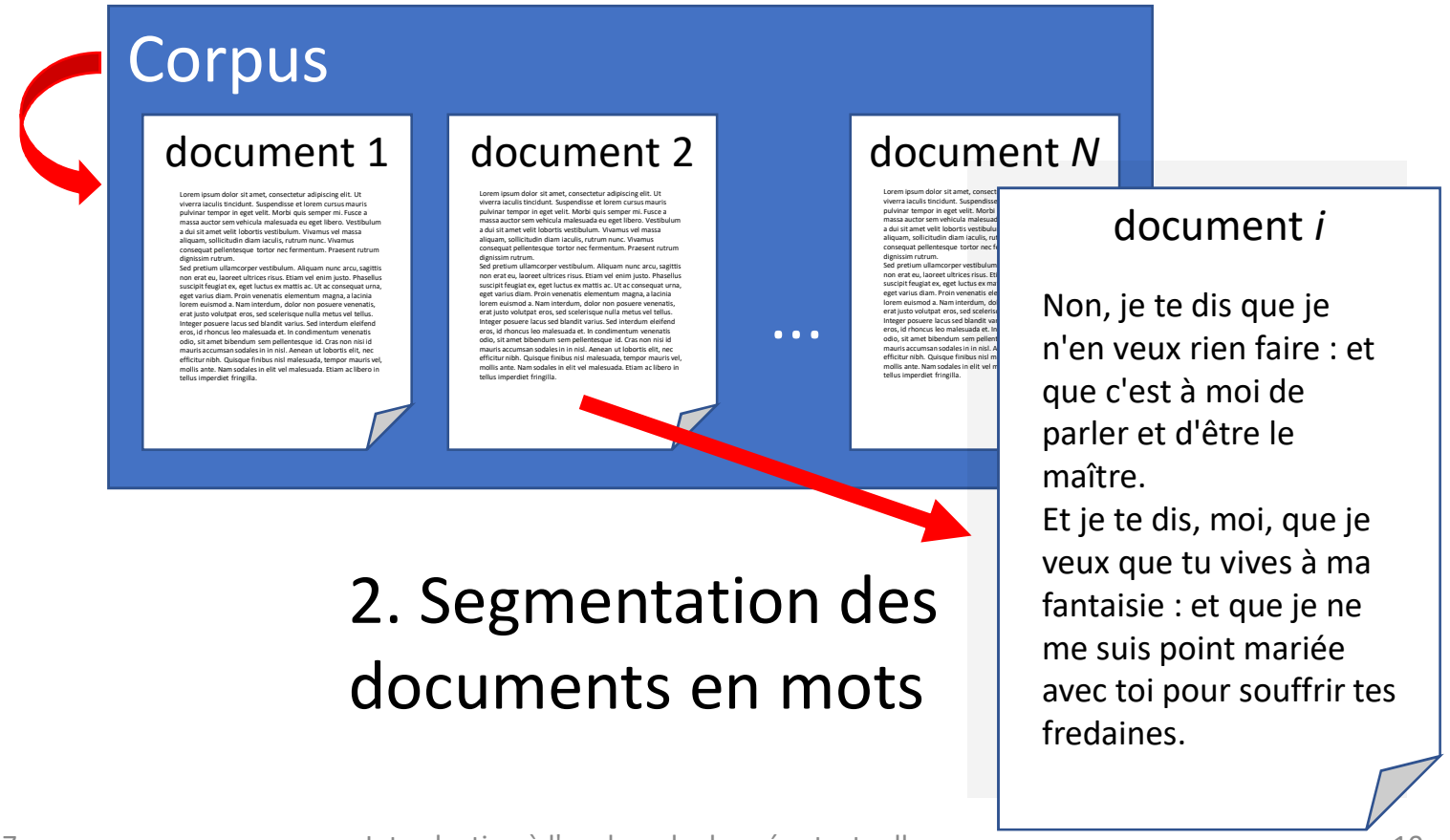
	Terme1	***Terme2***	***Terme3***	***...***
Document1	Nb. d'occurrences de *Terme1* dans *Document1*	Nb. d'occurrences de *Terme2* dans *Document1*	...	...
Document2	Nb. d'occurrences de *Terme1* dans *Document2*	...	...	...
Document3	...	...	...	...
...	...	...	...	...

Un exemple de MDT

	A	*Abandonnez*	*Acaste*	*Acceptez*	*...*	*total*
L'Avare	4	1	0	0	...	20 870
L'Ecole des femmes	4	0	0	0	...	16 494
Le Médecin malgré lui	0	0	0	0	...	9 219
Le Misanthrope	2	0	2	1	...	16 988
total	10	1	2	1	...	63 571

Données doublement segmentées

1. Segmentation du corpus en documents



2. Segmentation des documents en mots

Contextes et unités

- Contextes (contenants):

  typiquement documents, mais aussi groupe de documents ou parties de documents

- Unités (contenus):

  typiquement mots, mais aussi séquences de mots, parties de mots, lettres individuelles, etc...

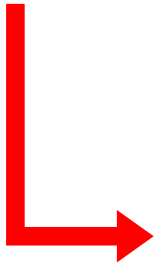
- Exemples:
  - fréquence des désinences verbales dans les pièces groupées par genre
  - fréquence des syntagmes nominaux dans les actes
  - etc.

Contenu textuel vs annotations

- Identité des segments traditionnellement basée sur leur contenu textuel, p.ex.:
  - une seule unité *le* dans *le bateau arrive* et *il le faut*
  - *vais* et *irai* considérés comme deux unités sans relation particulière
- Il est possible et parfois nécessaire de fonder l'identité des segments sur des *annotations* qui leur sont associées, p.ex.:
  - deux unités distinctes *le*|*art* vs. *le*|*pro*
  - un seul lemme (lexème/vocable) *ALLER* derrière la flexion

Une MDT peut en cacher une autre

	A	*Abandonnez*	*Acaste*	*Acceptez*	*...*
L'Avare	4	1	0	0	...
L'Ecole des femmes	4	0	0	0	...
Le Médecin malgré lui	0	0	0	0	...
Le Misanthrope	2	0	2	1	...



	Consonne	*Voyelle*
Prose	68 206	54 748
Vers	74 123	55 737

Cas particulier: fenêtre coulissante

- Construction de MDT à partir d'une seule segmentation, p.ex.:

le chien le suit, fenêtre de 2 mots

1. *le chien*
2. *chien le*
3. *le suit*

	le	*chien*	*suit*
1	1	1	0
2	1	1	0
3	1	0	1

Formats de MDT

	t1	*t2*	*t3*
d1	1	2	0
d2	0	1	1

tabulaire

document	*terme*
d1	*t1*
d1	*t2*
d1	*t2*
d2	*t2*
d2	*t3*

creux

document	*terme*	*occurrences*
d1	*t1*	1
d1	*t2*	2
d2	*t2*	1
d2	*t3*	1

creux pondéré

MDT, et après?

- Questionnements typiquement multivariés:
  - peut-on identifier des catégories de contextes ou d'unités?

    → clustering, topic models
  - peut-on visualiser les similarités entre contextes ou entre unités, ainsi que les attractions/répulsions entre contextes et unités?

    → analyse factorielle des correspondances
- Ici: mise en œuvre avec les logiciels Iramuteq et R

2. Annotations

Introduction à l'analyse de données textuelles

Définition

« L'annotation […] est la pratique consistant à ajouter des informations *interprétatives* […] à un corpus existant […], par le biais d'une forme de marquage associé […] à la *représentation* électronique des données »

Geoffrey Leech (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing, 8(4*): 275-281 (ma traduction)

Exemple

Catégories morpho-syntaxiques et lemmes:

Nous	PRO:PER	nous
misons	VER:pres	miser
sur	PRP	sur
les	DET:ART	le
transports	NOM	transport
publics	ADJ	public
et	KON	et
sur	PRP	sur
…		

Exemple (2)

Annotation selon la convention EmotionML
(2 émotions catégorisées et quantifiées):

```
<emotion>
  <category name="Disgust" value="0.82"/>
    ‘Come, there’s no use in crying like that!’
</emotion>
  said Alice to herself rather sharply;
<emotion>
  <category name="Anger" value="0.57"/>
    ‘I advise you to leave off this minute!’
</emotion>
```

Formats d'annotation

- Formats ad hoc vs standards
- XML (eXtensible Markup Language) :
  - norme d'encodage des données textuelles
  - standard
  - extensibilité
  - intégration (navigateurs, lang. de prog.)
  - ensemble de technologies (XSLT, XPath, ...)

XML: principes de base

- Balises :

```
<texte>un exemple simple</texte>
<fin_de_page></fin_de_page> → <fin_de_page/>
```

- Emboîtement (strict) :

```
<texte>
  <mot>un</mot>
  <mot>exemple</mot>
</texte>
```

- Attributs :

```
<mot lemme="exemple" cms="NOM">exemple</mot>
```

XML: document complet

- Déclaration XML
- Un et un seul élément racine (ici `<texte>`)

```
<?xml version="1.0" encoding="utf-8"?>
<texte>
  <mot type="art">un</mot>
  <mot type="nom">exemple</mot>
  <mot type="adj">simple</mot>
  et un fragment non annoté...
</texte>
```

3. Travaux pratiques

Introduction à l'analyse de données textuelles