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

ÉTUDE DE CASE : @BOTUS ET TRUMP&DUMP

EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

Analyse de sentiments de gazouillis

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

@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

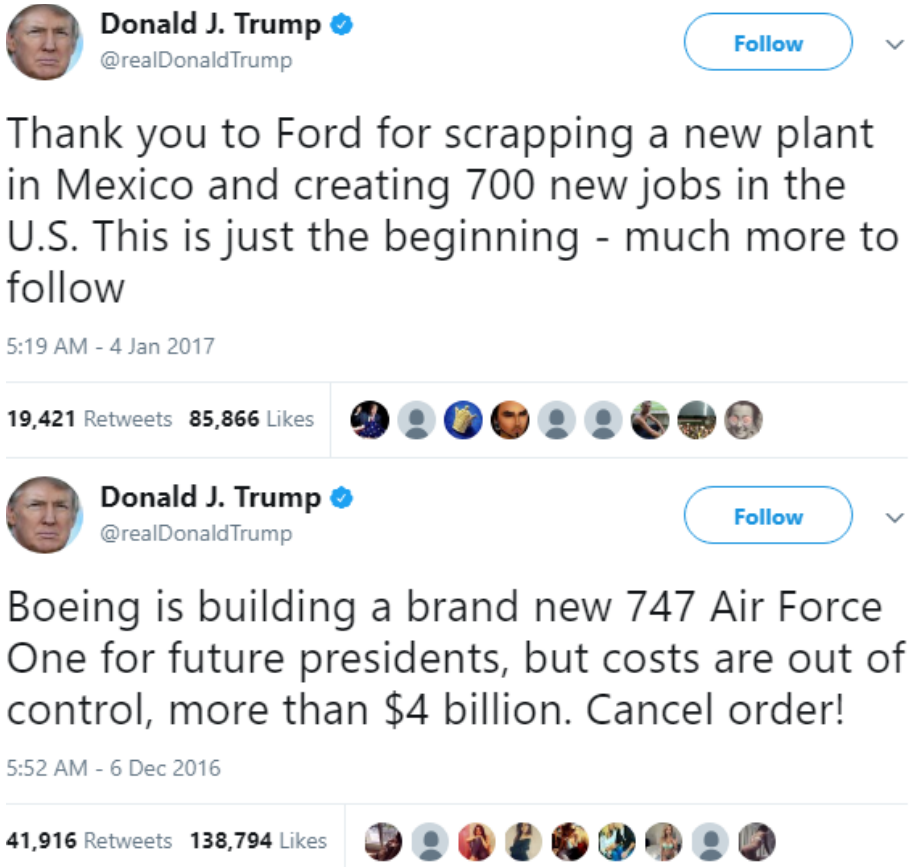
[]

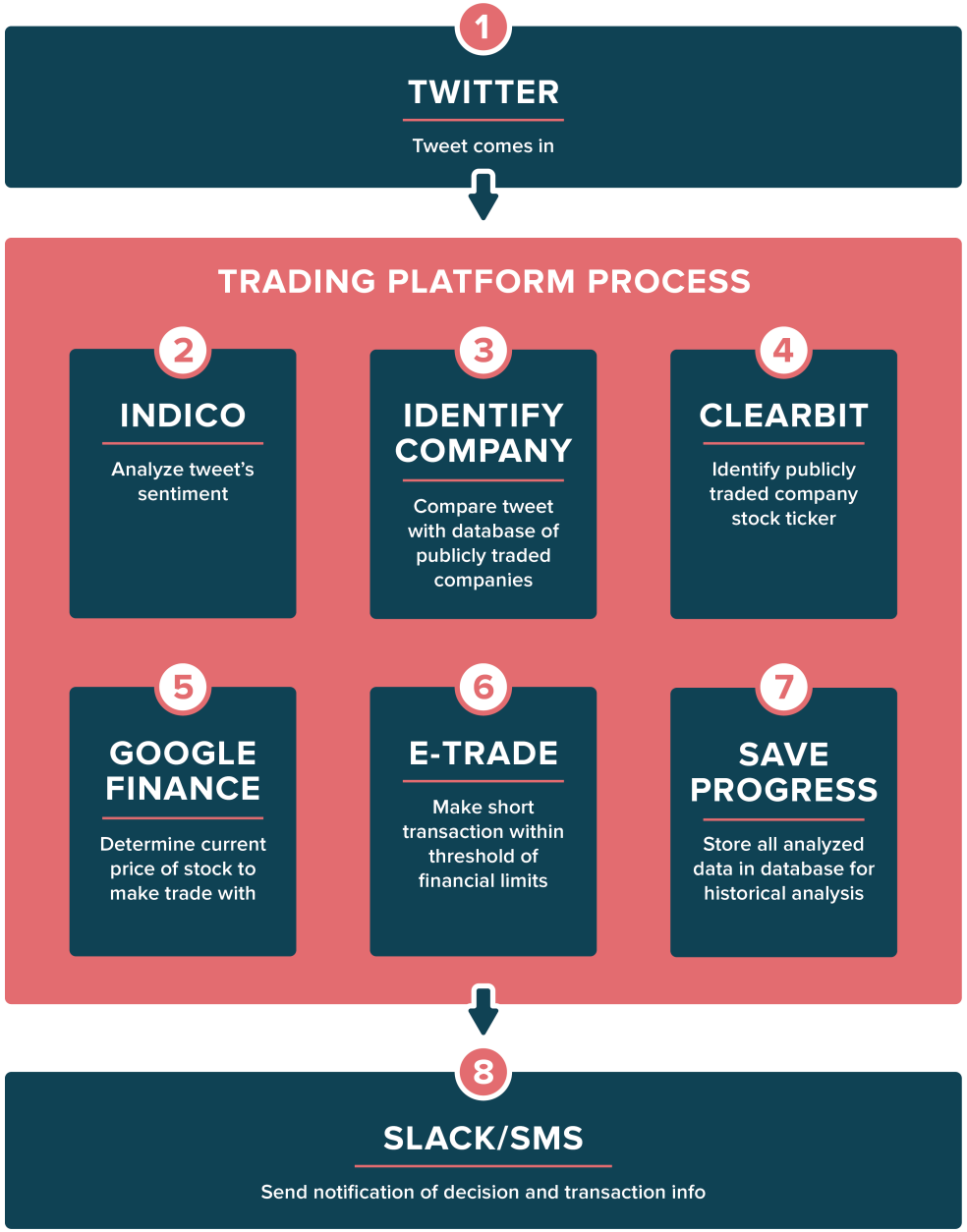
@BOTUS ET T&D

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!!! »

@BOTUS ET T&D





IDLEWYLD Sysabee DAVHILL uOttawa

data-action-lab.com

@BOTUS ET T&D

Les langages naturels sont riches, adaptables et permettent les variations syntaxiques (avantage pour les humains, désavantage pour les robots logiciels).

La signification d'un mot peut **dépendre** grandement **du contexte**.

Sarcasme, expressions idiomatiques, figures de style… les humains ne les détectent pas toujours.

Reconnaissance d'entités nommées : Le Château (entreprise) vs le château (bâtiment).

@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

@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).

DISCUSSION

Quelle est l'importance des indices visuels dans les communications et les négociations d'affaires? Quelle est l'importance du contexte?

Dans le même ordre d'idée, dans quelle mesure est-il facile d'apprendre d'une personne dont le contexte est différent du vôtre (plans culturel ET professionnel)?

EXPLORATION DE TEXTE ET TRAITEMENT AUTOMATIQUE DES LANGUES

EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

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

LE RÊVE DANS LE PAVILLON ROUGE (红楼梦)

宝玉道：“一言难尽.” 说者便把梦中之事细说与袭人听了. 然后说至警幻所授云雨之情，羞得袭人掩面伏身而笑.
(texte original par Cao Xueqin)

« C’est une longue histoire, » répondit Pao-yu, avant de raconter son rêve du début jusqu’à la fin, en concluant par son initiation au « sport des nuages et de la pluie » auprès de Désillusion. His-jen, entendant cela, se couvrit le visage et se mit à rire.
(traduction française de la traduction par Yang Xianyi)

Après avoir beaucoup hésité, il lui fit un récit détaillé de son rêve. Mais lorsqu’il lui raconta la partie où il faisait l’amour à Deux-en-un, Aroma fut prise d’un fou rire et cacha son visage dans ses mains.
(*traduction française de la traduction par D. Hawkes*)

Bao Yudao : « C’est difficile de dire un mot. » L’orateur a ensuite parlé de ce qui s’était passé dans le rêve et a écouté les gens. Il a ensuite déclaré que la police lui avait donné la sensation de nuage et de pluie et qu’il avait honte de cacher son visage et de rire.
(*traduction automatique*)

[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)

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

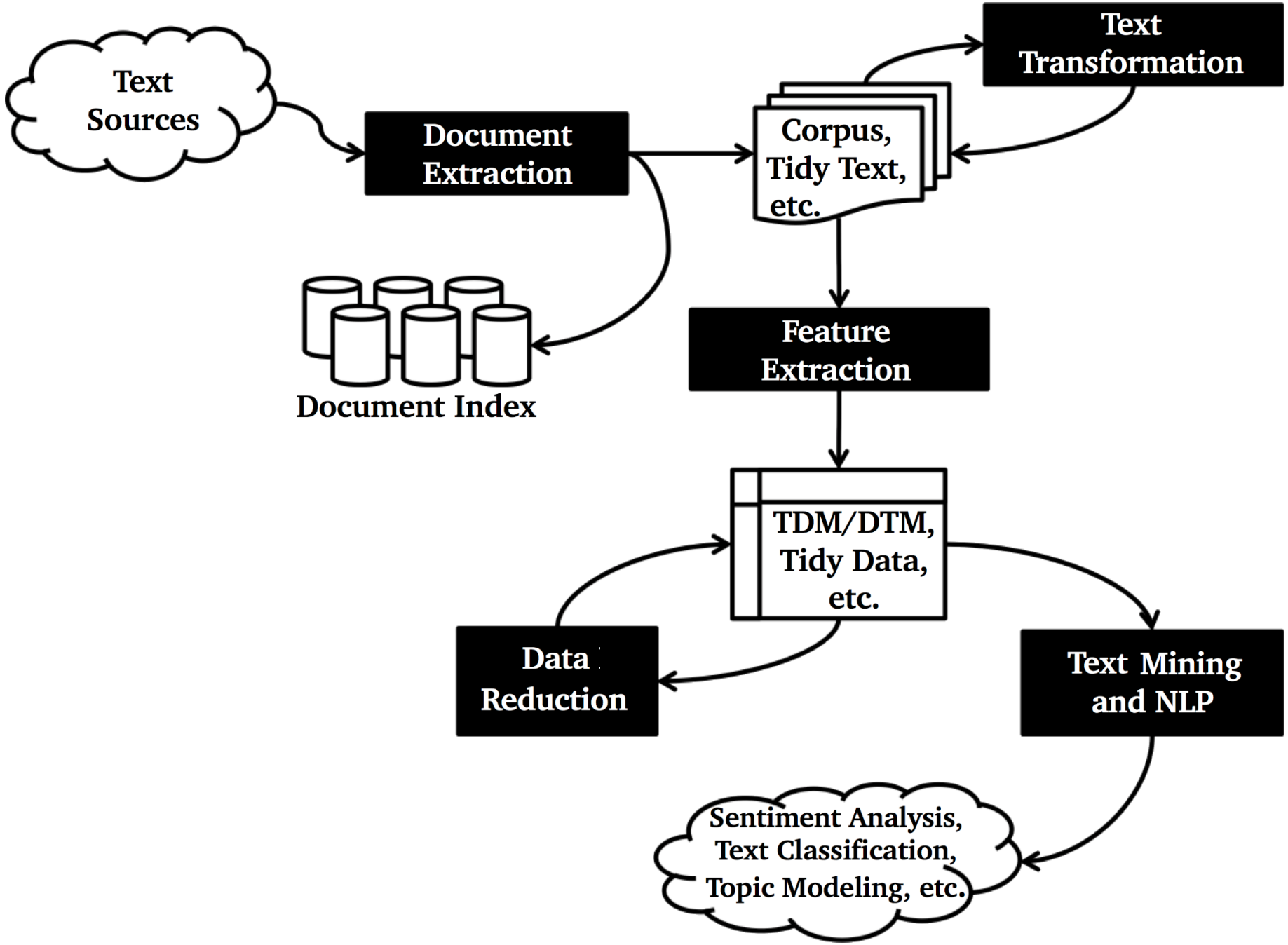
TRADUCTION AUTOMATIQUE

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.

DISCUSSION

Dans l'analyse de données numériques, il peut être difficile (même pour les experts), de déterminer lorsque les résultats n'ont aucun sens. Ce n'est pas le cas de l'analyse de données texte, puisque la plupart d'entre nous peuvent détecter d'un coup d'œil que quelque chose cloche.

Quelles mesures pouvez-vous prendre pour détecter les résultats sans queue ni tête afin qu'ils ne soient pas diffusés trop tôt?

BASES DE L'EXPLORATION DE TEXTE

EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

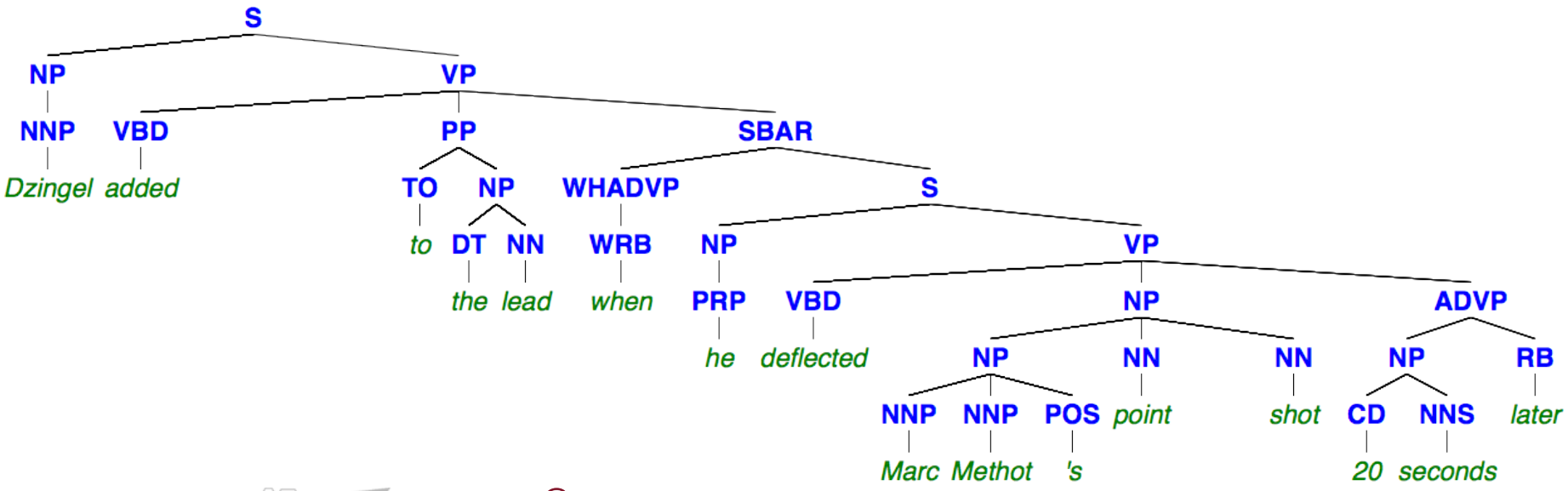
« Dzingel added to the lead when he deflected Marc Methot's point shot 20 seconds later. » (« *Dzingel a creusé l'écart lorsqu'il a fait dévier le tir au but de Marc Methot 20 secondes plus tard.* »)

(Associated Press, récapitulation de la joute entre les Sénateurs d'Ottawa et les Maple Leafs de Toronto le 18 février 2017)

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.

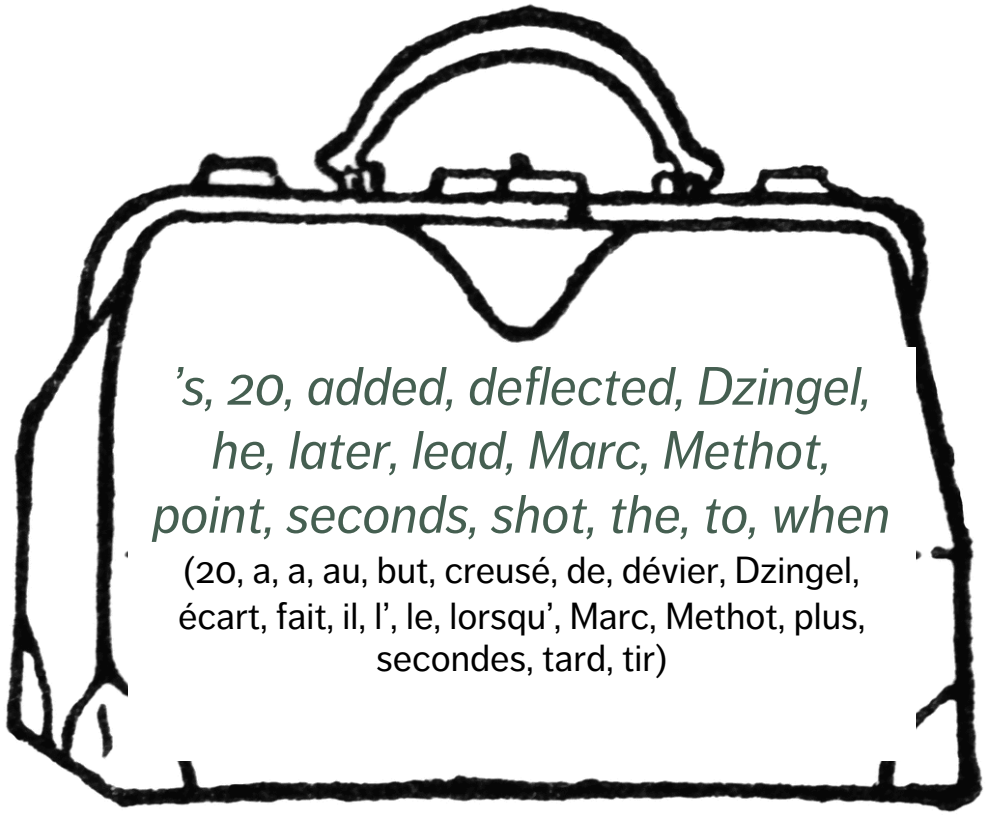


SAC DE « MOTS »

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

« Dzingel added lead deflected
Marc Methot point shot twenty seconds later »
« Dzingel creuser écart dévier tir
but Marc Methot vingt secondes plus tard »

---

added, deflected, Dzingel, later, lead, Marc, Methot, point, seconds, shot, twenty
but, creuser, dévier, Dzingel, écart, Marc, Methot, plus, secondes, tard, tir, vingt

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; #$&#!

EXERCICE

Comment traiteriez-vous ce court texte?

« <i>Il<i> est allé se coucher à 2 h. C\ » est beaucoup trop tard! Il était seulement endormi à 20 % au début, mais il a fini par s'endormir. »

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.

REPRÉSENTATION MATRICE DOCUMENT-TERME ET MATRICE TERME-DOCUMENT

	Document 1	Document 2	Document 3	…	Document *N*
Token 1	0	0	1	62	3
Token 2	0	1	0	61	2
Token 3	1	0	3	101	0
…	112	24	38	84	0
Token *M*	2	2	0	12	3
Sum	115	27	42	320	8

Sum
66
64
105
258
19

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

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

$tf^*_{t,d}$		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	1/7	0	0	1/7	0	1/7	1/7	0	0	1/7	0	0	1/7	1/7
	2	0	0	1/3	0	0	0	0	0	0	0	0	1/3	0	1/3
	3	0	1/7	0	2/7	1/7	0	0	1/7	1/7	0	1/7	0	0	0

STATISTIQUES TEXTE

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

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

df_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
	1/3	1/3	1/3	2/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3	2/3

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.

DISCUSSION

À l'étape de l'analyse, il est facile d'oublier d'où proviennent les données et ce à quoi elles s'appliquent réellement.

À l'arrivée le texte est non structuré et non organisé. Après le traitement, le texte est nettoyé, mais encore non structuré. Le sac de mots offre un cadre de représentation numérique structuré d'un texte.

Quelle est l'incidence sur le choix de statistique texte dans la matrice document-terme/matrice terme-document?

tf-idf n'est pas toujours idéal... une approche se fondant sur les **rapports de cotes pondérés** (« weighted log odds ») pourrait s'avérer préférable par moment.

ANALYSE DE SENTIMENTS

EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

« [...] classifier les messages sur les réseaux sociaux à la main n'est pas pratique à grande échelle (même si certaines firmes classent des échantillons à la main pour améliorer leurs algorithmes). Un simple pouce en l'air ou en bas comme "sentiment" est pire que simplement dénué de sens—il n'est simplement pas vrai. »

(S. Kessler, *The Problem With Sentiment Analysis*)

BASES

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**

EXERCICE

Trois évaluations (1 étoile, 3 étoiles, 5 étoiles) trouvées sur Amazon.ca. Pouvez-vous les identifier? Pouvez-vous déterminer le produit?

- « J'aime les jeans, le prix, la coupe, mais plus encore, j'aime les vendeurs. Non seulement ont-ils répondu immédiatement à toutes mes préoccupations, ils ont largement surpassé mes attentes! Je vais certainement effectuer d'autres achats par leur intermédiaire, fortement recommandé! »
- « N'ACHETEZ PAS. Même s'il s'agit d'une excellente série, cet ajout spécial est minable. Il s'agit en fait de livres brochés de grande diffusion : petits et inconfortables à tenir. Les versions régulières sont de loin supérieures et coûtent la même chose. »
- « À partir de la deuxième utilisation, le bol tombait constamment 30 secondes après le début du malaxage. Un peu déçu. »

[http://aiplaybook.a16z.com/docs/guides/nlp#user-content-apiexamples]

"Love the jeans, price, fit, but even more, love the suppliers. Simple concerns were not only answered immediately, they went beyond any expectations I had! Will definitely be buying through this supplier, highly recommended!"

Scores

API Name	Result	Total request time	API Time
+ Sentiment.JS (node.js library)	very positive (90)	299	0
+ Sentimental (node.js library)	very positive (90)	289	0
+ IBM Alchemy Language API	-1	341	43
+ IBM Watson Developer Cloud	positive (72)	1472	1128
+ Google Cloud APIs	-1	651	351
+ Microsoft Azure Cognitive Services	very positive (93)	789	482

[http://aiplaybook.a16z.com/docs/guides/nlp#user-content-apiexamples]

"DON'T BUY. Great series aside, this special addition is pathetic. They're basically mass-market paperbacks: small and uncomfortable to hold. The regular paperback versions are far superior for basically the same price."

Scores

API Name	Result	Total request time	API Time
+ Sentiment.JS (node.js library)	neutral (10)	163	0
+ Sentimental (node.js library)	neutral (10)	157	1
+ IBM Alchemy Language API	-1	240	79
+ IBM Watson Developer Cloud	neutral (-3)	1164	922
+ Google Cloud APIs	-1	436	218
+ Microsoft Azure Cognitive Services	negative (-69)	788	609

[http://aiplaybook.a16z.com/docs/guides/nlp#user-content-apiexamples]

"Beginning the second use, the bowl keeps falling out 30 seconds after the mixing starts. A bit disappointed."

Scores

API Name	Result	Total request time	API Time
+ Sentiment.JS (node.js library)	negative (-30)	65	0
+ Sentimental (node.js library)	negative (-30)	66	0
+ IBM Alchemy Language API	-1	379	211
+ IBM Watson Developer Cloud	neutral (-6)	1300	1026
+ Google Cloud APIs	-1	408	281
+ Microsoft Azure Cognitive Services	very negative (-78)	684	486

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 »**.

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.

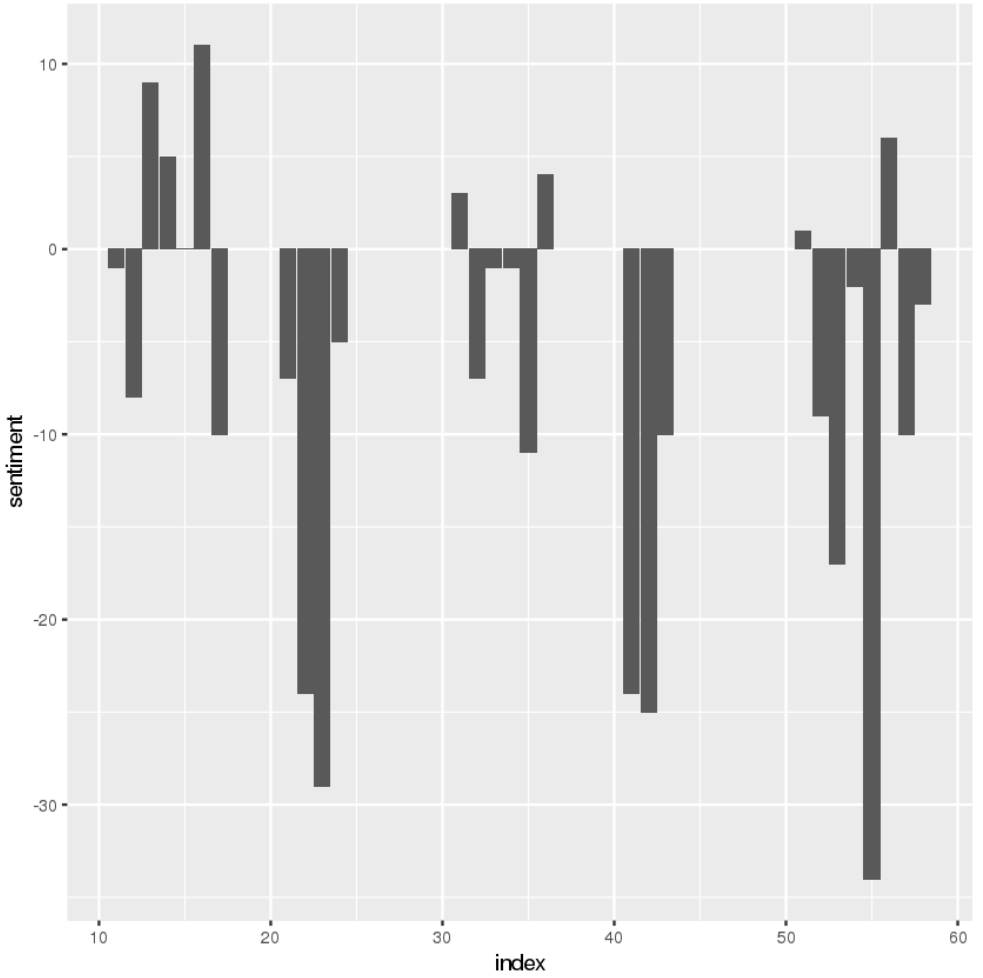
« flagrant »

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)

DISCUSSION

La plupart des mots de l'anglais sont neutres. Pourquoi est-ce que la plupart des mots qu'on trouve dans les lexiques sont négatifs? Est-ce aussi le cas pour les langues chinoises écrites? Pour le français?

Les lexiques sont-ils interchangeables (ère, culture, contexte)?

EXEMPLE : CRITIQUES DE FILM

EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

« Aucun bon film n'est trop long et aucun mauvais film n'est assez court. »

(Roger Ebert)

ÉNONCÉ DE L'EXERCICE

L'accent de cet exercice n'est pas mis sur la programmation, mais plutôt sur un flux de travail analytique complet au moyen de la trousse d'outils de langages naturels (NLTK) de Python.

Le but est de développer un modèle d'analyse de sentiments pour les critiques de film. Le jeu de données contient 50 000 critiques étiquetées **positive** ou **négative**.

Un modèle de sentiments exact nous permettra, par exemple, de classer automatiquement les nouvelles critiques afin de rassembler les données sur les critiques.

EXERCICE

1. **Renseignements sur le jeu de données**
   - Combien y a-t-il de critiques positives dans le jeu de données de formation? Combien de critiques négatives?
   - Combien y a-t-il de critiques positives dans le jeu de données d'essai? Combien de critiques négatives?
   - Quelle est la fourchette de notes pour les critiques positives et négatives dans les jeux de données de formation et d'essai?
   - Quels peuvent être les conséquences de l'absence de critiques neutres dans les jeux de données de formation et d'essai?
2. **Préparation des données**
   - Sélectionnez dix mots qui, selon vous, expliquent pourquoi la critique de *Haunted Boat* (3446_1.txt) est une critique une étoile.
   - Sélectionnez dix mots qui, selon vous, expliquent pourquoi la critique de *Night Listener* (10015_8.txt) est une critique huit étoiles.

EXERCICE

3. **Traitement du sac de mots**
   - Tous les jetons traités dans toutes les critiques du jeu de données de formation sont utilisés pour créer une matrice document-terme. Décrivez le processus pour passer d'un texte de critique complet à des jetons de la critique.
   - Combien de jetons sont conservés dans la matrice document-terme?
   - En quoi ce nombre dépend-il de la nature du générateur de jetons? (Veuillez noter l'erreur dans le cahier de notes : la forme produite par les extrants n'est pas la forme décrite dans l'explication.)
4. **Classification naïve bayésienne polynôme**
   - Est-ce que la classification naïve bayésienne polynôme créée pour la matrice document-terme de formation suggère que la critique 9999_1.txt est positive ou négative? Est-ce que des mots que vous aviez relevés à la question 2 se trouvent dans cette critique?
   - Même question, mais pour la critique 9999_10.txt.

EXERCICE

5. Évaluation de la performance

- Décrivez la performance de l'analyseur de sentiments fournie par le rapport de classification.
- Pourquoi croyez-vous que la performance de VADER (l'analyseur de sentiments déjà formé de la NLTK) était inférieure à celle du modèle que vous avez formé au moyen des données sur les critiques?

RÉFÉRENCES

EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

CAHIERS DE NOTES

Text Processing, Text Visualization, Text Clustering, Sentiment Analysis Notebooks (en format HTML)
https://www.data-action-lab.com/wp-content/uploads/2019/03/TMNotebooks.zip

RÉFÉRENCES

BASU, T. (2017). *NPR's Fascinating Plan to Use A.I. on Trump's Tweets*, sur le site inverse.com. Consulté le 12 septembre 2017. https://www.inverse.com/article/30149-npr-planet-money-bot-botus-donald-trump

GOLDMARK, A. (2017). *Episode 763: BOTUS, Planet Money podcast*, sur le site Planet Monet de NPR.org. Consulté le 12 septembre 2017. http://www.npr.org/sections/money/2017/04/07/522897876/meet-botus-planet-money-s-stock-trading-twitter-bot

GREENSTONE, S. (2017). *When Trump Tweets, This Bot Makes Money*, sur le site NPR.org. Consulté le 12 septembre 2017. http://www.npr.org/2017/02/04/513469456/when-trump-tweets-this-bot-makes-money

METTLER, K. (2017). « *Trump and Dump »: When POTUS tweets and stocks fall, this animal charity benefits*, sur le site Washington Post. Consulté le 19 septembre 2017. https://www.washingtonpost.com/news/morning-mix/wp/2017/01/31/trump-and-dump-when-potus-tweets-and-stocks-fall-this-animal-charity-benefits/

JOCKERS, M.L. (2014). *Text Analysis with R for Students of Literature*, Springer.

ANASTASIA, D.C., TAGARELLI, A. et G. KARYPIS (2014). « Document Clustering: The Next Frontier » dans AGGARWAL, C.C. et C.K. REDDY, Eds. *Data Clustering: Algorithms and Applications*, CRC Press.

RÉFÉRENCES

AGGARWAL, C.C. et C.X. ZHAI (2015). « Text Classification », dans AGGARWAL, C.C., Ed. *Data Classification: Algorithms and Applications*, CRC Press.

SRIVASTAVA, A.N. et M. SAHAMI, Eds (2009). *Text Mining: Classification, Clustering, and Applications*, CRC Press.

SILGE, J. et D. ROBINSON (2017). *Text Mining with R: a Tidy Approach*, O'Reilly. https://www.tidytextmining.com/

JURAFSKY, D. et J.H. MARTIN (2009). *Speech and Language Processing*, 2e éd., Pearson.

AGGARWAL, C.C. et C.X. ZHAI, Eds (2012). *Mining Text Data*, Springer.

BIRD, S., KLEIN, E. et E. LOPER (2009). *Natural Language Processing with Python*, O'Reilly. http://www.nltk.org/book/

http://aiplaybook.a16z.com/docs/guides/nlp#user-content-apiexamples

https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

MATÉRIEL SUPPLÉMENTAIRE – CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

EXPLORATION DE TEXTE ET ANALYSE DE SENTIMENTS

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

La **classification naïve bayésienne polynôme** est un algorithme où on assume que les vecteurs de caractéristiques de chaque classe ont une distribution polynôme (application la mieux connue : **filtres antipourriels**).

Le jeu de données M est un nombre de courriels (les **enregistrements**)

Chaque enregistrement comporte n caractéristiques (les **fréquences** de n termes sélectionnés dans le corps du courriel)

Chaque enregistrement est représenté par un **vecteur de caractéristique** dénoté par :

$$\boldsymbol{x} = (x_1, \dots, x_n)$$

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

Partez du principe qu'il y a K catégories dans lesquelles un enregistrement peut être **classé**

- Étiquettes : pourriel, quarantaine, personnel, affaires, etc.

Laissez $\{C_k : k = 1, \dots, K\}$ désigner les catégories

Le problème de classification est pour déterminer

$$P(\boldsymbol{x} \in C_k | \, x_1, \dots, x_n) \text{ pour chaque } k$$

La prédiction est donnée par la classe pour laquelle cette valeur est la plus **élevée**

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

Corrigez k. À partir du **théorème de Bayes**, nous avons

$$P(\boldsymbol{x} \in C_k | x_1, \dots, x_n) \propto P(C_k) \times P(x_1, \dots, x_n | \boldsymbol{x} \in C_k)$$

L'hypothèse **naïve** est

$$P(x_1, \dots, x_n | \boldsymbol{x} \in C_k) = P(x_1 | \boldsymbol{x} \in C_k) \times \cdots \times P(x_n | \boldsymbol{x} \in C_k)$$

afin que

$$P(\boldsymbol{x} \in C_k | x_1, \dots, x_n) \propto P(C_k) \times \prod_{i=1}^{n} P(x_i | \boldsymbol{x} \in C_k)$$

L'hypothèque **polynôme** est

$$P(x_i | \boldsymbol{x} \in C_k) \propto p_{k,i}^{x_i}, \text{ où } p_{k,i} \in [0,1] \text{ pour chaque mot } i$$

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

En combinant ces hypothèses, les « probabilités » a posteriori sont

$$P(\boldsymbol{x} \in C_k | x_1, \dots, x_n) \propto P(C_k) \times \prod_{i=1}^{n} p_{k,i}^{x_i}$$

Le modèle peut être linéarisé en prenant des logarithmes

$$\log P(\boldsymbol{x} \in C_k | x_1, \dots, x_n) \propto b_k + \sum_{i=1}^{n} x_i \cdot \log p_{k,i}$$

La classification est **formée** en estimant les paramètres $p_{k,i}$ dans un sous-ensemble de tous les enregistrements et en spécifiant les « valeurs » b_k a priori.

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

Si un message comporte un jeton qui n'a jamais été vu auparavant, il est **impossible** de prédire son appartenance probable à une classe au moyen d'un comportement antérieur (non existant)

Pour éviter les divisions par 0, vous pouvez utiliser l'estimation corrigée

$$\hat{p}_{k,i} = \frac{\sum_{\boldsymbol{x} \in C_k} x_i + 1}{\sum_{\boldsymbol{x} \in C_k} \sum_{j=1}^{n} x_j + |v|} = \frac{\#w_i \in C_k + 1}{W_k + |v|}$$

$|v|$: taille du vocabulaire, $\#w_i \in C_k$: compte de w_i dans C_k, W_k : compte de tous les mots dans C_k

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

Jeu de données de formation

cl	**ID**	**texte**
+	i1	J'aime ce téléphone
+	i2	excellente qualité audio
+	i3	J'aime ce téléphone génial
-	i4	je le déteste
-	i5	mauvaise qualité
-	i6	vraiment mauvaise Je le déteste

Jeu de données d'essai

?? i7 déteste déteste DÉTESTE la qualité du téléphone

Jeu de données de formation traité

cl	**ID**	**texte**
+	i1	aime téléphone
+	i2	excellente qualité audio
+	i3	aime téléphone génial
-	i4	déteste
-	i5	mauvaise qualité
-	i6	mauvaise déteste

Jeu de données d'essai traité

?? i7 déteste déteste déteste qualité téléphone

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

$$P(+) = \frac{3}{6} = 0.5 \quad \text{et} \quad P(-) = \frac{3}{6} = 0.5 \quad \text{donc} \quad b_+ = b_- = \ln 0.5$$

$$|v| = 8, W_+ = 8, W_- = 5$$

$$\hat{p}_{+,\text{excellente}} = \frac{(\#\text{excellente} \in +) + 1}{W_+ + 8} = \frac{1+1}{8+8} = \frac{1}{8}$$

$$\hat{p}_{-,\text{excellente}} = \frac{(\#\text{excellente} \in -) + 1}{W_- + 8} = \frac{0+1}{5+8} = \frac{1}{13}$$

…

CLASSIFICATION NAÏVE BAYÉSIENNE POLYNÔME

$\hat{p}$	excellente	mauvaise	géniale	déteste	aime	téléphone	qualité	audio
+	0.1250	0.6025	0.1250	0.0625	0.1875	0.1875	0.1250	0.1250
−	0.0769	0.2308	0.0769	0.2308	0.0769	0.0769	0.1538	0.0769

Jeu de données d'essai

	excellente	mauvaise	géniale	déteste	aime	téléphone	qualité	audio
i7	0	0	0	3	0	1	1	0

$$P(+ \mid \boldsymbol{x}) \propto 2.9 \times 10^{-6}$$

$$P(- \mid \boldsymbol{x}) \propto 9.7 \times 10^{-6}$$