Pré-projet – Data Mining

La composition de notre projet de Data Mining est la suivante : Daniel **KHARSA**, Mohamed-Amine **BOUREGA**, Samir **LAZZALI**, Thierry **GAMEIRO MARTINS** et Romain **KRAWCZYK**.

1 Choix de la base de données

Pour notre projet de Data Mining, nous avons choisi d'utiliser la base « Online Popularity News », disponible en libre accès ici. La base de données contient un ensemble d'informations relatives aux articles publiés sur le site Mashable entre janvier 2013 et janvier 2015.

La base contient 39644 observations et 61 colonnes. Voici un extrait du jeu de données (des 12 premières variables et des 6 dernières) :

	url	timedelta	n_tokens_title	n_tokens_content	n_unique_tokens	n_non_stop_words
0	http ://mashable.com/2013/01/07/amazon-instant-...	731.0	12.0	219.0	0.663594	1.0
1	http ://mashable.com/2013/01/07/ap-samsung-spon...	731.0	9.0	255.0	0.604743	1.0
2	http ://mashable.com/2013/01/07/apple-40-billio...	731.0	9.0	211.0	0.575130	1.0
3	http ://mashable.com/2013/01/07/astronaut-notre...	731.0	9.0	531.0	0.503788	1.0
4	http ://mashable.com/2013/01/07/att-u-verse-apps/	731.0	13.0	1072.0	0.415646	1.0

	n_non_stop_unique_tokens	num_hrefs	num_self_hrefs	num_imgs	num_videos	average_token_length
0	0.815385	4.0	2.0	1.0	0.0	4.680365
1	0.791946	3.0	1.0	1.0	0.0	4.913725
2	0.663866	3.0	1.0	1.0	0.0	4.393365
3	0.665635	9.0	0.0	1.0	0.0	4.404896
4	0.540890	19.0	19.0	20.0	0.0	4.682836

	max_negative_polarity	title_subjectivity	title_sentiment_polarity	abs_title_subjectivity	abs_title_sentiment_polarity	shares
0	-0.200000	0.500000	-0.187500	0.000000	0.187500	593
1	-0.100000	0.000000	0.000000	0.500000	0.000000	711
2	-0.133333	0.000000	0.000000	0.500000	0.000000	1500
3	-0.166667	0.000000	0.000000	0.500000	0.000000	1200
4	-0.050000	0.454545	0.136364	0.045455	0.136364	505

D'après une première observation de notre jeu de données, on peut constater que toutes les variables sont de type numérique, à l'exception de la première variable (url). Toutefois, on verra par la suite qu'il existe des variables catégorielles déjà binarisés.

2 Audit du jeu de données

2.1 Compréhension des variables

Notre base de données a été construite dans le cadre de travaux cherchant à prédire la popularité d'un article à partir de ses caractéristiques. Les différentes variables contenues dans notre base ont été créées en utilisant des méthodes d'analyse naturelle du langage (NLP). Avant toutes choses, il est donc important de nous assurer de la bonne compréhension de chacune des variables avant de réaliser un audit plus poussé. La liste des variables et leur signification est donnée ci-dessous :

- **url** : URL of the article
- **timedelta** : Days between the article publication and the dataset acquisition
- **n_tokens_title** : Number of words in the title
- **n_tokens_content** : Number of words in the content
- **n_unique_tokens** : Rate of unique words in the content
- **n_non_stop_words** : Rate of non-stop words in the content
- **n_non_stop_unique_tokens** : Rate of unique non-stop words in the content
- **num_hrefs** : Number of links
- **num_self_hrefs** : Number of links to other articles published by Mashable
- **num_imgs** : Number of images
- **num_videos** : Number of videos
- **average_token_length** : Average length of the words in the content
- **num_keywords** : Number of keywords in the metadata
- **data_channel_is_lifestyle** : Is data channel 'Lifestyle' ?
- **data_channel_is_entertainment** : Is data channel 'Entertainment' ?
- **data_channel_is_bus** : Is data channel 'Business' ?
- **data_channel_is_socmed** : Is data channel 'Social Media' ?
- **data_channel_is_tech** : Is data channel 'Tech' ?
- **data_channel_is_world** : Is data channel 'World' ?
- **kw_min_min** : Worst keyword (min. shares)
- **kw_max_min** : Worst keyword (max. shares)
- **kw_avg_min** : Worst keyword (avg. shares)
- **kw_min_max** : Best keyword (min. shares)
- **kw_max_max** : Best keyword (max. shares)

— **kw_avg_max** : Best keyword (avg. shares)
— **kw_min_avg** : Avg. keyword (min. shares)
— **kw_max_avg** : Avg. keyword (max. shares)
— **kw_avg_avg** : Avg. keyword (avg. shares)
— **self_reference_min_shares** : Min. shares of referenced articles in Mashable
— **self_reference_max_shares** : Max. shares of referenced articles in Mashable
— **self_reference_avg_sharess** : Avg. shares of referenced articles in Mashable
— **weekday_is_monday** : Was the article published on a Monday ?
— **weekday_is_tuesday** : Was the article published on a Tuesday ?
— **weekday_is_wednesday** : Was the article published on a Wednesday ?
— **weekday_is_thursday** : Was the article published on a Thursday ?
— **weekday_is_friday** : Was the article published on a Friday ?
— **weekday_is_saturday** : Was the article published on a Saturday ?
— **weekday_is_sunday** : Was the article published on a Sunday ?
— **is_weekend** : Was the article published on the weekend ?
— **LDA_00** : Closeness to LDA topic 0
— **LDA_01** : Closeness to LDA topic 1
— **LDA_02** : Closeness to LDA topic 2
— **LDA_03** : Closeness to LDA topic 3
— **LDA_04** : Closeness to LDA topic 4
— **global_subjectivity** : Text subjectivity
— **global_sentiment_polarity** : Text sentiment polarity
— **global_rate_positive_words** : Rate of positive words in the content
— **global_rate_negative_words** : Rate of negative words in the content
— **rate_positive_words** : Rate of positive words among non-neutral tokens
— **rate_negative_words** : Rate of negative words among non-neutral tokens
— **avg_positive_polarity** : Avg. polarity of positive words
— **min_positive_polarity** : Min. polarity of positive words
— **max_positive_polarity** : Max. polarity of positive words
— **avg_negative_polarity** : Avg. polarity of negative words
— **min_negative_polarity** : Min. polarity of negative words
— **max_negative_polarity** : Max. polarity of negative words
— **title_subjectivity** : Title subjectivity
— **title_sentiment_polarity** : Title polarity
— **abs_title_subjectivity** : Absolute subjectivity level
— **abs_title_sentiment_polarity** : Absolute polarity level
— **shares** : Number of shares

La variable 60 (la dernière variable) correspond au nombre de partages sur les réseaux sociaux (Facebook, Twitter, Google+, LinkedIn, Stumble-Upon et Pinterest), c'est la variable que nous essaierons de prédire (nous la noterons parfois Y dans ce qui suit). Les deux premières variables correspondent respectivement à l'URL de l'article, et au nombre de jours écoulés entre la publication de l'article et la création de la base de données. Ces deux variables ne sont pas supposées nous apporter de l'information utile à la prédiction de notre variable cible. Par conséquent, nous pensons les supprimer lors du passage à l'analyse prédictive.

Sachant que les autres variables ont été construites en utilisant des méthodes d'analyse naturelle du langage, un certain nombre de remarques préalables doivent être énoncées :

— les stop words correspondent à ce que l'on appelle des mots vides, c'est-à-dire des mots tellement communs que leur prise en compte n'est pas requise (les mots tels que « la » , « de », « un » en sont des exemples). Les variables de notre base de données sont exemptés des stopwords.

— Les créateurs de la base ont distingués sept catégories auxquelles les articles peuvent appartenir, ces catégories sont désignées par le mot channel (business, tech, lifestyle etc.) et sont représentées par les variables commençant par *data_channel_*.

— un algorithme LDA (Latent Dirichlet Allocation) a été utilisé sur l'ensemble des articles afin de mettre en évidence cinq sujets récurrents et mesurer la proximité des chaque article à ces cinq thèmes. Les variables correspondantes commencent par *LDA_*.

— la Polarité fait référence au degré de négativité ou de positivité d'un énoncé, tandis que la subjectivité fait référence au degré de jugement, d'émotion ou d'implication personnelle de l'auteur. Ces notions sont capturées par les variables se terminant par *_polarity* et *_subjectivity*.

— La description donnée par les auteurs pour les autres variables est claire.

2.2 Éléments simples de statistiques descriptives, données manquantes et aberrantes

La première information que nous retirons est que notre base de données ne contient pas de valeurs manquantes.

Var.	None	Var.	None	Var.	None
url	0	kw_max_min	0	LDA_01	0
timedelta	0	kw_avg_min	0	LDA_02	0
n_tokens_title	0	kw_min_max	0	LDA_03	0
n_tokens_content	0	kw_max_max	0	LDA_04	0
n_unique_tokens	0	kw_avg_max	0	global_subjectivity	0
n_non_stop_words	0	kw_min_avg	0	global_sentiment_polarity	0
n_non_stop_unique_tokens	0	kw_max_avg	0	global_rate_positive_words	0
num_hrefs	0	kw_avg_avg	0	global_rate_negative_words	0
num_self_hrefs	0	self_reference_min_shares	0	rate_positive_words	0
num_imgs	0	self_reference_max_shares	0	rate_negative_words	0
num_videos	0	self_reference_avg_sharess	0	avg_positive_polarity	0
average_token_length	0	weekday_is_monday	0	min_positive_polarity	0
num_keywords	0	weekday_is_tuesday	0	max_positive_polarity	0
data_channel_is_lifestyle	0	weekday_is_wednesday	0	avg_negative_polarity	0
data_channel_is_entertainment	0	weekday_is_thursday	0	min_negative_polarity	0
data_channel_is_bus	0	weekday_is_friday	0	max_negative_polarity	0
data_channel_is_socmed	0	weekday_is_saturday	0	title_subjectivity	0
data_channel_is_tech	0	weekday_is_sunday	0	title_sentiment_polarity	0
data_channel_is_world	0	is_weekend	0	abs_title_subjectivity	0
kw_min_min	0	LDA_00	0	abs_title_sentiment_polarity	0

Concernant notre variable cible *share*, il n'y a également pas de valeur manquantes. Il est possible d'observer quelques statistiques descriptives de certaines variables, afin de vérifier si des données sont aberrantes :

	min_negative_polarity	max_negative_polarity	title_subjectivity	title_sentiment_polarity	abs_title_subjectivity	abs_title_sentiment_polarity	shares
count	39644	39644	39644	39644	39644	39644	39644
mean	-0.521944	-0.1075	0.282353	0.0714254	0.341843	0.156064	3395.38
std	0.29029	0.095373	0.324247	0.26545	0.188791	0.226294	11627
min	-1	-1	0	-1	0	0	1
25%	-0.7	-0.125	0	0	0.166667	0	946
50%	-0.5	-0.1	0.15	0	0.5	0	1400
75%	-0.3	-0.05	0.5	0.15	0.5	0.25	2800
max	0	0	1	1	0.5	1	843300

Nous remarquons que toutes les variables relatives à la polarité sont bien comprises en -1 et 1 tandis que celles relatives à la subjectivité sont comprises en 0 et 1. Il en est de même pour les variables correspondant à des ratios (non affiché ici). On peut donc en conclure que la base de ne contient pas de valeurs aberrantes au sens des variables qualitatives.

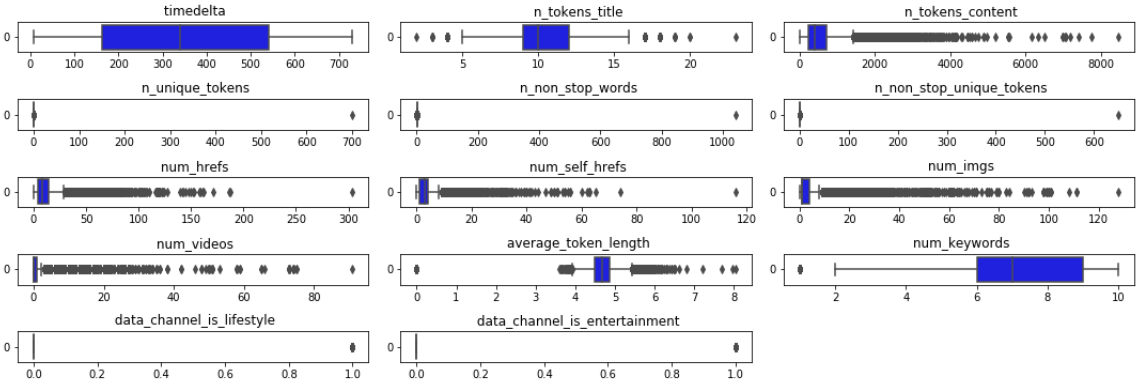

FIGURE 1 Boxplot – 1

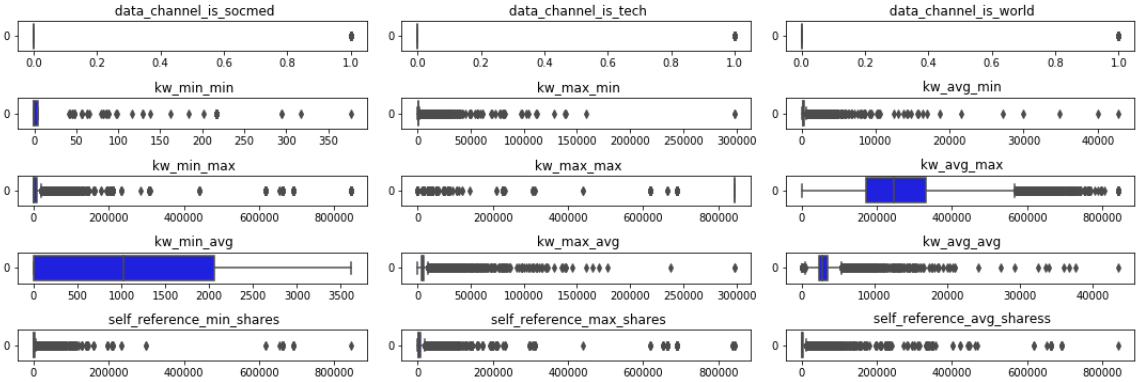

FIGURE 2 Boxplot – 2

Toutefois, à partir des *boxplot*, on peut constater la présence de plusieurs variables quantitatives contenant des valeurs extrêmes (les points affichés en dehors des rectangles bleus). Ceci pourra entraîner des erreurs de prédiction par la suite, mais ces valeurs pourront aussi révéler des observations intéressantes.

Il est également possible d'afficher une simple matrice des corrélations entre l'ensemble des variables du jeu de données :

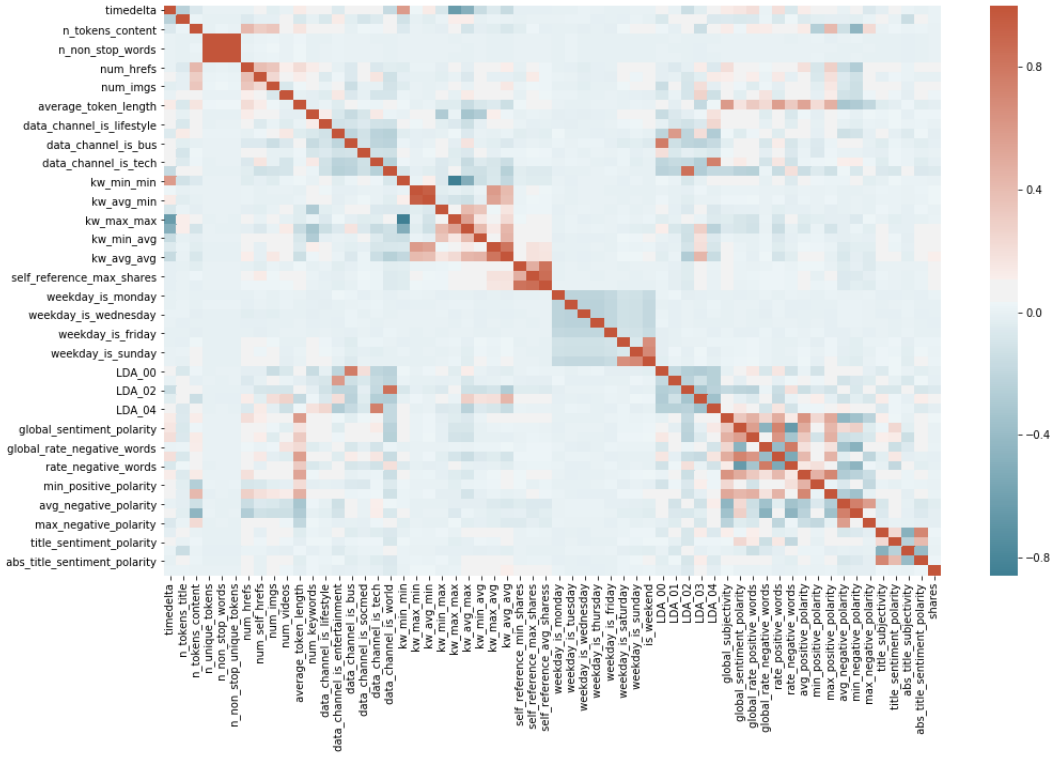

FIGURE 3 Corrélations

De manière globale, on peut observer que certaines variables sont très corrélées entre elles (par exemple les premières variables à propos des mots de l'article en question, des statistiques à propos des mots clés, ou encore à propos des degrés de négativité et de positivité des articles). On peut en déduire une forte présence de multicolinéarité dans notre jeu de donnés pour plusieurs blocs de variables.

Enfin, on peut terminer cette analyse par l'observation de la distribution de notre variable cible :

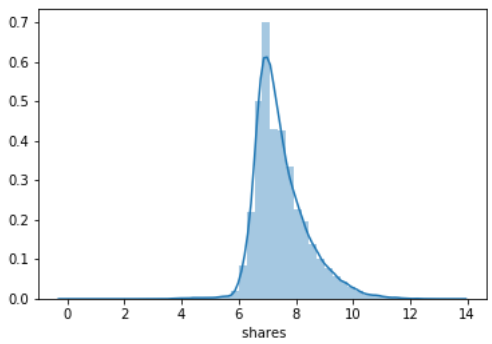

FIGURE 4 Distribution – *log_share*

Nous constatons que pour avoir une distributions plutôt « normale ». Pour cela, il a fallu

transformer notre variable cible en logarithme. Il faudra donc réfléchir par la suite si nous conservons cette forme des valeurs lors de nos prédictions.

2.3 Ecueils potentiels

D'après nous, le principal écueil est lié au nombre des variables explicatives à disposition. L'inclusion de toutes les variables lors de la modélisation risque de mener à du sur-apprentissage. Une phase de réduction de la dimensionnalité nous semble donc nécessaire, d'autant plus que certaines variables sont très corrélées entre elles. Deux méthodes pourront être utilisées afin d'opérer cette réduction en amont :

— **Etudes des corrélations avec la variable cible (test de Student, V de Cramer, test de Kruskal-Wallis etc.)** : sélection des variables les plus corrélées avec la cible, élimination des variables trop corrélées entre elles. Cette méthode reste tout de même arbitraire, mais permet potentiellement d'augmenter le pouvoir explicatif de notre modèle final.

— **A l'aide de méthodes telles que l'ACP ou l'algortihme t-SNE** : afin de ne pas nuire à l'interprétaiblité des variables utilisées et donc du modèle final, la phase préalable de réduction de la dimensionnalité pourrait être menée sur un ensemble de variables liées (par exemple les variables relatives à la subjectivité d'un article). On pourra observer des groupes de variables présentant la même information.
— Une alternative consisterait à utiliser des méthodes robustes à la multicolinéarité et à un nombre important de prédicteurs (LASSO ou RandomForests par exemple). Toutefois, cette dernière option nous forcerait à nous limiter à un nombre restreint de modèles prédictifs.

Enfin, nous prévoyons d'intégrer une partie destinée à l'interprétabilité des sorties de notre modèle final. Pour cela, il serait préférable de discrétiser la variable cible en deux classes : « populaire » et « non populaire ». En plus de faciliter la lecture des résultats, la discrétisation constitue également une réponse idoine au problème des valeurs extrêmes. Nous aurons donc à déterminer la méthode de discrétisation optimale (KMean, CAH, par l'étendue ou les quantiles par exemple).

3 Problématique

En plus de notre objectif de prédiction (la popularité d'un article), la richesse de notre base de données nous permet également de déterminer quels types d'articles sont les plus populaires : les articles négatifs ou positifs ? Les articles subjectifs ou objectifs ? Est-ce qu'un thème particulier est plus populaire qu'un autre ? Ces dernières questions sont

davantage des analyses descriptives.

A propos de notre analyse explicative, celle ci consiste donc à un problème de régression : tenter de prédire le nombre de partages d'un nouvel article. Toutefois, grâce à la possibilité d'une discrétisation de notre variable cible, cette problématique pourrait tendre à déterminer si un nouvel article peut être très populaire, ou non (variable binaire).