EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

APERÇU

1. Étude de cas : @BOTUS
2. Exploration de texte et traitement automatique des langues
3. Bases de l'exploration de textes
4. Analyse de sentiments
5. Exemple : Critiques de film

[Greenstone, S. [2017]. Mettler, K. [2017]]

ÉTUDE DE CAS: @BOTUS ET T&D

D'après quelques données, les gazouillis du 45e président des États-Unis ont une incidence sur le marché boursier.

Est-ce que l'analyse de sentiments et l'intelligence artificielle (IA) peuvent être utilisées pour tirer profit en temps réel (rapidement) de la nature imprévisible de ses gazouillis?

Entre en scène **@BOTUS** du balado *Planet Money* de NPR et **Trump&Dump** de T3.

https://www.youtube.com/watch?v=ZqI89XDNqg8

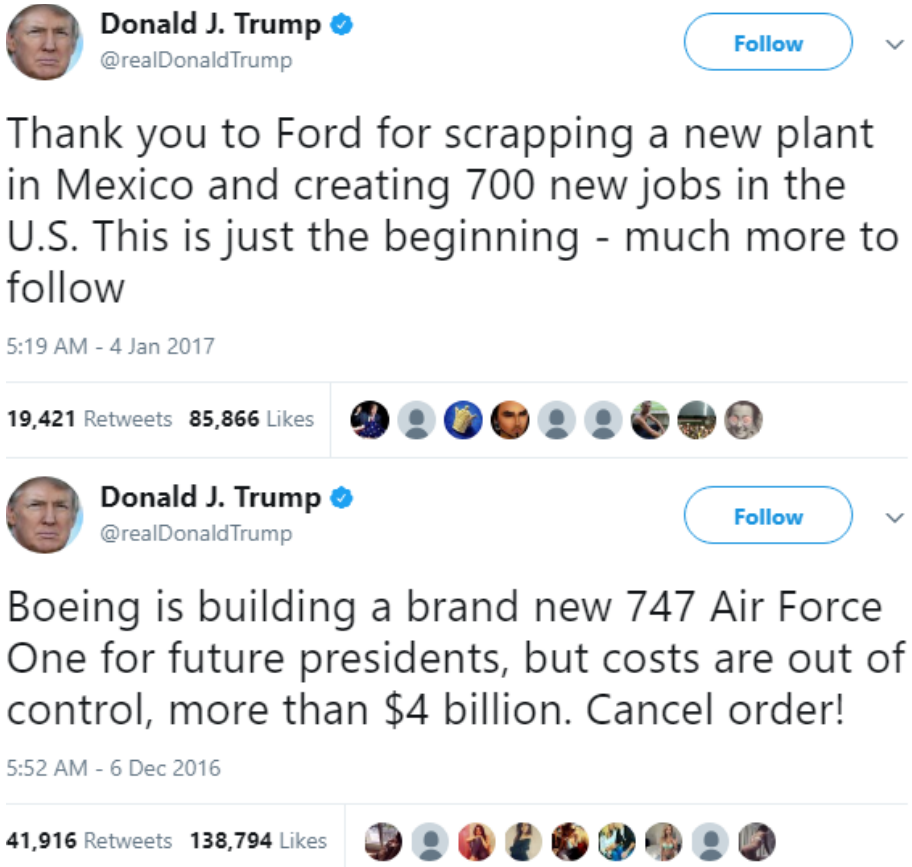
[]

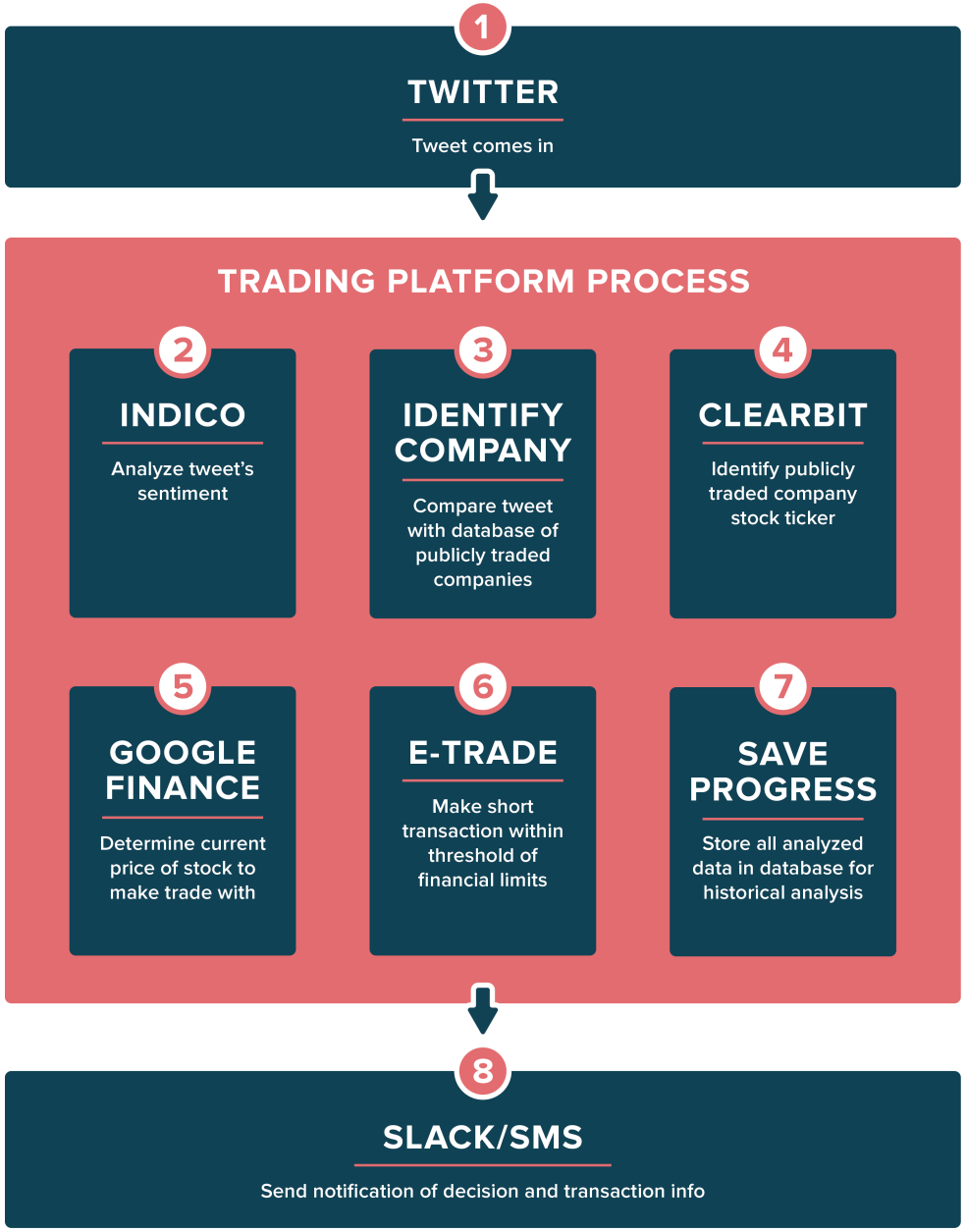
ÉTUDE DE CAS: @BOTUS ET T&D

L'analyse de sentiments (ou fouille d'opinion) est l'ensemble d'algorithmes utilisé pour déterminer l'attitude (positive, négative, neutre, etc.) de l'auteur d'un texte par rapport à un sujet ou un produit donné.

« Je ne peux pas croire que VOUS êtes le président!!! » vs « Je ne peux pas croire que vous êtes le PRÉSIDENT!!! »

ÉTUDE DE CAS: @BOTUS ET T&D





ÉTUDE DE CAS: @BOTUS ET T&D

Le président de T3 affirme que T&D est rentable, mais aucun détail n'a été fourni et le site Web a récemment été fermé.

Au cours de ses quatre premiers mois d'activités, @BOTUS n'a pas effectué une seule transaction (pour différentes raisons).

La stratégie de transactions était souple… ce qui a entraîné une perte lors de la première transaction.

Bot of the U.S.
@BOTUS
Follow

I see a company name. ✔ I know the stock ticker (AMZN) ✔ I can analyze the sentiment. ✔ (It's pretty negative). But market wasn't open. 🚫

> **Donald J. Trump** @realDonaldTrump
> The #AmazonWashingtonPost, sometimes referred to as the guardian of Amazon not paying internet taxes (which they should) is FAKE NEWS!

7:24 AM - 28 Jun 2017

Bot of the U.S.
@BOTUS
Follow

Replying to @realDonaldTrump

.@realdonaldtrump tweeted about Facebook, Inc. I shorted the stock at $168.67 and lost $0.30.

> 
>
> **Donald J. Trump** @realDonaldTrump
> Thank you Arizona. Beautiful turnout of 15,000 in Phoenix tonight! Full coverage of rally via my Facebook at: facebook.com/DonaldTrump/vi...

7:01 AM - 23 Aug 2017

ÉTUDE DE CAS: @BOTUS ET T&D

Réussites :

- Présentation d'analyses de sentiments bien exécutées
- Simulation d'un processus qui trouve la meilleure stratégie de transactions

Mais, n'est pas aussi bon qu'un outil de **prévision** (sans lien avec l'exploration de texte et le traitement automatique des langues).

L'analyse des données descriptives peut expliquer ce qui s'est produit.

Les hypothèses de modélisation ne sont pas toujours applicables dans le monde réel (domaine prédictif).

EXPLORATION DE TEXTE VS TRAITEMENT AUTOMATIQUE DES LANGUES

L'**exploration de texte** est l'ensemble de processus quantitatifs par lesquels nous essayons d'extraire des renseignements **utiles** (exploitables) à partir d'un texte.

À bien des égards, l'exploration de texte porte sur la transition d'un état **désorganisé** à un état **organisé** (données non structurées à données structurées). Le traitement automatique des langues consiste à faire réagir les machines de façon **« appropriée »** lorsqu'elles interagissent avec du langage naturel.

Dans le cadre du présent cours :

- L'**exploration de texte** renvoie à l'application de tâches liées à la science des données à des données texte.
- Le **traitement automatique des langues** est réservé aux tâches qui cherchent à « comprendre » les langues.

APPLICATIONS DE L'EXPLORATION DE TEXTE

Classification

- Questions sur l'auteur, distinction entre les énoncés vrais ou faux, etc.

Estimation de la valeur

- Analyse de sentiments, détection d'un préjugé, etc.

Agrégation

- Modélisation des sujets, récupération des renseignements et recommandations, etc.

Autres

- Description du texte, visualisation du texte, etc.

COMPRENDRE LE LANGAGE

Syntaxe

- Lemmatisation, marquage des parties du discours, désambiguïsation des limites d'une phrase, etc.

Sémantique

- Traduction automatique, génération de langage, reconnaissance d'entités nommées, segmentation des sujets, questions et réponses, etc.

Discours

- Analyse du discours, récapitulation, etc.

Parole

- Reconnaissance, segmentation, synthèse texte-parole, etc.

L'EXPLORATION DE TEXTE EST FACILE, LE TRAITEMENT AUTOMATIQUE DES LANGUES EST IA-COMPLET

[https://www.youtube.com/watch?time_continue=13&v=EdsIMJg1m2E]

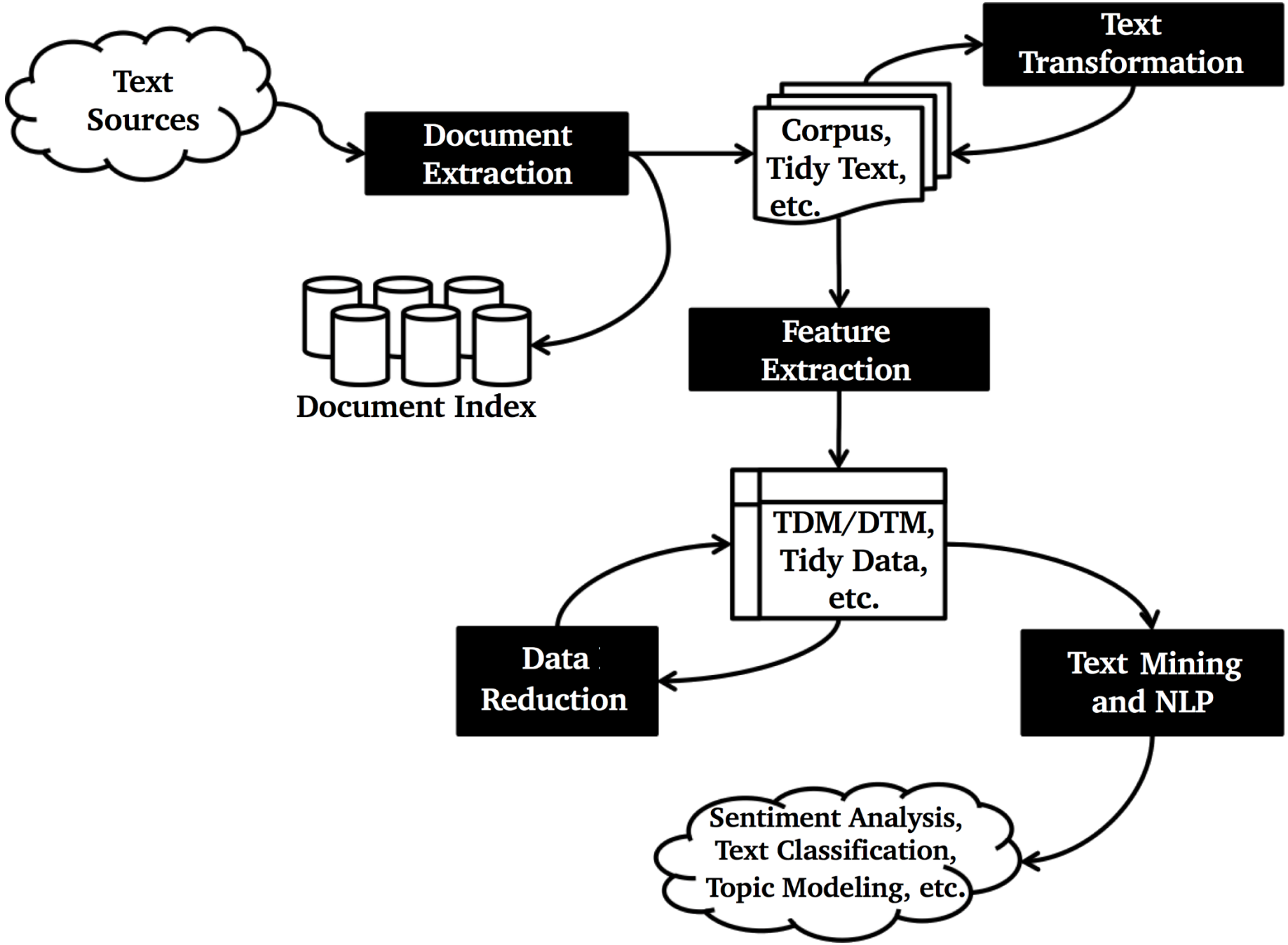
TRADUCTION AUTOMATIQUE

J'ai été au sud du sud au soleil
Bleu blanc rouge les palmiers
Et les cocotiers glacés
Dans les pôles aux Esquimaux bronzés
Qui tricotent des ceintures fléchées
Farcies
Et toujours la Sophie
Qui venait de partir

(*Lindberg*, R. Charlebois)

I was south of south in the sun
Blue white red palm trees
And frozen coconut palms
In the poles to the tanned Eskimos
Who knit arrow belts
Stuffed
And always Sophie
Who had just left

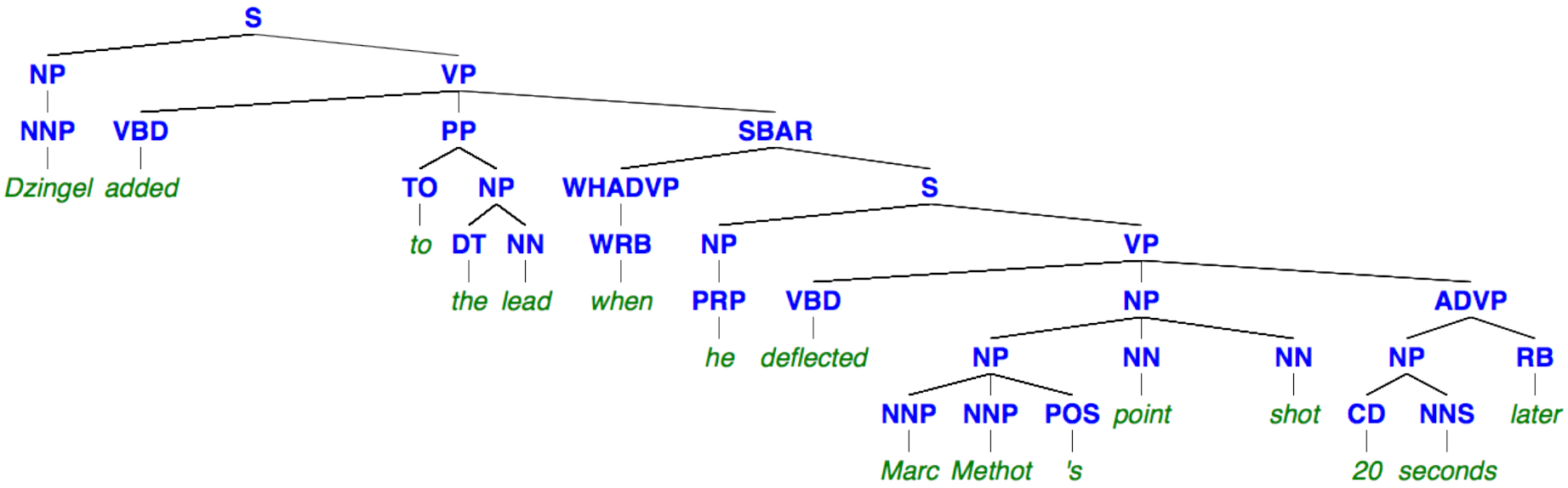
???


Text Sources
Document Extraction
Text Transformation
Corpus, Tidy Text, etc.
Document Index
Feature Extraction
TDM/DTM, Tidy Data, etc.
Data Reduction
Text Mining and NLP
Sentiment Analysis, Text Classification, Topic Modeling, etc.

ANALYSE SYNTAXIQUE DE LA SÉMANTIQUE

Le processus de conversion d'une phrase en langage naturel vers une **représentation rigoureuse du sens**.

L'**ordre** et le **type**/rôle du mot définissent les **attributs** du mot.

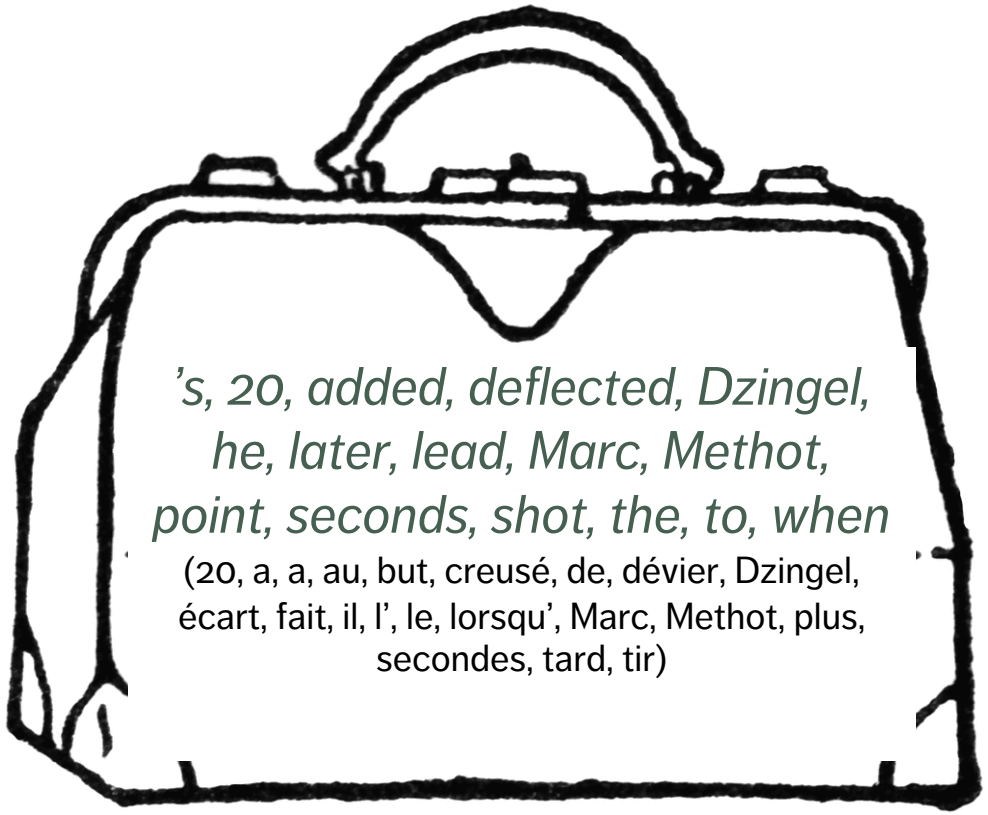


SAC DE « MOTS »

Seule la **présence** (ou l'**absence**) de « mots » (racines, *n*-grammes, phrases, etc.) est importante.

Les **fréquences** relatives donnent de l'information (intention, thème, sentiment, etc.) sur le corpus.

Les mots **eux-mêmes** sont des attributs du document.



TRAITEMENT DE TEXTE

Les données texte nécessitent un nettoyage exhaustif et un traitement complexe.

La nature des données soulève de nombreuses difficultés :

- Qu'est-ce qu'une anomalie dans le texte?
- Qu'est-ce qu'une observation aberrante?
- Est-il possible de définir ces concepts?
- Que faire en cas d'erreur d'encodage?

Les fautes d'orthographe et les erreurs typographiques sont difficiles à relever dans les documents volumineux, même au moyen d'un correcteur orthographique.

TRAITEMENT DE TEXTE

Le processus peut être simplifié dans une certaine mesure à l'aide d'**expressions rationnelles** et de **fonctions de prétraitement de texte**.

Les étapes propres au prétraitement varient selon le problème :

- Le *jargon* des utilisateurs de *Twitter* diffère de la *langue de bois* des juristes
- De même, un enfant qui apprend à parler et un candidat au doctorat n'ont pas le même vocabulaire

Comme presque tout ce qui touche à l'exploration de texte, le processus de nettoyage **dépend grandement du contexte**.

Veuillez noter que l'ordre des tâches de prétraitement peut avoir une incidence sur les résultats.

TRAITEMENT DE TEXTE – OPTIONS

Convertir toutes les lettres **en minuscules** (à éviter pour rechercher des noms)

Retirer tous les **signes de ponctuation** (à éviter pour rechercher des émojis)

Supprimer tous les **chiffres** (à éviter pour explorer des quantités)

Supprimer tous les **espaces blancs superflus**

Supprimer tous les **caractères entre crochets** (à éviter pour rechercher des balises)

Remplacer tous les chiffres par des **mots**

TRAITEMENT DE TEXTE – OPTIONS

Remplacer les **abréviations**

Remplacer les **contractions** (éviter pour rechercher des paroles informelles)

Remplacer tous les **symboles par des mots**

Supprimer tous les **mots vides** ou **non informatifs** (selon la langue, l'ère et le contexte)

Utiliser des **mots racines** et des **racines complètes** pour supprimer les variations vides

- « conductif », « conductible », « conductibilité », « conducteur » sont porteurs du sens de « conduction »
- dans « recherche opérationnelle », « systèmes opérationnels » et « dentisterie opératoire », la racine « opérat » représente des **sens différents**

TRAITEMENT DE TEXTE

Représentation de l'accent phonétique
eille chus à boutte là, écoute-moé!

Néologismes et mots-valises
Mais quel adulescent!

Mauvaises traductions/mots étrangers

Calembours et jeux de mots

Mots-clés, balises et texte non informatif
<b>; \includegraphics; résumé ISBN

Vocabulaire spécialisé
logithèque; codec; Turboencabulator

Noms et lieux fictifs
Qo'noS; Kilgore Trout

Argot et jurons
fou raide; #$&#!

REPRÉSENTATION TEXTUELLE

Le texte doit être stocké dans les structures de données avec les propriétés adéquates :

- une **chaîne** ou un vecteur de caractères, avec un encodage propre au langage
- un **corpus** (une collection) de documents texte (avec des métadonnées)
- une **matrice document-terme** où les rangées sont les documents, les colonnes sont les termes et les entrées sont une statistique texte appropriée (ou la **matrice terme-document** transposée)
- un **jeu de données texte organisé** avec un **jeton** (uniterme, *n*-gramme, phrase, paragraphe) par rangée

Il n'y a pas de formule magique : le meilleur format dépend du problème encouru. Mais cette étape est **essentielle**, tant pour l'analyse sémantique que le sac de mots.

STATISTIQUES TEXTE

Prenons un corpus $\mathcal{C} = \{d_1, ..., d_N\}$ qui comporte N **documents** et M **termes** de sac de mots $\mathcal{C} = \{t_1, ..., t_M\}$.

Par exemple, si

$$\mathcal{C} = \left\{ \begin{array}{c} \text{“the dogs who have been let out”}, \\ \text{“who did that”}, \\ \text{“my dogs breath smells like dogs food”} \end{array} \right\},$$

(Traductions : « les chiens qui sont sortis », « qui a fait ça », « l'haleine de mon chien sent la moulée »)

alors

$$N = 3, d_1 = \text{“the dogs who have been let out”},$$

$$d_2 = \text{“who did that”}, d_3 = \text{“my dogs breath smells like dogs food”}$$

STATISTIQUES TEXTE

La **fréquence relative d'un terme** de t dans d est

$$tf^*_{t,d} = \frac{\text{nombre de fois que } t \text{ se répète dans } d}{M_d}$$

La **fréquence relative d'un document** de t est

$$df^*_t = \frac{\text{nombre de documents dans lesquels } t \text{ se répète}}{N} = \frac{\sum_d \text{sign}(tf^*_{t,d})}{N}$$

STATISTIQUES TEXTE

La **fréquence de terme – fréquence de document inverse** de t est dans d est

$$tf\text{–}idf^*_{t,d} = -tf^*_{t,d} \times \ln(df^*_t)$$

$tf\text{–}idf^*_t$		t													
		1 been	2 breath	3 did	4 dogs	5 food	6 have	7 let	8 like	9 my	10 out	11 smells	12 that	13 the	14 who
d	1	0.16	0	0	0.06	0	0.16	0.16	0	0	0.16	0	0	0.16	0.06
	2	0	0	0.37	0	0	0	0	0	0	0	0	0.37	0	0.14
	3	0	0.16	0	0.12	0.16	0	0	0.16	0.16	0	0.16	0	0	0

STATISTIQUES TEXTE

Si **tous les documents** contiennent le terme t, alors $\mathrm{df}_t^* = 1$ et

$$tf\text{–}idf_{t,d}^* = -tf_{t,d}^* \times \ln(1) = 0$$

(ce terme ne fournit pas d'information)

Si un terme t **apparaît rarement** dans un document d, alors $tf_{t,d}^* \approx 0$ et

$$tf\text{–}idf_{t,d}^* \approx -0 \times \ln(df_t^*) \approx 0.$$

Les termes qui apparaissent relativement souvent seulement dans un petit sous-ensemble de document sont essentiels à la compréhension de ces documents **dans le contexte général** du corpus.

BASES DE L'ANALYSE DES SENTIMENTS

La plupart d'entre nous avons une bonne compréhension innée de l'intention émotionnelle des mots, ce qui nous permet de présumer **la surprise, le dégoût, la joie, la douleur,** etc. à partir d'un segment de texte.

Le processus, lorsqu'il est appliqué par des machines à un bloc de texte, s'appelle **l'analyse de sentiments** (fouille d'opinion).

Questions typiques de l'analyse de sentiments :

- « Cette critique de film est-elle positive ou négative? »
- « Ce courriel d'un client est-il une plainte? »
- « Est-ce que l'attitude des journaux au sujet du premier ministre a changé depuis les élections? »

DIFFICULTÉS

La plupart des humains seraient **habituellement** en mesure de répondre à ces questions s'ils avaient en main les documents texte appropriés. Pour les machines, ce problème n'est pas facile à résoudre.

Difficultés :

- Nous ne nous entendons pas toujours sur le contenu émotionnel d'un texte
- Les mots peuvent avoir une signification/valeur émotionnelle différente selon le contexte (anti-antonymes)
- Les qualificatifs peuvent changer drastiquement la valeur émotionnelle d'un terme
- Les changements de sujet
- Figures de rhétorique

TÂCHES CONNEXES

L'analyse de sentiments est un problème d'**apprentissage supervisé**, qui nécessite des dictionnaires de contenu émotionnel compilés au préalable (à l'interne ou à l'externe).

Tâches connexes :

- Rejeter l'information subjective (extraction de l'information)
- Reconnaître les questions axées sur des opinions (réponse aux questions)
- Tenir compte de nombreux points de vue (résumé)
- Déterminer si les vidéos conviennent aux enfants, s'il y a des partis pris dans les sources de nouvelles et si le contenu est approprié pour un placement publicitaire

Élément de **subjectivité**

TYPES D'ANALYSE DE SENTIMENTS

Dans le présent cours, nous faisons la distinction entre deux types d'analyse de sentiments :

- l'analyse **terme par terme** évalue le contenu émotionnel de jetons et essaie de déduire une note pour les passages qui les contiennent;
- l'analyse **document par document** évalue les passages notés et essaie de trouver les jetons qui portent la charge émotionnelle ou de prédire quelle note serait attribuée à un nouveau passage sur un spectre émotionnel.

L'analyse terme par terme n'est pas une tâche technique complexe : elle nécessite seulement la capacité de faire correspondre une note de lexique à un terme, et de faire la somme des notes.

L'analyse document pas document est, à la base, un problème de classification. Elle nécessite des données texte étiquetées, mais le principe est exactement le même : prédire les étiquettes **« positives/négatives »** (consulter les exercises).

LEXIQUES DE SENTIMENTS

L'analyse de sentiments terme par terme repose largement sur des **lexiques**, c'est-à-dire des listes de termes qui ont été classés sur une échelle émotionnelle.

- AFINN : Les mots sont placés sur une échelle qui va de -5 (négatif) à 5 (positif)
- BING : Binaire négatif/positif
- NRC : Les mots se voient attribuer une ou des catégories de sentiments
- LOUGHRAN : Contenants catégoriques

Chacun de ces lexiques contient une majorité de termes **négatifs**.

La sélection du meilleur lexique est dictée par le **contexte**.

LEXIQUES DE SENTIMENTS

« abandon »

AFINN : -2
BING : S.O.
NRC : peur, négatif, tristesse
LOUGHRAN : négatif

« mauvais »

AFINN : -3
BING : négatif
NRC : colère, dégoût, peur, etc.
LOUGHRAN : négatif

« pas »

AFINN : S.O.
BING : S.O.
NRC : S.O.
LOUGHRAN : S.O.

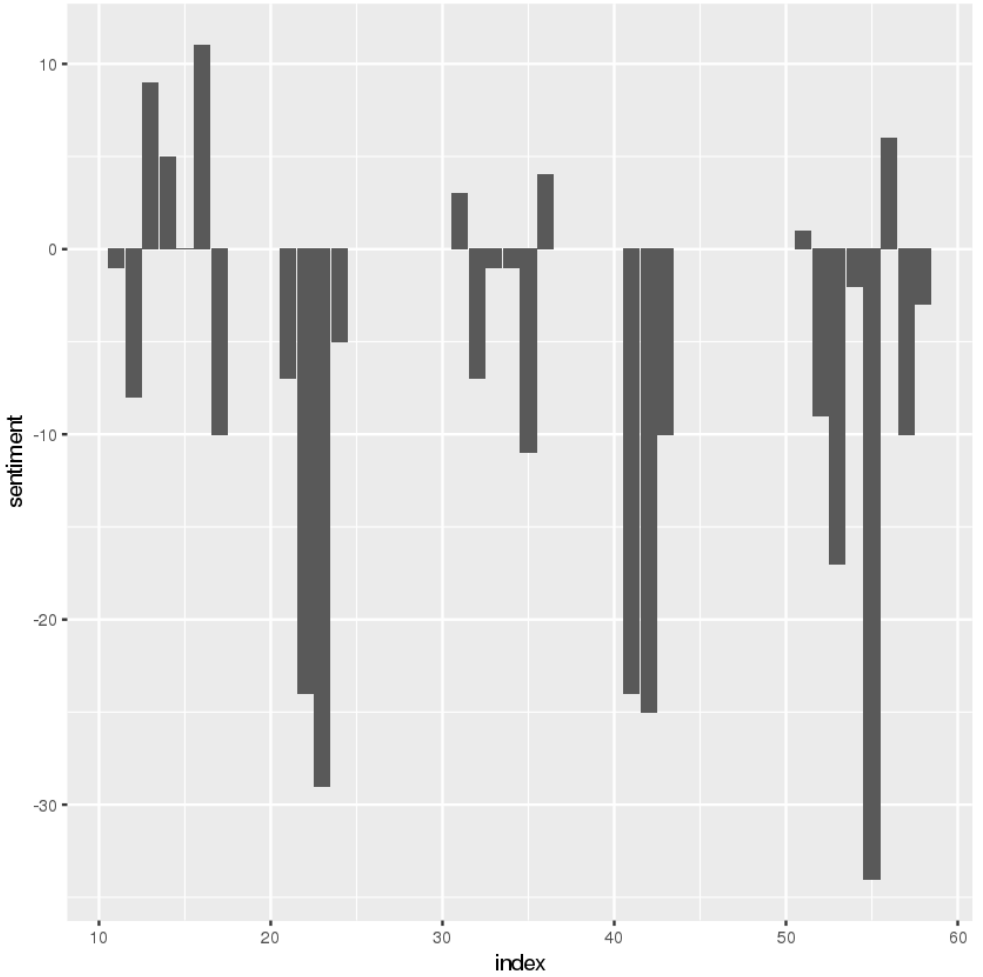
« flagrant »

AFINN : ?
BING : ?
NRC : ?
LOUGHRAN : ?

LEXIQUES DE SENTIMENTS

Une fois qu'un lexique est sélectionné, l'analyse terme par terme s'effectue tout simplement **en morcelant le texte** et en calculant les notes de sentiments pour chaque bloc (environ 100 mots, chaque 100 lignes, chaque chapitre, etc.).

Y a-t-il des raisons de s'attendre à ce que les différents lexiques donnent les mêmes notes?



(*Macbeth* par Shakespeare, notes par scène selon le lexique AFINN)