Data Challenge

Sentiment Analysis

Université Paris Dauphine-PSL

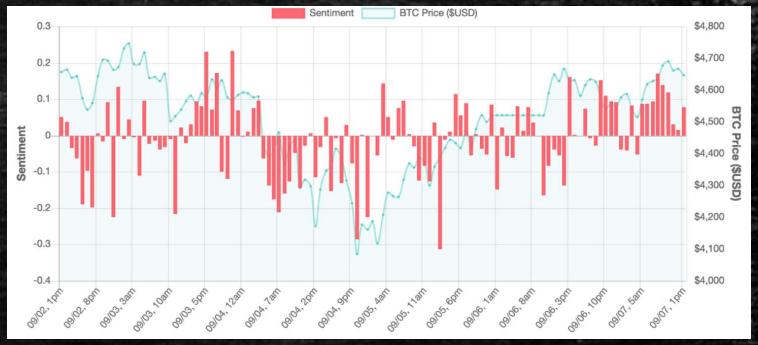
Sommaire

- Définition et objectif
- Mesure de sentiment
- Lexiques de sentiments ou émotions
- Graphiques de sentiments ou émotions
- Algorithme : Naive Bayes Classifier
- Applications numériques

Définition et objectif

- Définition
  - Sentiment Analysis = processus d'extraction de l'intention émotionnelle d'un auteur à partir d'un texte
- Objectif :
  - Avoir un avis objectif sur un livre, un film, un produit, l'avis d'une population sur une thématique…
  - Résultats sous différentes formes :
    - Elémentaire : avis positif ou négatif sur un livre, un film, un produit…. Ex : faut-il acheter tel produit ?

    - Plus complexe : échelle d'attitude de 1 à 5 Ex : avis sur un film, un séminaire….
    - Avancé : attitudes complexes Ex : avis d'internautes….



Superposition cours bitcoin et sentiment pour la monnaie

Mesure de sentiment

- Notion de polarité
  - Degré de sentiment positif et négatif
  - Métrique constitué à partir de mots, de groupe de mots d'une phrase
- Classification des mots :
  - Mots orientés positivement ou négativement (Polarized)

    joie colère
  - Mots neutres ➔ pas d'effet émotionnel x_i^0 Le, autour, chez, livre, film…
  - Mots inversants (negator) ➔ mots qui inversent l'orientation du sentiment x_i^N Pas gentil, pas mauvais…
  - Mots de valence ➔ mots qui intensifient ou désintensifient l'émotion

    x_i^a x_i^d Peu, guère…

    Très, beaucoup….

Mesure de sentiment

- Détermination de la polarité
  - À partir d'une phrase, d'un groupe de mots (context cluster)
  - Formule :

$$\delta = \frac{x_i^T}{\sqrt{n}}$$

Indice de polarité

(Hu & Liu, 2004)

Nombre de mots

Constante (0.8)

Nombre de mots inversants

$$x_i^T = \sum ((1 + c(x_i^A - x_i^D)) \cdot w(-1)^{\sum x_i^N})$$

$$x_i^A = \sum (w_{neg} \cdot x_i^a)$$

$$x_i^{D'} = \sum (-w_{neg} \cdot x_i^a + x_i^d)$$

$$w_{neg} = \left(\sum x_i^N\right) \bmod 2$$

Mot intensifiant et désintensifiant

NB : le calcul de la polarité (δ) dépend du dictionnaire de polarité des mots (lexique de sentiments)

Mesure de sentiment

- Détermination de la polarité (emojis)
  - À partir de tweets…

  - Classification des emoji en 3 groupes :

$$c \in \{-1, 0, +1\}$$

Négatif, neutre, positif

  - Polarité (score)

$$\overline{s} = -1 \cdot p_{-} + 0 \cdot p_0 + 1 \cdot p_{+}$$

Moyenne pondérée

$$-1 < \overline{s} < +1$$

Score de sentiment

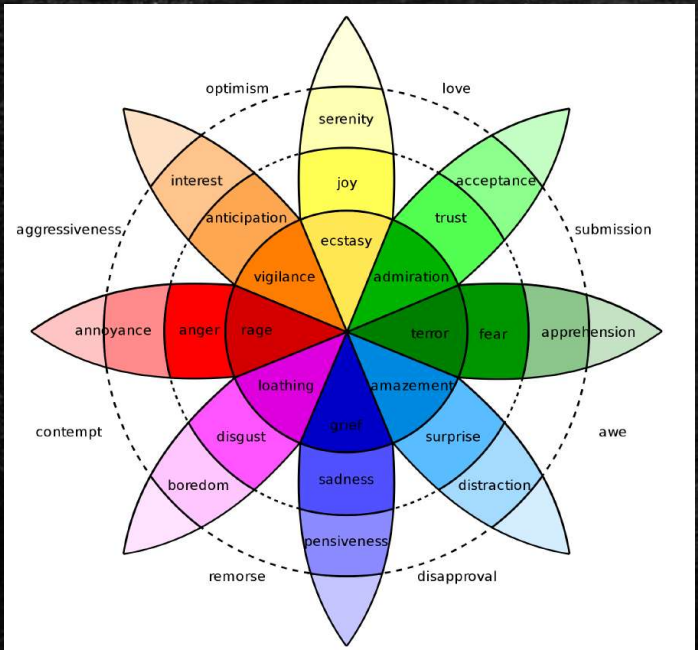
(Novak, Smailovic, Sluban, & Mozetic, 2015)

Probabilités discrètes (évaluation empirique)

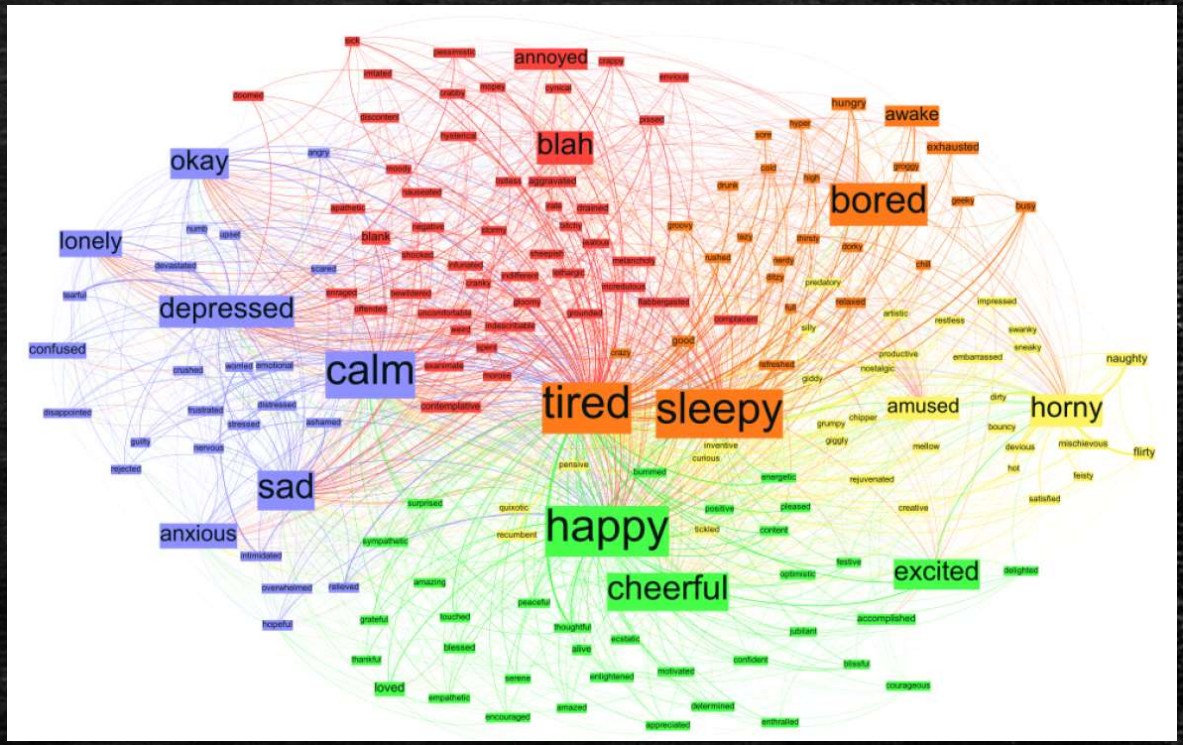
Lexiques de sentiments ou émotions

- Inventaire des sentiments (émotions)

Roue des émotions de Plutchik



Classification de Kanjoya



Lexiques de sentiments ou émotions

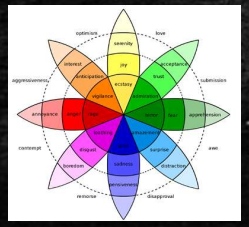
- Lexiques
  - AFINN (2477 mots)
    - Mots anglais classés sur une échelle d'entier compris entre -5 " negative " et 5 " positive"
  - BING (6 787 mots)
    - Mots anglais généraux classifiée en "positive" ou "negative"

```
> afinn[sample(1:nrow(afinn), 20),]
# A tibble: 20 x 2
   word          value
   <chr>         <dbl>
 1 immobilized      -1
 2 appreciation      2
 3 awful            -3
 4 inquisition      -2
 5 stimulate         1
 6 defiant          -1
 7 nervous          -2
 8 apocalyptic      -2
 9 disparaging      -2
10 appease           2
11 boastful         -2
12 blamed           -2
13 lurks            -1
14 accusing         -2
15 strengthening     2
16 breathtaking      5
17 hesitate         -2
18 distrust         -3
19 charged          -3
20 reassuring        2
>
```

```
> bing[sample(1:nrow(bing), 20),]
# A tibble: 20 x 2
   word          sentiment
   <chr>         <chr>
 1 transgress    negative
 2 erroneously   negative
 3 pleases       positive
 4 mistress      negative
 5 exalting      positive
 6 trapped       negative
 7 terribly      negative
 8 calumniate    negative
 9 vile          negative
10 simpler       positive
11 sags          negative
12 conscientious positive
13 furiously     negative
14 faze          negative
15 destructive   negative
16 manic         negative
17 inexperience  negative
18 matchless     positive
19 proper        positive
20 undependability negative
>
```

Lexiques de sentiments ou émotions

- Lexiques
  - Loughran-McDonald (4 150 mots)
    - Lexique de mots financiers classes en 6 états : "negative", "positive", "litigious", "uncertainty", "constraining " et "superfluous".
  - NRC (13 901 mots)
    - Classification des mots anglais standards en 10 sentiments ou émotions : "negative", "positive", "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise " et "trust"



```
> loughran[sample(1:nrow(loughran), 20),]
# A tibble: 20 x 2
   word          sentiment
   <chr>         <chr>
 1 imbalances    negative
 2 crime         litigious
 3 scrutinized   negative
 4 misstatements negative
 5 statutorily   litigious
 6 revoke        negative
 7 hinders       negative
 8 grievance     negative
 9 certiorari    litigious
10 litigations   negative
11 usurping      negative
12 advantaged    positive
13 preconditions constraining
14 unsure        negative
15 disclosed     negative
16 unencumbered  litigious
17 downgrade     negative
18 expropriate   negative
19 barred        negative
20 overestimated negative
>
```

```
> nrc[sample(1:nrow(nrc), 20),]
# A tibble: 20 x 2
   word          sentiment
   <chr>         <chr>
 1 miracle       trust
 2 inhuman       sadness
 3 comfort       joy
 4 rigorous      negative
 5 bang          disgust
 6 nectar        positive
 7 sufferer      sadness
 8 lash          anger
 9 sea           positive
10 louse         disgust
11 massacre      disgust
12 chastisement  negative
13 miracle       positive
14 frolic        positive
15 halfway       negative
16 drab          sadness
17 whirlwind     negative
18 bacterium     fear
19 hollow        sadness
20 bacterium     negative
>
```

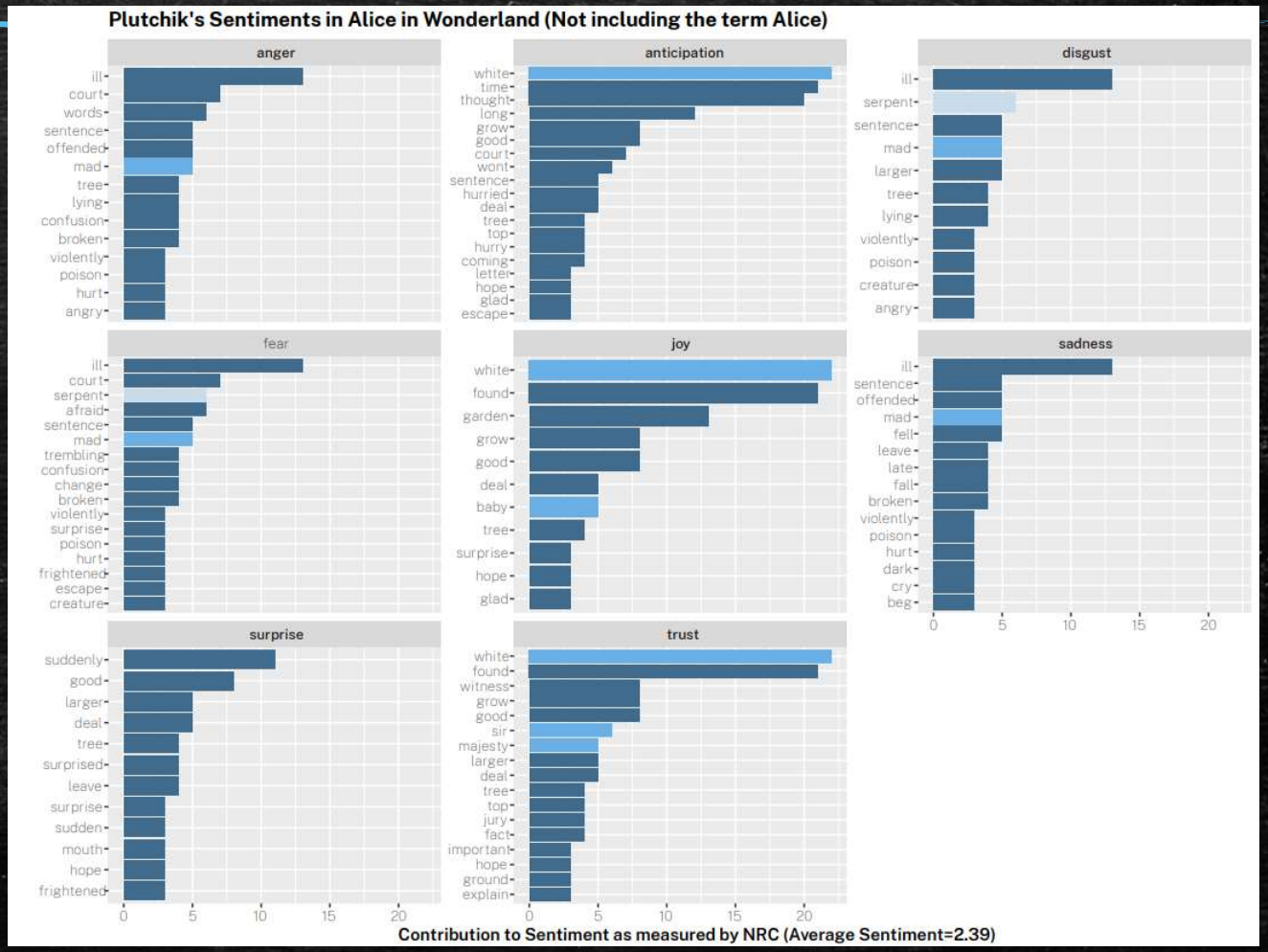
Lexiques de sentiments ou émotions

- Remarques :
  - Choix du lexique influe sur l'analyse
  - Intéressant de mixer l'analyse avec plusieurs lexiques
  - Possibilité de créer son propre lexique en fonction du domaine d'analyse (ex : aérien, environnement…)
  - Ressources logicielles :
    - Python : **pysentiment2**
    - R : package **lexicon**

Applications numériques

- Livre « Alice au pays des merveilles »

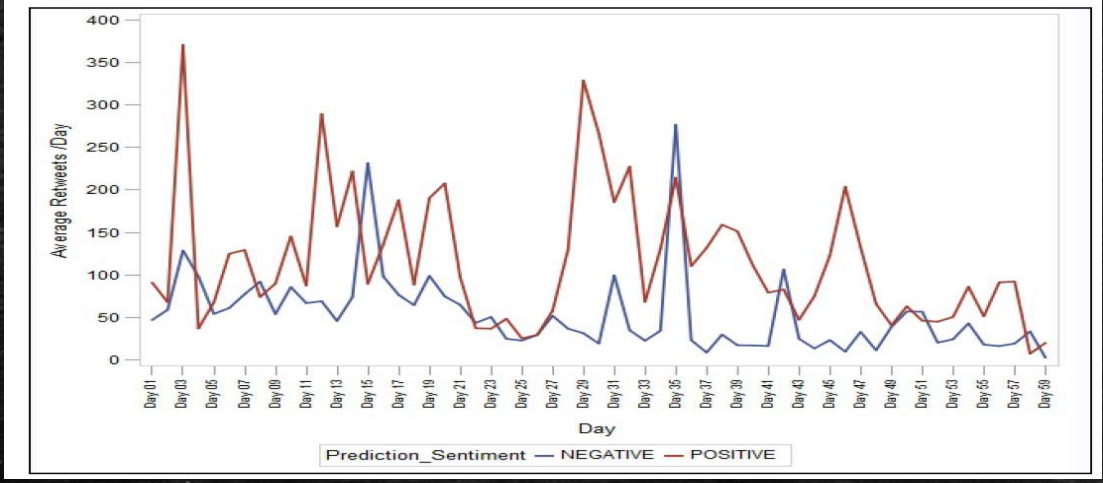
Lexicon : NRC



Text Analysis of Lewis Carroll's Alice in Wonderland, Brown & McSweeney, December 2019

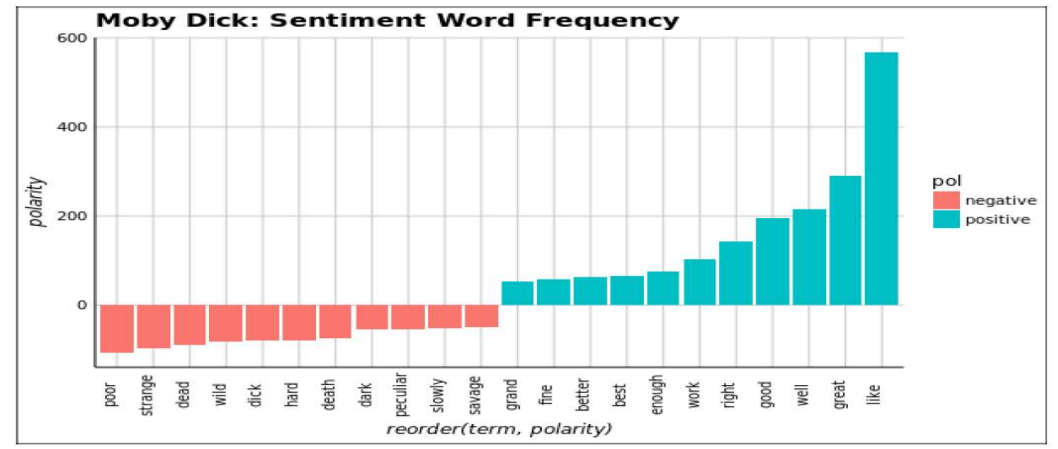
Graphiques de sentiments ou émotions

- Suivi temporel



Evolution journalière sentiment/émotion des tweets

- Statique

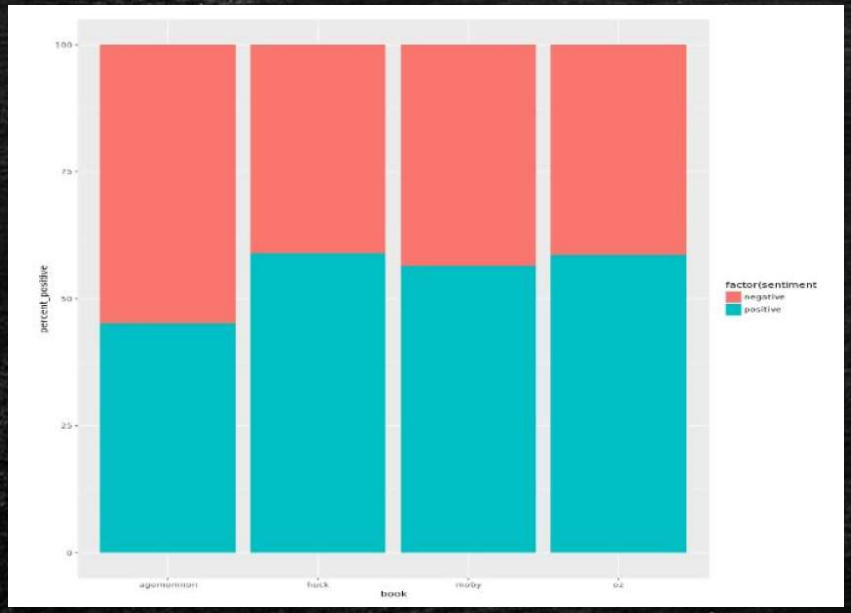


Synthèse sentiments/émotion pour un document (ici roman Moby Dick)

Graphiques de sentiments ou émotions

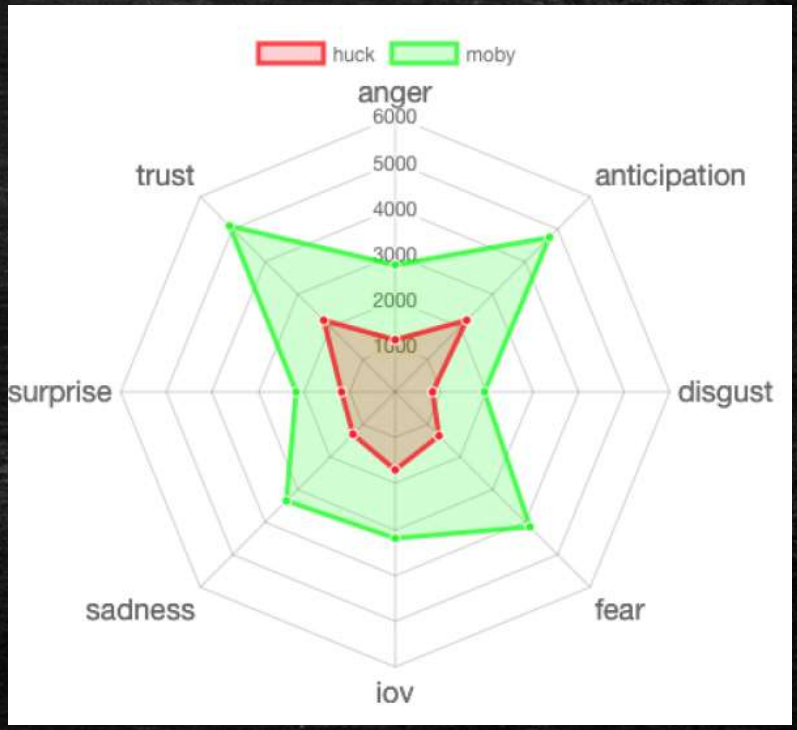
- Comparaison

Histogramme empilé



Analyse synthétique entre différents documents (ici des romans)

Radar

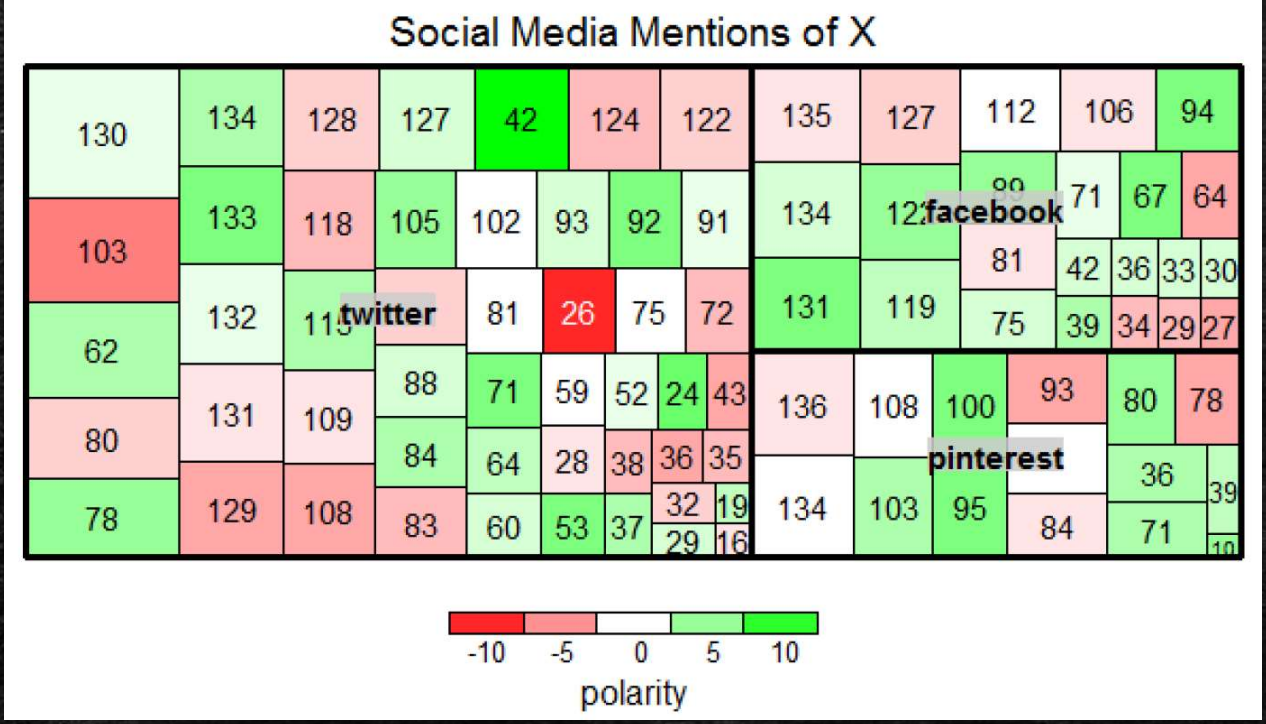


Analyse détaillée en fonction des sentiments

Graphiques de sentiments ou émotions

- Comparaison

Treemap



Chaque case représente un article (post) dont la taille est liée à une statistique

La couleur de chaque case dépend de l'échelle de sentiment (rouge : « negative », vert « positive »)

Algorithme : Naive Bayes Classifier

- Estimation de $P(C_{k,})$
  - Nombre d'occurrences d'observations appartenant à la classe C_k par rapport au nombre total d'observations toutes classes confondues

Effectif de la classe k

$$P(C_k) = \frac{n_k}{N}$$

Effectif total

- Estimation de $P(x \mid C_{k,})$ (x = mot)
  - Probabilité conditionnelle (règle de Laplace)

Fréquence du mot w attribuée à la classe k

$$P(C_k) = \frac{\#(w,k) + 1}{\#k + |V|}$$

Estimateur de Laplace

Nombre de mots de la classe k

Nombre total de mots

Algorithme : Naive Bayes Classifier

- Estimation de $P(x \mid C_{k,})$
  - Calcul compliqué
  - Hypothèse : indépendance des composantes des caractéristiques de X ($\in R^d$) i.e. indépendance des variables Xi et Xj d'où le terme « Naïve »
  - Simplification du calcul de

$$P(x \mid C_{k,})$$

$$P(x \mid C_{k,}) = \mathrm{P}(x_{[1]} \mid C_{k,})\mathrm{P}(x_{[2]} \mid C_{k,}) * \cdots * \mathrm{P}(x_{[d]} \mid C_{k,})$$

- Implémentation informatique
  - R : `e1071`
  - Python : sklearn.naive_bayes

Remarque :

Transformation des calculs en logarithmes dans les logiciels pour des raisons de mise en œuvre (dimension élevée (d) de l'espace des caractéristiques, probabilité faible (débordement de capacité))

Algorithme : Naive Bayes Classifier

- Exemple :
  - Question : est ce que je dois acheter le dernier téléphone de la marque XYZ ?
  - Données :
    - Collecte d'avis sur internet

Collecte avis sur 5 sites internet

DOC	WORDS	CLASS
1	DONT BUY	NEGATIVE
2	PHONE GOT HANGED	NEGATIVE
3	BATTERY DRAINS FAST	NEGATIVE
4	DURABLE PHONE	POSITIVE
5	GREAT CAMERA	POSITIVE
6	GREAT PHONE BUY IT	?

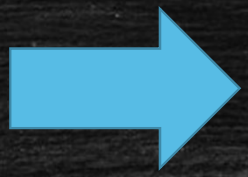
Ma question :

achat ou non achat ?

Algorithme : Naive Bayes Classifier

- Exemple :
  - Tableau de contingence des mots

DOC	WORDS	CLASS
1	DONT BUY	NEGATIVE
2	PHONE GOT HANGED	NEGATIVE
3	BATTERY DRAINS FAST	NEGATIVE
4	DURABLE PHONE	POSITIVE
5	GREAT CAMERA	POSITIVE



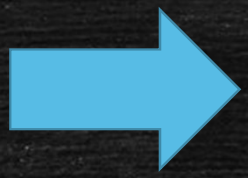
Conversion des avis dans une table de contingence

WORDS	POSITIVE	NEGATIVE
DONT	0	1
BUY	0	1
PHONE	1	1
GOT	0	1
HANGED	0	1
BATTERY	0	1
DRAINS	0	1
FAST	0	1
DURABLE	1	0
GREAT	1	0
CAMERA	1	0

Algorithme : Naive Bayes Classifier

- Exemple :
  - Calcul des probabilités a priori

DOC	WORDS	CLASS
1	DONT BUY	NEGATIVE
2	PHONE GOT HANGED	NEGATIVE
3	BATTERY DRAINS FAST	NEGATIVE
4	DURABLE PHONE	POSITIVE
5	GREAT CAMERA	POSITIVE



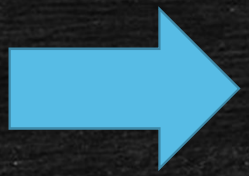
Probabilités a priori

Classe	Effectif classe	Proba a priori	
POSITIVE	2	**0.4**	= 2 /5
NEGATIVE	3	**0.6**	= 3 /5

Algorithme : Naive Bayes Classifier

- Exemple :
  - Calcul des probabilités conditionnelles

WORDS	POSITIVE	NEGATIVE
DONT	0	1
BUY	0	1
PHONE	1	1
GOT	0	1
HANGED	0	1
BATTERY	0	1
DRAINS	0	1
FAST	0	1
DURABLE	1	0
GREAT	1	0
CAMERA	1	0



= (0+1) /(4+11)

WORD	Eff POSITIVE	Eff NEGATIVE	P(WORD / POSITIVE)	P(WORD / NEGATIVE)
DONT	0	1	**0.066666667**	**0.105263158**
BUY	0	1	**0.066666667**	**0.105263158**
PHONE	1	1	**0.133333333**	**0.105263158**
GOT	0	1	**0.066666667**	**0.105263158**
HANGED	0	1	**0.066666667**	**0.105263158**
BATTERY	0	1	**0.066666667**	**0.105263158**
DRAINS	0	1	**0.066666667**	**0.105263158**
FAST	0	1	**0.066666667**	**0.105263158**
DURABLE	1	0	**0.133333333**	**0.052631579**
GREAT	1	0	**0.133333333**	**0.052631579**
CAMERA	1	0	**0.133333333**	**0.052631579**
Total	**4**	**8**		

= (1+1) /(8+11)

Algorithme : Naive Bayes Classifier

- Exemple :
  - Calcul des probabilités a posteriori et décision

Probabilités conditionnelles

Classe	a priori	GREAT	PHONE	BUY	Posteriori (num)	Posteriori
POSITIVE	0.4	0.133333333	0.13333333	0.06666667	0.000474074	0.575347062
NEGATIVE	0.6	0.052631579	0.10526316	0.10526316	0.000349905	0.424652938

  - Décision : critère Maximum A Posteriori (MAP)
    - positive

Classe	Posteriori (num)	Posteriori
POSITIVE	0.000474074	0.575347062
NEGATIVE	0.000349905	0.424652938

Proba(POSITIVE/ mots) > Proba(NEGATIVE/ mots)

Algorithme : Naive Bayes Classifier

- Exemple internet :
  - **Sentiment Analysis of a Tweet With Naive Bayes**
  - https://towardsdatascience.com/sentiment-analysis-of-a-tweet-with-naive-bayes-ff9bdb2949c7