L’étude du contexte des émissions englobant les écrans publicitaires ;

Optimisation de la performance d’achat des écrans publicitaire

Mémoire préparé et présenté par Fatemeh Sajadi Ansari

Directeur de recherche :

Monsieur Damien Nouvel

Sommaire

Résumé :

Au cours du projet de ce mémoire, nous allons étudier l'impact des écrans publicitaires sur le trafic des sites e-commerces. À ces écrans sont associés certains types d'information, comme les émissions avant et après le passage de l'écran publicitaire. Chacune de ces informations peut impacter la performance de ces écrans.

Dans un premier temps, le but de ce mémoire est de présenter un modèle prédictif de la performance des écrans publicitaires par le biais des émissions, avant ou après le passage de ceux-ci. Dans un second temps, nous nous pencherons sur la description de ces émissions pour mettre l'accent sur l'analyse des mots utilisés.

Enfin, nous verrons s'il existe un lien établi entre le contexte de l'émission et la réussite de ces écrans.

Mot clés :

Campagne publicitaire, Ecran publicitaire, spot télé, DRTV, réseaux de Neurones, Perceptron, textométrie

1. Introduction

Aujourd'hui dans le monde commercial et plus particulièrement, dans le domaine de l'E-Commerce, on « monitore » de très près et parfois d'une manière très précise le trafic de son site. L'analyse de ces données est primordiale dans les décisions de marketing et business planning.

De nos jours, il existe un nombre très important d'outils qui permettent aux entreprises d'effectuer ce type d'analyse. Par contre quand il est question d'analyser les résultats d'une campagne publicitaire, la tache se complique surtout quand on parle de la DRTV[1].

Cette complexité vient du fait que pour une étude précise, une récolte immense de données est nécessaire[2]. Ces données (big data) doivent être accessibles rapidement et sous le format attendu.

Après la récolte et l'accessibilité de ces données, le travail des analystes commence. Ces études vont permettre aux entreprises de vérifier si leurs campagnes publicitaires[3] ont été efficaces ou non.

Or, l'objectif de ce mémoire de recherche est de présenter le processus d'analyse des données, afin d'en déduire les résultats d'une campagne publicitaire et les solutions pour optimiser ces résultats.

Tout en se basant sur les analyses existantes, nous proposerons d'autres études numériques et textométriques pour optimiser la performance de ces campagnes.

Pour les analyses numériques, les données sont modélisées avec les réseaux neuronaux, vu la performance de ceux-ci. Entre ces différents réseaux, le choix a été porté sur le *Perceptron mono couche,* qui est un modèle simple à paramétrer.

Pour cette modélisation, il a fallu une étude de « feature engineering[4] », dû au fait qu'il existe un certain nombre de facteurs qui impactent (directement ou non) les résultats des études. L'analyse d'impact d'un facteur nécessite d'enlever l'influence de tout autre élément du système.

Cette recherche comprend deux parties majeures contenant chacune un corpus propre. Les données des émissions associées à chacun des écrans publicitaires sont étudiées sous deux angles différents. Premièrement, les émissions sont présentées comme des ensembles numériques. Deuxièmement, les titres et la description de ces émissions ont été considérés pour une étude contextuelle (plus textuelle que numérique).

[1] « DRTV » est la publicité télévisée (dans son ensemble) qui invite les téléspectateurs à répondre directement à l'entreprise - généralement soit en appelant un numéro 800 soit en visitant un site Web
[2] Le sujet est discuté en détail dans le chapitre des données numériques
[3] Désormais quand nous parlerons de la campagne publicitaire cela veut dire la campagne télé
[4] « feature engineering » n'est pas un terme défini officiellement. Il peut être défini comme l'ensemble des taches liées à la conception de fonctionnalités, largement considérées comme l'une des étapes cruciales de chaque application de Machine learning.

Nous nous poserons les questions suivantes: est ce qu'il y a un lien entre le contexte de l'écran et la probabilité de réussite de cet écran ; que faire pour rendre un spot plus performant contextuellement ?

2. Présentation et les enjeux du groupe MyMedia

2.1 Le groupe MyMedia

Afin de mieux comprendre les activités auxquelles j'ai participé et le contexte dans lequel j'ai baigné durant ces six mois, je commencerai par présenter le groupe MyMedia à travers ses différentes activités et son évolution. Je présenterai ensuite l'agence MyMedia, plus particulièrement le département Business Innovation dans lequel j'ai effectué ce stage.

Le groupe MyMedia est le premier groupe de communication indépendant en France, fondé en 2005 par 2 anciens annonceurs. Le digital ainsi que les nouveaux médias et technologies bousculent la relation aux consommateurs et les opportunités medias. Ils sont actuellement au cœur de la réflexion marketing. C'est pour cette raison que MyMedia a développé une expertise unique en E-commerce et en stratégie digitale pour accompagner les entreprises face aux défis d'internet. MyMedia est ainsi une véritable agence média interactive.

Les activités du groupe :

Le groupe MyMedia est constitué de cinq agences :

– MyMedia (média)

– Géovista (média de proximité)

– Eggs (création)

– Stratco

– Search Foresight.

2.2 GéoVista

GéoVista est la branche du groupe MyMedia spécialisée dans le conseil et la mise en œuvre des plans médias à destination de ses clients, en faisant converger une double approche de la proximité : la géolocalisation et l'affinité avec une cible donnée. Le personnel de l'agence est composé d'experts du média et de la géolocalisation. GéoVista s'attache à développer des solutions de médiaplanning pour les annonceurs recherchant une visibilité supérieure sur leur territoire de communication.

L'équipe de GéoVista répond à deux problématiques différentes dans la recherche de proximité :

- Pour les réseaux de points de vente qui désirent émerger sur leur zone de chalandise (zone géographique d'influence, d'où provient la majorité de la clientèle) : obtenir un meilleur ciblage territorial.

- Pour les annonceurs institutionnels : grâce à des analyses socio-comportementales, adapter les solutions pour permettre à la cible d'être exposée de la façon la plus efficace au message souhaité.

2.3 Eggs

Eggs est l'agence de publicité du groupe MyMedia, elle associe professionnels de la publicité et experts du digital.

Cette configuration lui permet de proposer aux annonceurs des dispositifs créatifs, stratégiques et innovants autour d'un "one stop" shopping des métiers de la communication: création publicitaire, buzz marketing, site internet, événementiel. Eggs compte aujourd'hui 15 collaborateurs et intervient entre autres pour : Orange, Areva, Le coq sportif, Sofinco, France 4, Mtv, LaPoste, Warner, Nexity, NRJ mobile, Primagaz...

Grâce à la culture et l'expérience des équipes, Eggs réconcilie "copystrategy" et " création digitale". L'agence propose à ses clients le positionnement, un plan de communication, la création publicitaire print, audiovisuelle et digitale, la création de sites, le buzz marketing et l'évènementiel.

2.4 Search Foresight

Search Foresight accompagne ses clients dans le développement du trafic naturel de leur site internet. Ses compétences couvrent tout le spectre de "l'inbound marketing" (stratégie marketing visant à faire venir le client vers soi plutôt que d'aller le chercher) : SEO mais aussi, l'optimisation pour les réseaux sociaux, le référencement local, le référencement des sites et applications mobiles, l'e-réputation... L'équipe, composée de consultants expérimentés, est particulièrement rompue à l'accompagnement stratégique de marques ou de grands comptes confrontés à des problématiques complexes : de nombreuses versions pays et /ou versions linguistiques, des sites à forte volumétrie d'audience ou de pages, des plateformes techniques complexes.

Les consultants disposent d'outils et de solutions permettant de faciliter l'intégration technique des recommandations sur tout type de plateforme, de framework, ou de CMS. Search Foresight dispose d'une expertise dans l'exploitation des technologies liées au traitement automatisé du langage et de la linguistique informatisée à des fins de création d'audience.

L'agence propose plusieurs services notamment :

– SEO / Référencement naturel

– Audits approfondis

– Crawl et analyse de logs

– Stratégie SEA (avec les équipes MyMedia)

– SMO : optimisation pour les outils sociaux, référencement mobile (sites et applications), Netlinking, optimisation de plateformes, framework, CMS, AMOA / AMOE sur migrations de sites

– Gestion de projets, accompagnement stratégique, gestion du multilinguisme et du multipays, formation au référencement, reporting et dashboards avancés.

2.5 MyMedia.

L'agence MyMedia est aujourd'hui la première agence média indépendante française avec un volume d'achat d'espace brut prévisionnel de 700 millions d'euros en 2013 et un portefeuille de plus de 100 marques issues de tous secteurs, principalement du E-commerce.

A l'heure où l'offre media s'atomise littéralement, MyMedia propose une solution marketing et media plus souple et réactive.

2.5 Les enjeux du groupe MyMedia

MyMedia est une agence media capable d'intervenir sur l'ensemble des problématiques de communication. Elle est attachée à répondre à des problématiques très ROIStes[5] (Return On Investment ou retour sur investissement) afin de démontrer directement et concrètement l'efficacité des médias.

Après plus de huit années d'activité, My Media est la 1ère agence media indépendante française avec la plus forte croissance du secteur et plus de 100 clients actifs.

Quels sont alors les enjeux de cette entreprise ?

Les clients (des marques) cherchent à gagner en visibilité par le biais de campagnes publicitaires. Mais il ne suffit pas uniquement d'investir de l'argent dans une campagne pour qu'elle soit réussie. Encore faut-il être en mesure de dire si elle fut bénéfique ou non. Les enjeux sont donc de modéliser toutes ces données et d'en tirer le maximum d'informations utiles.

Intervient alors MyMedia afin d'étudier (de monitorer) lesdites campagnes.

En effet, MyMedia achète des espaces publicitaires télévisés afin de les vendre, sous forme de campagnes publicitaires aux marques.

De plus, avec le lancement d'un nouveau projet, leadsmonitor version 2 dans lequel s'inscrit le projet de ce mémoire de recherche, MyMedia développe un nouvel outil de business intelligence qui permet à ses clients de suivre rigoureusement l'avancée de leur campagne. Elle leur donne la possibilité d'avoir des informations qui étaient jusque-là difficiles d'accès à savoir : le nombre de visite suite à la diffusion d'un spot publicitaire (dans les cinq minutes

[5] ROI est un acronyme utilisé pour le terme anglais Return On Investment ou retour sur investissement en français. La notion de R.O.I. est très présente pour mesurer la rentabilité des actions de marketing, notamment dans les domaines du marketing direct et du webmarketing où il est possible de mettre en relation de manière précise les coûts de campagne et l'activité commerciale générée. Le ROI s'exprime souvent à l'aide du chiffre d'affaires généré (ex : 1€ investi a rapporté 5€ de chiffre d'affaires.)
C'est évidemment une simplification de la mesure du ROI puisqu'il faut idéalement aussi prendre en compte la marge générée et les notions de life time value.

qui suivent la diffusion, comme sur le long terme), le trafic du site en temps réel, le taux de transformation des visiteurs en clients, le nombre de commande, le coût par visite immédiate, le nombre d'inscription…

Toutes ces informations permettent alors au client de déterminer l'impact de sa campagne.

3. *Problématique d'Etudes d'une campagne publicitaire :*

Etudier l'impact d'une campagne publicitaire télévisée sur les visites et d'autres KPIs [6] d'un site E-commerce nécessite la disposition de plusieurs informations. En France ces informations sont fournies par les différentes sources de Médiamétrie. Certaines de ces sources permettent une récupération immédiate et en temps réel des données, d'autres donnent des informations en différentes fréquences.

L'outil de la Pige:

C'est un outil mis en place par l'institut anglais Kantar. Cet outil est en grande partie utilisé pour obtenir des informations concernant les versions et les créations des écrans publicitaires.

Médiamétrie :

Le GRP est un indicateur de pression publicitaire essentiellement utilisé pour le media TV. Le GRP acronyme de "Gross Rating Point" qui correspond au nombre moyen de contacts publicitaires obtenus sur 100 individus de la cible visée. GRP = (couverture en %) x (répétitions moyennes).[7] Par ailleurs il existe plusieurs types de GRP tout comme homme15+ (qui veut dire les hommes de plus de 15 ans)

Toutes les informations des GRP de référence et les GRP de cible sont fournis par Médiamétrie qui traite les informations de tous les médias: Télévision, Radio, Internet.

Le GRP de cible définit la structure de l'audience de l'écran. Par le biais du GRP, le client peut connaître le nombre de contact généré par son écran pour chaque cible.

L'outil d'achat d'espace télé PEAKTIME:

Le logiciel PEAKTIME, permet aux agences de média actives dans le domaine DRTV d'acheter des écrans publicitaires auprès des régies des chaines.

[6] KPI est un acronyme pour Key Performance Indicator. Les KPIs ou ICPs (indicateurs clés de performance) peuvent être utilisés, entre autres, dans le domaine du management au sens large, dans le domaine du marketing ou dans le domaine de l'analyse d'audience d'un site web.
Dans un contexte marketing, les KPIs sont utilisés pour déterminer les facteurs pris en compte pour mesurer l'efficacité ou la rentabilité d'une campagne. Pour une action de marketing direct, les KPIs retenus peuvent être par exemple le nombre de catalogues demandés, le nombre de commandes effectuées et le C.A. généré.
[7] http://www.definitions-marketing.com

Généralement, la vente des écrans se fait entre 3 à 7 jours avant le passage des spots. En ce moment, l'outil de PEAKTIME fournit une estimation de l'heure de passage des écrans et GRPs.

Le lendemain, lors de la prochaine étape, le logiciel fournit une nouvelle version, plus précise que la première, et généralement, une troisième version (dite consolidée) est fournie huit jours plus tard. Cette troisième version demeure la version la plus précise.

L'outil de TvTy :

Tvty est un outil qui contrairement aux autres outils cités ci-dessus, fournit les informations concernant la version et l'heure de passage des écrans en temps réel. Les heures de passages sont fournies par cet outil avec une précision à la seconde.

Tvty est un outil de « pattern matching » (vérification de la présence de constituants d'un motif par un programme informatique, ou parfois par un matériel spécialisé) qui écoute la télévision française, et à la seconde du passage d'écran publicitaire, enregistre l'heure et la date de passage, la création du spot et le nom du spot.

Malgré tout, il est difficile de trouver l'heure exacte de passage d'un écran. D'une part, il se peut que l'écran passe avec une différence d'une demi-heure avec l'heure indiquée dans le plan d'achat et d'autre part, le code écran peut ne pas se retrouver dans la liste des informations générées par Tvty. Or, faire la correspondance entre l'heure du plan et l'heure exacte de passage de l'écran a toujours été une tache délicate. C'est une question d'autant plus importante que cette information s'avère primordiale pour un calcul correct des gains immédiats[8].

Après avoir récolté toutes ces données, il faut savoir les interpréter ou les utiliser dans les calculs du gain de KPIs. Quand on en vient à parler de gain, il faut savoir par rapport à quelle mesure l'évaluer. Ce gain est alors calculé par rapport à une Baseline (base du trafic) qui est une mesure probabilistique qui définit quel aurait été le trafic du site s'il n'y avait pas de campagne publicitaire.

[8] Dans le domaine du DRTV, ce gain, nommé gain immédiat ou visites immédiates, est calculé à partir de la différence des visites cinq minutes après le passage d'écran et une Baseline avant le passage d'écran.

4. corpus

4.1 Corpus de description des émissions

Récoltes des données :

- Aspirer un des sites des programmes télé: PARIS PREMIERE et Télé 7 Jours
- Nettoyage des données aspirées : [9]
    - Mettre les pages html en xhtml
    - Récupérer la balise titre de ces fichiers
    - Récupérer les balises de textes de ces fichiers xhtml
    - Constituer une liste en format de fichiers texte (descriptions des émissions en format de csv) [10]
- Supprimer les mots vides des textes du corpus.

Cette étape de prétraitement du corpus a été effectuée à l'aide d'un programme en java qui prend en entrée une liste des mots vides et un corpus nettoyé de tous les mots de la liste.[11]

La liste des mots vides supprimés des textes est similaire à la liste ci-dessous :

- à au aux
- De Du Des
- Et, ou, en
- Dans
- L'
- La
- Le
- Les
- Un, une
- Que, Qui,
- Par, avec, dans
- Sous, Sur
- Cette, ce, ça, cela

[9] Aujourd'hui, toute la procédure est automatisée par les .jars. Ces jars se lancent automatiquement tous les jours pour enrichir la base de la description des émissions et garder cette base à jour.
[10] Cette dernière étape est effectué pour rematcher les émissions à leurs description.
[11] Le programme java (les mots vides) a été développé dans le cadre du projet universitaire pour le cours de java.

La stemmisation des textes pourrait être l'étape finale de la constitution du corpus mais l'étude des différentes formes de mots pourrait s'avérer nécessaire durant le projet. Or en premier lieu, on va garder les mots dans leur forme d'origine.

Le dernier traitement sur le corpus, porte sur la suppression des fichiers ne contenant qu'une ligne dont le titre de l'émission. A ce jour, le nombre total de documents dans ce corpus s'élève à 3705 fichiers et le corpus d'étude, contenant 68 documents, est extrait de ce premier. Ces 68 fichiers sont le résultat du matching de la première ligne des fichiers de description (qui est le titre de l'émission) et le titre des émissions dans le plan de campagne du client de test. Dans le plan de campagne de client, 2407 émissions distinctes ont été repérées.

Mais les émissions ont été aspirées à partir du mois de septembre et la campagne du client a commencé au mois de mai, d'où la différence entre le nombre des émissions dans le plan de campagne publicitaire du client et le nombre des émissions matchées.

Pour rester impartial dans les analyses, il a été nécessaire qu'on garde le nom complet de l'émission avec le titre de l'épisode ; une information supplémentaire qui rend le matching des titres des émissions avec les fichiers de descriptions des émissions plus compliqué.

Les étapes de prétraitement du corpus ont été toutes effectuées au cours de ce projet avec les outils et les scripts développés spécialement pour cette étude. Les bases de données MYSQL et les serveurs UNIX représentent l'environnement technique et les outils ont été développés en JAVA (.jars) et les scripts en PHP.

Un exemple de ces fichiers de description :

Emission
LUC FERRY DANS MEDIAS LE MAG FRANCE 5 ARRETEZ CES CONNERIES PUTAIN
MICHEL DRUCKER LE MAGNIFIQUE
L'AMOUR EST DANS LE PRE 9 M6
LES REINES DU SHOPPING M6 CORALIE FAIT SENSATION
UNE NOUVELLE SAISON POUR NOS CHERS VOISINS

Un fichier extrait du corpus d'étude :

Les reines du shopping (M6) : Coralie fait sensation en lingerie sexy - News Télé 7 Jours

Tout avait pourtant bien commencé sur M6 dans

Les reines du shopping, mardi 30 septembre. Sur le thème

C'est le jour de votre mariage, Coralie, gracieusement baptisée "Lady Gaga du n'importe quoi" par l'une de ses concurrentes, a d'abord présenté une robe noire satinée du plus bel effet. Arpentant le catwalk avec assurance sur ses chaussures à talons rouges, parsemant sa

performance de moult ondulations lascives, elle a laissé les autres participantes de marbre et enthousiasmé Cristina Cordula. La présentatrice n'était pas au bout de ses surprises.

Car tout ceci n'était en fait que le premier volet du défilé ; après un bref retour en coulisses, Coralie a surgi en lingerie fine face à des concurrentes aussi hilares que médusées. Un deuxième aller-retour osé qui montre bien que la candidate nordiste, trois fois mariée, n'a pas oublié de quoi retourne une union pour la vie. L'une des candidates lâche entre deux larmes de rire :

"C'est la cerise sur le gâteau". Cristine Cordula, elle, a adoré. Et Coralie s'est visiblement bien amusée. Encore une preuve que l'essentiel est définitivement d'être bien dans ses vêtements - même quand on en a peu.

Cet épisode amusant ne peut que servir les audiences des Reines du shopping, déjà excellentes : l'émission a battu son propre record la semaine dernière en rassemblant en moyenne 1,4 million de spectateurs par jour entre le 22 et le 26 septembre.

Sébastien Wesolowski

4.2 Les données des spots

Les données associées aux spots publicitaires arrivent chez l'agence sous la forme de plans de campagne en format dit Popcorn. Ce sont des plans détaillés de spots comme nous pouvons le constater dans le tableau 1. Ce qui nous intéresse dans ces plans de campagne est la date et l'heure de diffusion dans le calcul des gains, les typologies des émissions ainsi que leurs titres dans la modélisation des réseaux neuronaux.

Date	Ecran	Support	H diff	Format	GRP 25 +	GRP 4+	Emission	TAP	Emission après
20/10/2014	739	PUISSANCE TNT	07:09:00	30	0.1	0.1	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-LA BANDE A RENAUD	W9:MAGAZ.QUOI DE NEUF - W9	
31/10/2014	739	PUISSANCE TNT	07:09:00	30	0.1	0.1	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-ZAZ	W9:MAGAZ.QUOI DE NEUF - W9	
29/10/2014	739	PUISSANCE TNT	07:12:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-ZA	
30/10/2014	809	PUISSANCE TNT	07:31:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:CLIP.WAKE UP	
20/10/2014	809	PUISSANCE TNT	07:40:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:CLIP.WAKE UP	
21/10/2014	809	PUISSANCE TNT	07:40:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:CLIP.WAKE UP	
22/10/2014	809	PUISSANCE TNT	07:41:00	30	0.0	0.0	W9:CLIP.WAKE UP	W9:MAGAZ.QUOI DE NEUF - W9	
23/10/2014	809	PUISSANCE TNT	07:41:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:CLIP.WAKE UP	
24/10/2014	809	PUISSANCE TNT	07:41:00	30	0.0	0.0	W9:CLIP.WAKE UP	W9:CLIP.WAKE UP	
31/10/2014	809	PUISSANCE TNT	07:41:00	30	0.2	0.2	W9:CLIP.WAKE UP	W9:CLIP.WAKE UP	
29/10/2014	809	PUISSANCE TNT	07:50:00	30	0.0	0.0	W9:CLIP.WAKE UP	W9:CLIP.WAKE UP	
23/10/2014	839	PUISSANCE TNT	08:12:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-LA	
22/10/2014	839	PUISSANCE TNT	08:14:00	30	0.0	0.0	W9:SPECT.W9 HOME FESTIVAL-SELAH SUE	W9:SPECT.W9 HOME FESTIVAL-HOLLYSIZ	
28/10/2014	839	PUISSANCE TNT	08:14:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-ZA	
21/10/2014	839	PUISSANCE TNT	08:15:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-LA	
20/10/2014	839	PUISSANCE TNT	08:17:00	30	0.1	0.1	W9:CLIP.WAKE UP	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-LA	
24/10/2014	839	PUISSANCE TNT	08:18:00	30	0.1	0.0	W9:CLIP.WAKE UP	W9:MAGAZ.PLUS VITE QUE LA MUSIQUE-LA	
26/11/2014	900	T.F.1	09:11:00	15	0.1	0.1	TELESHOPPING	METEO	
29/10/2014	930	PUISSANCE TNT	09:13:00	30	0.1	0.1	6TER:SERIE.SYDNEY FOX AVENTURI,W9:SPECT.W9 HOME FE	6TER:SERIE.SYDNEY FOX AVENTURI,W9:CLI	
31/10/2014	930	PUISSANCE TNT	09:14:00	30	0.2	0.2	6TER:SERIE.SYDNEY FOX AVENTURI,W9:CLIP.W9 HITS	6TER:SERIE.SYDNEY FOX AVENTURI,W9:CLI	
28/11/2014	900	T.F.1	09:15:00	15	0.3	0.2	TELESHOPPING	METEO	
25/11/2014	900	T.F.1	09:16:00	15	0.3	0.2	TELESHOPPING	METEO	
27/11/2014	900	T.F.1	09:16:00	15	0.2	0.2	TELESHOPPING	METEO	

Tableau 1 Extrait de plan de campagne

La colonne TAP du tableau 1 représente la catégorie de l'émission, comme par exemple TELESHOPPING.

GRP 25+ est le GRP cible de client et comme c'est expliqué dans la section 3 de ce mémoire, grâce à GRP cible client peux définir sa structure d'audience.

4.3 Les données des visites et d'autres KPIs

Les données des visites et d'autres KPIs sont récoltées grâce à la cookie posé sur le site du client. À chaque visite, une ligne est insérée dans les bases de données avec un id de session et beaucoup d'autres informations. Ces données permettent à l'entreprise de calculer les gains de visites ou d'autres KPIs au cours de leur campagne publicitaire ou même après. Ici nous allons montrer le lien entre ces gains et le contexte d'écran publicitaire.

5. Modélisation des émissions

5.1 Les réseaux de neurone

Dans ce chapitre nous modéliserons les émissions (le contexte des spots) mais avant cela nous présenterons une définition brève des réseaux de neurones et le perceptron.

5.1.1 Un neurone artificiel

Les réseaux de neurones sont inspirés par les neurones biologiques. Le schéma montre la composition d'un neurone artificiel. Chaque neurone faisant partie du réseau est une machine élémentaire. Il reçoit un certain nombre de variables d'entrées provenant de neurones en amont. À chacune de ces entrées, est associé un poids qui représente la force de la connexion entre les deux neurones. Chaque neurone est doté d'une unique sortie qui se ramifie ensuite pour représenter les entrées qui alimenteront d'autres neurones en aval. Pour résumer, chaque neurone calcule une sortie unique en se basant sur les informations qui lui sont données.

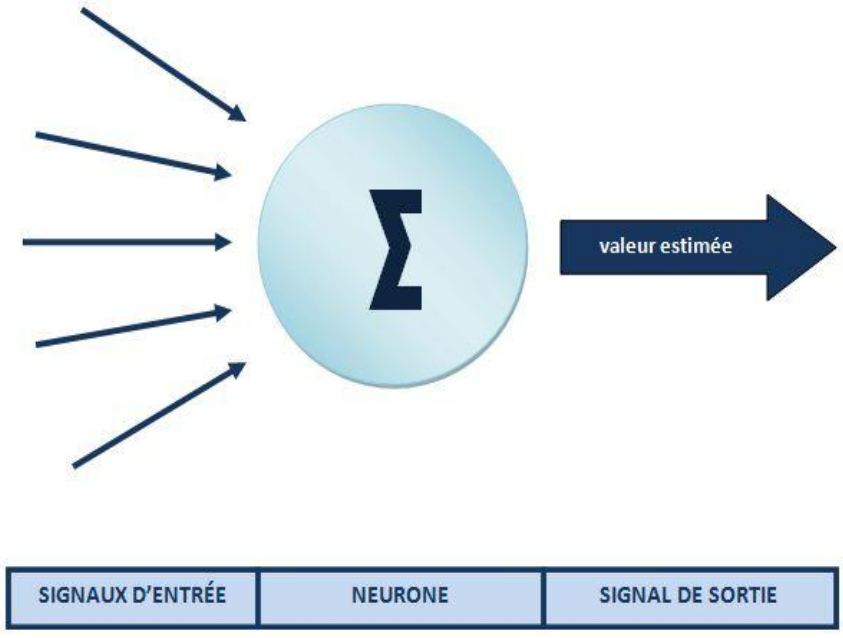
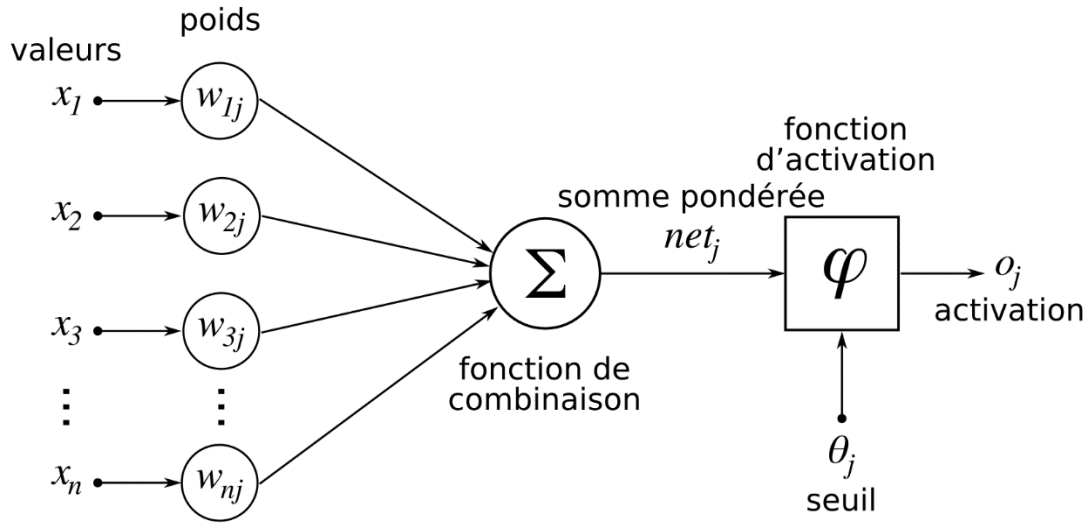


Figure 1. Schéma d'un neurone artificiel



Le neurone en tant qu'unité élémentaire agit de la façon suivante :

1ère phase : Le neurone fait le calcul de la somme pondérée des entrées (en fonction de la force des connexions). L'apprentissage ayant été réalisé auparavant, le poids des connexions est ici déjà déterminé et fixe.

2ème phase : À partir de la valeur obtenue, une fonction d'activation ou de transfert va calculer la valeur de l'état du neurone. Le neurone compare la somme pondérée des entrées à une valeur de seuil et fournit alors une réponse en sortie.

La majorité des fonctions sont continues et offrent donc une infinité de valeurs possibles comprises dans les intervalles [0;+1] ou [-1;+1]. Il existe différents types de fonctions de transfert pour le neurone artificiel à savoir la fonction à seuil, la fonction linéaire par morceaux, la fonction sigmoïde, la fonction gaussienne…

5.2 Perceptron

5.2.1 Définition :

Le perceptron est le premier réseau neuronal conçu par **Rosenblatt** (1958). Il vise à associer des réponses à des patterns présentés en entrée (problèmes de reconnaissance). Il se compose de deux couches de neurones:

1) La première couche, ou **rétine** est composée de cellules binaires perceptives (OUT = 0 ou 1).

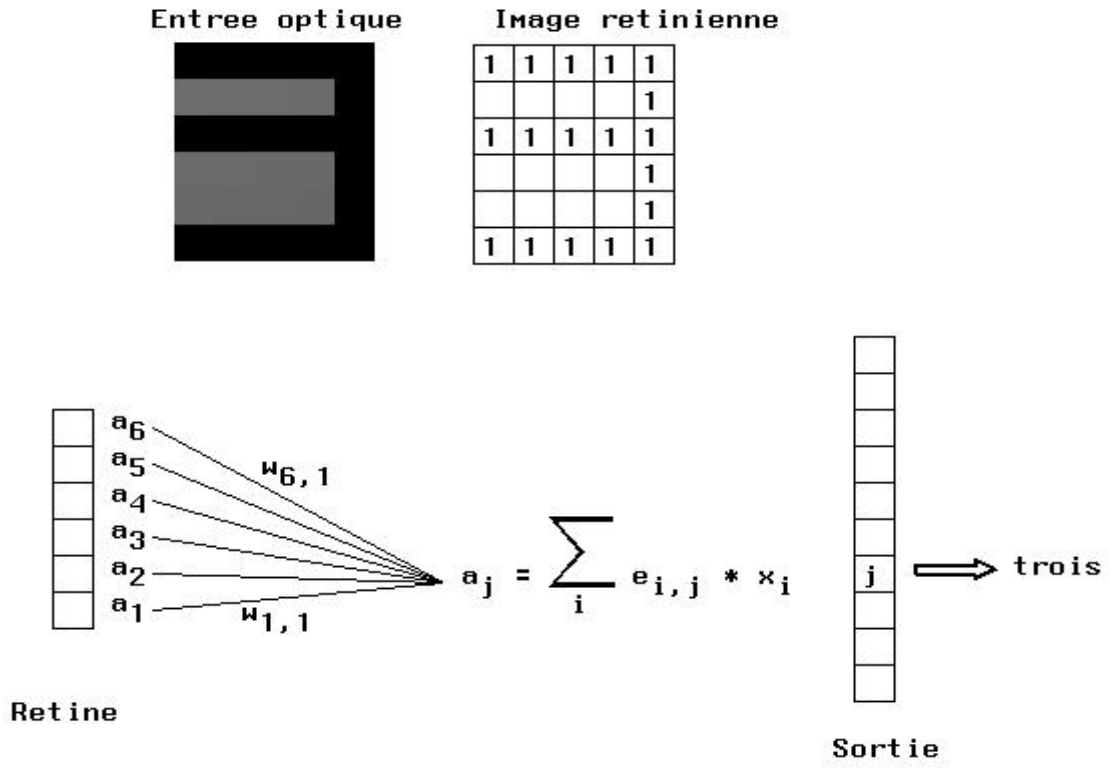
2) La deuxième couche fournit la réponse.

Par exemple la rétine peut comporter 5 * 6 cellules et la sortie 10 cellules. Le problème

est de reconnaître un chiffre (parmi 0, 1, 2, 3, 4, 5, 6, 7, 8,9) impressionnant la rétine. Si le chiffre *4* est présenté, la cinquième cellule de sortie doit être active et toutes les autres doivent être passives.
Toutes les cellules de la rétine sont reliées à chacune des cellules de la sortie par des **poids** variables. Si les cellules de la rétine sont indexées par i (variant de 0 a 29) et celles de la sortie sont indexées par j (variant de 0 a 9), l´entrée a_j d´une cellule de sortie j est la somme pondérée de toutes les sorties des cellules de la rétine (voir figure 4-1):

$$a_j = \sum_{i=0}^{i=29} w_{i,j} * X_i$$

x_i = sortie (0 ou 1) de la cellule i de la rétine.
$w_{i,j}$ = poids de la liaison i-> j.
Une fonction seuil permet alors de déterminer la sortie de la cellule j:
$x_j = (a_j <= T)\ ?\ 0 : 1$
On peut remplacer le seuil T par un seuil nul à condition d´ajouter une cellule de la rétine toujours active et de poids -T.



5.2.2 Règle d´apprentissage

La règle la plus connue est celle dite de **Widrow-Hoff** (1960) ou **règle du delta**. Elle est locale, c´est à dire que chaque cellule de sortie apprend indépendamment des autres.
Une cellule de sortie ne modifie les poids des liaisons qui aboutissent à elle que si elle se trompe:
Si elle est active alors qu´elle devrait être passive alors elle diminue les poids correspondants aux cellules actives de la rétine.
Si elle est passive alors qu´elle devrait être active alors elle augmente les poids correspondants aux cellules actives de la rétine.
L´algorithme de l´apprentissage se déroule de la façon suivante:

1) On présente au perceptron (dans un ordre arbitraire) les couples (image sur la rétine, réponse).

2) S´il y a des erreurs alors les poids sont corrigés selon la règle du delta et on retourne en 1, sinon le perceptron a appris.

La règle de Widrow-Hoff peut s´écrire:

$w_{i,j}(t + 1) = w_{i,j}(t) + n * (t_j - o_j) * x_i = w_{i,j}(t) + dw_{i,j}$

$w_{i,j}(t + 1)$ = poids de la liaison i -> j au temps t + 1
$w_{i,j}(t)$ = poids de la liaison i -> j au temps t
x_i = sortie (0 ou 1) de la cellule i de la rétine
o_j = réponse de la cellule j de sortie
t_j = réponse théorique (souhaitée) de la cellule j de sortie
n = constante positive (entre 0.0 et 1.0): Une bonne valeur est 0.75, mais il est préférable de faire évoluer n au cours du temps: D´abord grande (0.9) puis décroissante au cours de l´apprentissage.

5.2.3 Théorème de convergence du perceptron

On démontre que si une solution au problème existe alors le perceptron la trouvera en un nombre fini de cycles d´apprentissages.
Mais on montre que, comme pour un neurone artificiel, le perceptron ne peut résoudre que des problèmes linéairement séparables.
Pour traiter des problèmes non linéairement séparables, on complexifie le perceptron en lui ajoutant une couche (dite **couche cachée**) de neurones, en faisant un réseau neuronal plus général. C´est ce que fit Rosenblatt en 1962 pour résoudre le problème du XOR. Mais la correction des poids devient problématique puisqu´elle modifie les entrées de la couche cachée et donc ses sorties...
La technique, dite de **rétro-propagation de l'erreur** fut trouvée en 1969 par Bryson et Ho, puis redécouverte en 1986 en particulier par **Rumelhart**.

5.3 Modélisation des émissions avec le réseau de neurones Perceptron

Au cours de ce projet, nous avons porté notre choix sur l'implémentation du perceptron sous MATLAB. (L'implémentation détaillée du perceptron sous MATLAB est ajoutée dans l'annex.1 de ce document.)

Pour modéliser le contexte des écrans publicitaires de notre client de choix, il y a un certain nombre de features à considérer :

- La typologie du programme avant l'écran publicitaire
- Le contexte avant l'écran passé (l'émission avant)
- La typologie du programme après l'écran publicitaire
- Le contexte après l'écran passé (émission après)
- La situation de l'écran : soit, entre deux émissions ou durant la même émission.

Le modèle de la typologie d'émission avant le passage d'écran est comme celui-ci :

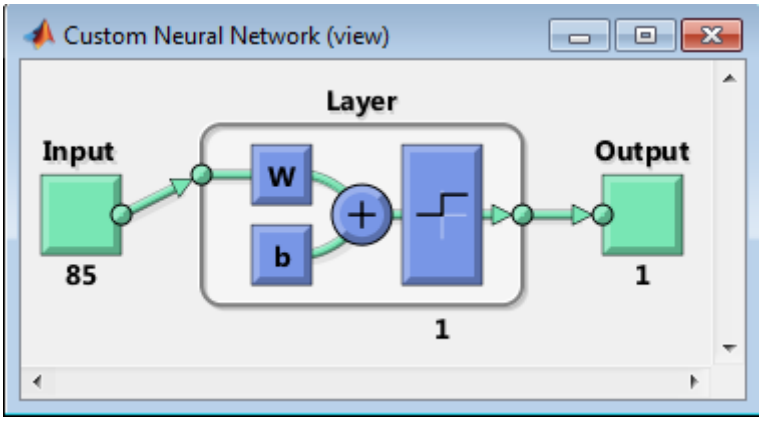


Figure 2. Le modèle du perceptron pour typologie avant

Les données d'entrée du modèle sont sous forme d'une matrice de 0 et 1. Les typologies des émissions sont les lignes et les spots (les samples) sont les colonnes. Les lignes de matrice d'entrainement contiennent toutes les typologies de tous les spots du client de test[12].

Pour chaque spot passé, la typologie de l'émission avant est initialisée à 1 et le reste des lignes de matrice de départ à 0.

[12] Le client de test œuvre dans le domaine de vacances et de voyages

Pour les données de cibles, qui sont modélisées sous forme d'une matrice contenant une ligne et autant de colonne que de spots, les spots qui apportent un gain peu satisfaisant ou pas satisfaisant appartiennent à la classe 0 et les spots satisfaisants appartiennent à la classe 1. Dans le domaine de la DRTV, ce gain, nommé gain immédiat ou visites immédiates, est calculé à partir de la différence des visites cinq minutes après le passage d'écran et une Baseline avant le passage d'écran.

Pour pouvoir déterminer si ce gain a été satisfaisant, plusieurs facteurs importent :

- L'heure de passage d'écran
- Le budget d'écran
- La chaine
- L'audience (les contacts générés)
- La création de la pub (ou le film)[13]
- L'émission avant et après l'écran.

C'est pourquoi les gains immédiats des écrans sont normalisés par la moyenne regroupée, la création et la durée d'écran et uniformisés[14] par l'heure de passage des spots (les valeurs entre 0 et 23) et le jour de la semaine (les valeurs entre 0 et 6 qui correspondent au lundi au dimanche).

Les gains sont débruités et seuls, sont gardés, les spots sur lesquels le signal de l'écran publicitaire est lisible. Le seuil de satisfaction est basé sur ce facteur. C'est-à-dire quand le

[13] La création de spot signifie les images ou le film de l'écran.

[14] Le processus d'uniformisation permet de calculer des notes plus robustes pour chaque
Index. En effet, le calcul de la note peut résoudre le problème d'une répartition inéquitable des
index, qui est un co-occurrente non uniforme de l'indice. Si nous considérons que nous avons un sac
Indice de Ii, I ∈ {1, 2} (notamment dans le cas de jour indice sera dans {0, 1,2,...,6}) chacun indice prend sa valeur dans S = {1, 2} puis envisager le calcul du score de la valeur de l'indice I_1égal à 1, cela signifie $S_1^{(1)} = \frac{1}{N} \sum_{i=1}^{N} \frac{Gain_i}{GRP_i}$

où N est le nombre de tuple I1 et qui a l'indice égal à 1, cette moyenne peut être décomposée comme suit:

$$S_1^{(1)} = \frac{1}{N} \times \left(\sum_{i=1}^{N^{(1)}} \frac{Gain_i}{GRP_i} + \sum_{j=1}^{N^{(2)}} \frac{Gain_j}{GRP_j}\right)$$

Où N (1) et N (2) sont respectivement le nombre de tuple qui a indice I2 égal à 1 et
a Indice I2 égal à 2. S'il arrive que la co-occurrence de (I1: 1, I2: 1) est très fréquente, alors le score $S_k^{(1)}$ reflète une indésirable priorité de l'indice I_1qui va à l'encontre de l'hypothèse 1 uniforme avant. Le processus d'uniformisation consiste en informatique à calculer un sous-score de chaque index, fixer la valeur d'autres index et enfin de calculer la moyenne du sous-score pour obtenir le score. Dans l'exemple précédent, la formule serait:

$$S_k^{(1)} = \frac{1}{2} \times \left(\frac{1}{N^{(1)}} \sum_{i=1}^{N^{(1)}} \frac{Gain_i}{GRP_i} + \frac{1}{N^{(2)}} \sum_{i=1}^{N^{(2)}} \frac{Gain_i}{GRP_i}\right)$$

Ce processus nous permet d'établir une certaine équité dans la répartition de la co-occurrence de
L'Index et pour améliorer la robustesse du score sans ajouter de partialité.

signal est lisible, et avant cela, l'écran est considéré comme non satisfaisant et il appartient à la classe de 0.

Les tests et l'évaluation effectués sur ce modèle montrent une performance entre 40 à 50%, ce qui n'est pas idéal.

Dans le but d'avoir une meilleure performance et une meilleure précision, intégrer les émissions avant ou après dans le modèle comme feature, s'est avéré nécessaire.

En Médiamétrie, l'importance est portée sur les émissions après. Or, le modèle de l'émission avant ne répondant pas, ce choix d'émission a été exclu.

Finalement, le projet se basera sur les données des émissions après le passage de l'écran publicitaire.

Le modèle perceptron pour les émissions après est comme ceci :

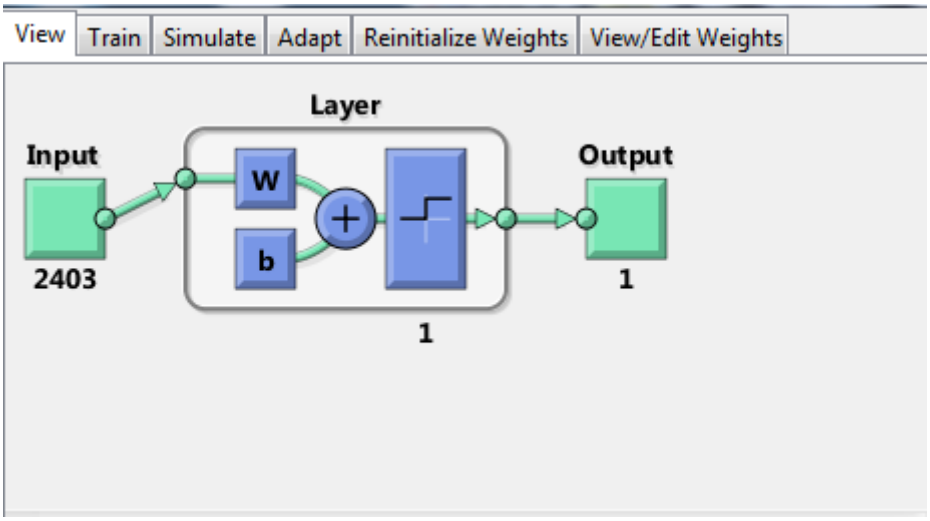


Figure 3. Le modèle de perceptron pour l'émission après

La matrice d'entrée est constituée à partir des données de spots et la matrice des données de cibles se base sur les gains de chacun des spots.

Les lignes de la matrice d'entrée représente chacune une émission et il y a autant de colonnes que de spots (samples).

L'entrainement pour le modèle a été effectué sur 20% des spots (l'outil MATLAB n'acceptant pas une matrice plus grande que 2403 × 2500, j'ai été dans l'obligation d'entrainer le modèle sur 20% des données, ce qui va influencer la performance du réseau de neurone) et le test comprend également 20% des données. Une évaluation simple du test sur les données montre une exactitude de 47.33%.

Modèle	Entrainement	Test	Evaluation
Les typologies avant	20%[15]	20%	30,33%
Les émissions après	20%	20%	47,66%

Tableau 2.Récapitulatif des modèles

Maintenant qu'on a réussi à modéliser la performance des émissions pour les spots, nous pouvons analyser le contexte des émissions les plus performantes pour le client de test durant sa dernière campagne.

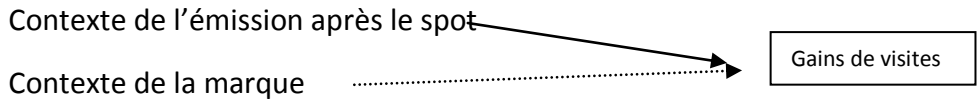
6. Etudes contextuelles :

Ce qu'on appelle dans ce projet le contexte des émissions, c'est le contexte des émissions avant et après chaque passage d'écran.

La question qui se pose est : Y a-t-il un lien établi entre le contexte de spots, le contexte de la marque et la performance de l'écran ?

Premièrement, le rapprochement du contexte du spot et de la marque a-t-il un impact sur la performance du spot.

Deuxièmement, le contexte de l'écran à lui seul a-t-il un impact sur le gain du spot.



Après avoir modélisé les gains des spots et les avoir classifié dans deux groupes 0 et 1, j'ai réussi à matcher 68 sur 400 émissions se trouvant dans le corpus d'étude.

Ci-dessous nous voyons un tableau de spécificité de notre corpus d'étude :

Description du corpus d'étude :

Statistiques générales du corpus y compris les mots vides:

- 18 937 mots
- 3756 mots distincts
- Moyenne de 278 mots par texte
- Et 68 fichiers
- Spécificités du corpus [16] (sans les mots vides)

[15] L'outil de Matlab n'a pas permis d'entrainer le modèle sur plus de 2500 spots c'est-à-dire 20 pourcent des données.

mots	Fréquence
TF1	45
soit	44
amour	41
ai	41
cette	40
En	40
Amour	40
millions	40
je	40
audience	39
part	35
deux	34
elle	33
leur	33
ans	30
tout	30
fait	30
film	30
semaine	27
Une	26
ses	26
mais	26
Bien	25
Place	25
Nouvelle	25
Public	25
Famille	25
Comme	24
Nous	24

Tableau 3. Spécificités du corpus d'étude

Pour étudier les textes de description des émissions, je vais utiliser le modèle bag of words.[17] Dans une première étape, on va considérer chaque fichier de description comme un sac de mots. Le score associé à chaque mot représente l'occurrence du mot dans le fichier en question, normalisé par la taille du fichier et le classement des 30 premiers mots (hors mots vides et doublons), comme illustré dans le tableau 4.

[16] Exporté dans TXM

[17] Dans l'algorithme de bag of words on considère que le monde peut être décrit au moyen d'un dictionnaire (de « mots »). Dans sa version la plus simple, un document particulier est représenté par l'histogramme des occurrences des mots le composant: pour un document donné, chaque mot se voit affecté le nombre de fois qu'il apparaît dans le document. Un document est donc représenté par un vecteur de la même taille que le dictionnaire, dont la composante *i* indique le nombre d'occurrences de l'i-ème mot du dictionnaire dans le document.

mots	score
Est	4.9382716
Pré	3.7037037
Millions	3.7037037
Soit	3.7037037
Série	3.7037037
téléspectateurs	3.7037037
Saison	3.26171369
Chaîne	3.21543408
France	3.21482602
Pas	3.21114814
Télé	3.20945946
On	3.20097245
Tv	3.18965517
Star	3.18471338
People	3.17936413
Sont	3.17700454
Son	3.16831683
Cest	3.16613305
Amour	3.16205534
Je	3.15416801
Plus	3.14919979
News	3.14269536
Il	3.13801722
Septembre	3.12198262
Disponible	3.1201248
Lancement	3.11688312
Nouvelle	3.11643836
Famille	3.11566368

Tableau 4. Occurrences normalisées des mots

Mais ce classement s'est avéré insuffisant pour trouver un ou des centres de cluster pour les émissions du corpus. De plus, il y des mots qui n'ont pas une occurrence très forte, cependant ils sont présents dans plusieurs documents. C'est pourquoi j'ai décidé de calculer une fréquence pondérée par le nombre de documents présents dans le corpus.

A partir de la méthode de bag of words et en s'inspirant de la mesure TF-IDF, un classement pondérés de l'occurrence des mots est présenté.

La pondération se fait sur la base du total des documents dans lesquels le mot a été présent.

Score du mot = Freq(mot) * (doc_nombre) / (total doc)

Où le Freq(mot) est la fréquence totale du mot dans tout le corpus,

doc_nombre est le nombre de documents dans lesquels le mot est présent et total_doc est le nombre total de fichiers dans le corpus.

Ne voulant pas mettre l'accent sur un mot en particulier et dire quelle description d'émission est la plus pertinente pour tel ou tel mot, j'ai jugé préférable de ne pas calculer la mesure TF-IDF.

Voici le résultat pour les 30 premiers mots les plus fréquents dans le corpus .

mot	doc_nombre	occurrence_total	nombre total de documents	score
est	51	270	68	202.5
amour	21	150	68	46.3235294
tv	29	105	68	44.7794118
pre	2	104	68	3.05882353
audiences	25	100	68	36.7647059
france	32	100	68	47.0588235
qui	44	93	68	60.1764706
pas	34	84	68	42
télé	65	73	68	69.7794118
son	36	67	68	35.4705882
plus	34	64	68	32
téléspectateurs	24	61	68	21.5294118
je	12	61	68	10.7647059
news	59	59	68	51.1911765
on	25	59	68	21.6911765
pré	29	54	68	23.0294118
sa	28	54	68	22.2352941
famille	13	52	68	9.94117647
saison	25	50	68	18.3823529
mais	21	50	68	15.4411765
ont	28	48	68	19.7647059
il	21	47	68	14.5147059
people	9	40	68	5.29411765
soit	19	39	68	10.8970588
toujours	15	39	68	8.60294118
série	20	38	68	11.1764706
deux	22	38	68	12.2941176
leur	20	37	68	10.8823529
cest	19	36	68	10.0588235

Tableau 5. Classement pondéré des mots

C’est un corpus de description, or :

- Les verbes sont conjugués à la troisième personne
- Les mots sont, la plus part du temps, positifs
- Le mot famille remonte ce nouveau classement
- Pas de rapprochement avec le contexte du client qui est « vacances et voyage »
- Pas de thématique particulière pour que le contexte d’un spot soit réussi.

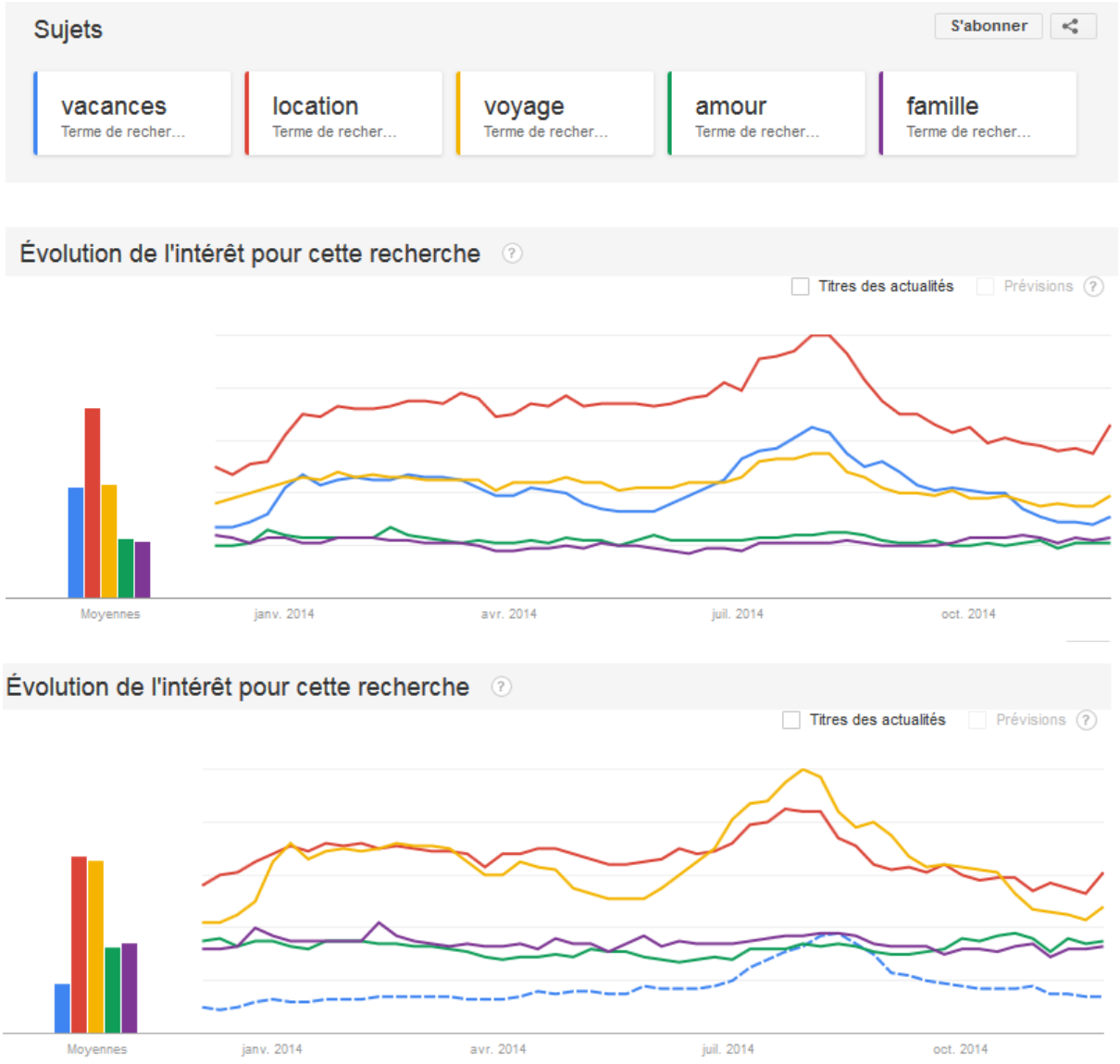


Figure 4. Les mots clé

La figure 4 illustre les courbes des mots clé vacances, location, amour, voyage et famille exportées de GOOGLE TRENDS. La courbe en pointillée montre les mots clé du client test.

Sur ces courbes, on voit bien la corrélation des mots clé du client et les mots vacances, location et voyages, mais visiblement il n'y a pas de corrélation entre les mots les plus significatifs en haut du classement des mots du corpus d'étude.

En revanche la corrélation entre le mot amour et famille est évidente. Ce qui signifie qu'il existe des centres de cluster dans notre corpus d'études qui ne sont pas forcément corrélés au contexte du client.

Cela n'empêche que la présence récurrente du mot famille et amour dans le corpus reste très intéressant car le client de test cible le grand public. Le contexte des émissions montre que les émissions qui se regardent en famille ont plus de succès pour notre client test. Selon les résultats de cette étude, on ne voit pas de lien direct entre le contexte des écrans publicitaires et le contexte du client en question. Cependant d'une manière indirecte, il y a un lien entre le contexte de l'émission et le succès de l'écran qui passe par la structure de l'audience et la cible du client.

En perspective de ce travail, nous envisageons de regarder si, par l'utilisation de la structure d'audience (cibles), il serait possible d'établir des corrélations entre des descriptions des contextes (émissions) et les gains générés.

7. Conclusion

Pendant cette recherche, on a essayé d'étudier d'abord l'impact des émissions avant ou après sur les gains des écrans publicitaires. La question était : Y a t-il un impact du contexte de l'écran sur le gain d'écran ? Si oui, que faire pour améliorer cet impact ?

La modélisation des gains des spots, par les émissions en utilisant les réseaux de neurones, a montré que oui, par le biais des émissions, il est possible de prédire si un écran impact d'une façon positive le trafic du site des clients.

Puis, nous nous sommes focalisés sur le contexte des émissions pour trouver le facteur influençant les émissions. Certes, la typologie des émissions peut être très importante mais nous avons essayé de regarder le sujet sous un angle purement textuel, afin de mettre au clair, si la thématique des émissions influence l'effet de l'écran sur le trafic du site du client.

Les études ont démontré l'impact du contexte de l'émission sur la structure potentielle de l'audience. Cela veut dire que par l'étude contextuelle d'un spot, ce serait possible de recommander une émission qui est susceptible d'engendrer et d'attirer plus de contacts dans les cibles des clients.

Nous pensons par ailleurs qu'une clusterisation des émissions, non pas selon leurs thèmes mais selon leurs structures d'audience, pourrait être envisagée comme une piste d'amélioration. Cela pourrait aider à trouver des corrélations entre les émissions (contextes) et leurs audiences ? Répondre à cette question pourrait alors nous aiderà optimiser des plans d'achat de spots.

Bibliographie

Donghahi Guan, Weiwei Yuan, Young-Koo Lee, Andrey Gavrilov: ICIC 2007, CCIS 2, pp 1220-1226, Springer-Verlag Berlin Heidelberg 2007

R. Rojas: Neural Networks, Spring-Verlag, Berlin, 1996

R. Rajendra Prasath, Sudeshna Sarkar, “Unsupervised Feature Generation Using Knowledge Repositories for effective Text Categorization. ECAI 2010: 1101-1102

S. Goswami, M. Singh Shishodia, “A fuzzy based approachto text mining and document clustering”, Indian Institut of technology

INReV - Université Paris 8, https://www.inrev.org

wikiversity.org

lrde.epita.fr

http://aimotion.blogspot.fr/2011/12/machine-learning-with-python-meeting-tf.html

Annexe 1.

Rosenblatt [Rose61] créé de nombreuses variantes du perceptron. L'une des plus simples est un réseau à une seule couche dont les poids et les biases pourraient être formées pour produire un vecteur de cible correcte lorsqu'ils sont présentés avec le vecteur d'entrée correspondant. La technique de formation utilisé est appelé ''la règle d'apprentissage du perceptron''. Le perceptron a suscité un grand intérêt en raison de sa capacité à généraliser à partir de ses vecteurs de formation et d'apprendre des connexions initialement réparties de façon aléatoire. Les perceptrons, réseaux rapides et fiables, sont particulièrement adaptés à la résolution de problèmes simples classifiés par motifs. En outre, une compréhension des opérations du perceptron, fournit une bonne base pour la compréhension des réseaux plus complexes.

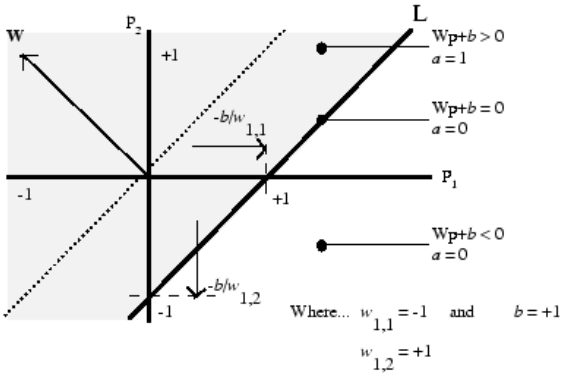
Le Modèle de Neurone

Chaque entrée externe est pondérée avec un poids de W1J approprié, et la somme pondérée des entrées est envoyée à la fonction de transfert de disque de réponse, qui a également une entrée de transmission à travers la polarisation. La fonction de transfert dur-limite, qui renvoie un 0 ou un 1, est présentée ci-dessous.

Le neurone de perceptron produit un 1 si l'entrée nette dans la fonction de transfert est égale ou supérieure à 0; sinon, il produit un 0.

La fonction de transfert de disque limité donne au perceptron la possibilité de classer les vecteurs d'entrée en divisant l'espace d'entrée en deux sections. Plus précisément, les résultats seront de 0 si l'entrée nette n est inférieure à 0, ou 1 si l'entrée nette n est 0 ou plus. La figure suivante montre l'espace d'une à deux entrées dur-limité de neurone avec les poids W1, 1 = -1, W1, 2 = 1 et un biais b = 1 entrée.

Deux sections de classification sont formées par la ligne frontalière de décision à L Wp + b = 0. Cette ligne est perpendiculaire à la matrice de pondération W et déplacée en fonction de la polarisation b. Les vecteurs d'entrée ci-dessus et à gauche de la ligne L se traduiront par une entrée nette supérieure à 0 et, par conséquent, provoqueront le disque limité de neurone à la sortie un 1. Les vecteurs d'entrée ci-dessous et à droite de la ligne L provoquent le neurone à la sortie 0. Vous pouvez choisir les valeurs de poids et les biases à orienter et déplacer la ligne de démarcation afin de classer l'espace d'entrée comme vous le souhaitez.



Les Neurones Hard-limites sans biases auront toujours une ligne de classement passant par l'origine. L'ajout d'un neurone de la polarisation permet de résoudre des problèmes lorsque les deux ensembles de vecteurs d'entrée ne sont pas situés sur des côtés différents de l'origine. Le parti pris permet la frontière de décision pour être déplacée loin de l'origine. Vous pouvez exécuter le nnd4db comme l'exemple de programme. Avec cela, vous pouvez déplacer une frontière de décision, prendre de nouvelles entrées à classer, et de voir comment l'application répétée de la règle de l'apprentissage donne un réseau qui ne classifie pas les vecteurs d'entrée correctement.

Architecture du perceptron

Le réseau est constitué d'une simple couche de perceptron de neurones de perceptron S connectés à des entrées de R au moyen d'un ensemble de poids wi, j, comme indiqué ci-dessous sous deux formes. Comme précédemment, les indices de réseau i et j indiquent que wi, j est l'intensité de la connexion de l'entrée à la jième neurone i.

La règle d'apprentissage du perceptron brièvement décrit est capable de former une seule couche. Ainsi, ici les réseaux à une seule couche sont seulement pris en compte. Cette restriction impose des limites sur le calcul d'un perceptron. Les types de problèmes de perceptrons sont capables de résoudre sont discutés dans les limitations et précautions.

Créer un Perceptron :

Vous pouvez créer un perceptron de la manière suivante:

net = perceptron;
net = configure (net, P, T);

où les arguments d'entrée sont les suivants:

P est une matrice R par Q de Q vecteurs d'entrée d'éléments de R chacun.

T S est une matrice Q par Q de vecteurs cibles de chacun des éléments de S.

Généralement, la fonction hardlim est utilisé dans les perceptrons, elle est donc la valeur par défaut.

Les commandes suivantes créent un réseau de perceptron avec un vecteur d'entrée d'un seul élément avec les valeurs 0 et 2, et un neurone de sortie qui peut être soit 0 ou 1:

```
P = [0 2];
T = [0 1];
net = perceptron;
net = configure (net, P, T);
```

Vous pouvez voir ce réseau créé en exécutant la commande suivante:

inputweights = {1,1} net.inputweights

qui donne les rendements

```
inputweights =
        retards: 0
       initFcn: 'initzero'
         apprendre: true
      learnFcn: «LEARNP '
    learnParam: (aucun)
         taille: [1 1]
     weightFcn: «dotprod '
   weightParam: (aucun)
      userdata: (votre information personnalisée)
```

La fonction d'apprentissage par défaut est LEARNP. L'entrée nette de la fonction de transfert est hardlim dotprod, ce qui génère un produit de la matrice de vecteur d'entrée et du poids et ajoute la polarisation d'entrée pour calculer le filet.

La fonction d'initialisation par défaut initzero est utilisée pour définir les valeurs initiales des poids à zéro.

De même,

biases s = net.biases {1}

donne

```
biais =
       initFcn: 'initzero'
```

apprendre: 1
learnFcn: «LEARNP '
learnParam: []
taille: 1
userdata: [1x1 structure]

Le Perceptron est formé sur des exemples de comportement souhaité. Le comportement souhaité peut être résumé par un ensemble de paires d'entrée, de sortie

p1t1, p2t1, ..., pQtQ

où p est une entrée pour le réseau, et t est la cible de sortie correspondant correct. L'objectif est de réduire l'erreur e qui est la différence t - entre un neurone de réponse et le vecteur cible t.

Le règle d'apprentissage de perceptron LEARNP calcule les changements désirés aux poids et les biases de la Perceptron, étant donné un vecteur d'entrée p et l'erreur e associé. Le vecteur cible t doit contenir des valeurs de 0 ou 1, car Perceptrons (avec des fonctions de transfert hardlim) peut reproduire ces valeurs.

Chaque LEARNP est exécuté, le perceptron a une meilleure chance de produire des résultats corrects. La règle du perceptron est prouvée à converger vers une solution en un nombre fini d'itérations si une solution existe.

Si un biais n'est pas utilisé, LEARNP essaye de trouver une solution en modifiant seulement le vecteur de poids w à un point vers vecteurs d'entrée pour être classé comme 1 et loin de vecteurs pour être classé comme 0. Il en résulte une frontière de décision qui est perpendiculaire à w et qui classifie correctement les vecteurs d'entrée.

Il y a trois conditions qui peuvent se produire pour un seul neurone fois un vecteur d'entrée p est présenté et la réponse un du réseau est calculé:

CAS 1. Si un vecteur d'entrée est présenté et la sortie du neurone est correct (a = e et t = t - a = 0), alors le vecteur de poids w est pas altérée.

Cas 2. Si la sortie du neurone est égal à 0 et aurait dû être une (a = 0 et t = 1 et t = e - a = 1), le vecteur d'entrée p est ajouté au vecteur de poids w. Cela rend le point de vecteur de poids de plus près au vecteur d'entrée, ce qui augmente la probabilité que le vecteur d'entrée soit classé comme une à l'avenir.

Cas 3. Si la sortie du neurone est égal à 1 et aurait dû être 0 (a = 1 et t = 0 et t = e - a = -1), le vecteur d'entrée p est soustrait du vecteur de pondération w. Cela rend le point de vecteur de pondération plus loin à partir du vecteur d'entrée, ce qui augmente la probabilité que le vecteur d'entrée soit classé comme un 0 dans le futur.

La règle d'apprentissage du perceptron peut être écrit de manière plus succincte en termes de l'erreur e = t - a et la modification à apporter au vecteur de poids Aw:

CAS 1. Si e = 0, puis faire un changement Aw égale à 0.

Cas 2. Si e = 1, puis faire un changement Aw égale à PT.

CAS 3. Si e = -1, puis faire un changement Aw égale à -pt.

Tous les trois cas peuvent alors être écrites en une seule expression:

Aw = (t-α) pT = epT

Vous pouvez obtenir l'expression des changements dans la biais d'un neurone en notant que le biais est tout simplement un poids qui a toujours une entrée de 1:

Ab = (t-α) (1) = e

Dans le cas d'une couche de neurones il faut

AW = (t-a) (p) = T e (p) T

et

Ab = (t-a) = e

La règle d'apprentissage du perceptron peut être résumée comme suit:

Wnew = Wold + epT

et

```
ps = LEARNP (w, p, [], [], [], [], e, [], [], [], [], [])
ps =
     1 2
```

Les nouvelles pondérations, alors, on obtient comme

```
w = w + dw
w =
    2.0000 1.2000
```

Le processus de trouver de nouveaux poids (et les biases) peut être répété jusqu'à ce qu'il n'y ait pas d'erreurs. Rappelons que la règle d'apprentissage du perceptron est garantie à converger en un nombre d'étapes fini pour tous les problèmes qui peuvent être résolus par un perceptron. Ils comprennent tous les problèmes de classification qui sont linéairement

séparables. Les objets à classer dans de tels cas peuvent être séparés par une seule ligne.

Vous pourriez vouloir essayer l'exemple nnd4pr. Il vous permet de choisir de nouveaux vecteurs d'entrée et d'appliquer la règle d'apprentissage pour les classer.

Entrainement

Si sim et LEARNP sont utilisés à plusieurs reprises pour présenter des entrées à un perceptron, et à modifier les poids de Perceptron et les biases s selon l'erreur, le perceptron finira par trouver le poids et les biases s. Les valeurs qui permettent de résoudre le problème, étant donné que les perceptrons peuvent résoudre. Chaque traversée par toutes les entrées et cibles des vecteurs de formation est appelé une passe.

Notez que le train ne garantit pas que le réseau qui en résulte. Vous devez vérifier les nouvelles valeurs de W et b en calculant la sortie du réseau pour chaque vecteur d'entrée pour voir si tous les objectifs sont atteints. Si un réseau ne fonctionne pas correctement, vous pouvez former davantage en appelant le train avec les nouveaux poids d'entrainement, ou vous pouvez analyser le problème pour voir s'il y a un problème approprié pour le perceptron.

Pour illustrer la procédure d'entrainement, de résoudre un problème simple. Considérons un perceptron du neurone avec une entrée de vecteur unique ayant deux éléments:

Ce réseau, et le problème que vous êtes sur le point de prendre en considération, Sont suffisamment simples pour que vous puissiez les suivre grâce à des calculs manuels. Le problème décrit ci-dessous suit que l'on trouve dans [HDB1996].

Supposons que vous avez le problème de la classification suivante et que vous souhaitez le résoudre avec une entrée de vecteur unique, réseau de perceptron à deux éléments.

{p1 = [22], t1 = 0} {p2 = [1-2], t2 = 1} {p3 = [- 22], t3 = 0} {p4 = [- 11], t4 = 1}

Utilisez les poids initiaux et les biases On note les variables à chaque étape de ce calcul à l'aide d'un nombre entre parenthèses après la variable. Ainsi, ci-dessus, les valeurs initiales sont W (0) et b (0).