

RAPPORT PROJET ENCADRÉ

Analyse de Comportements avec Twitter



PROMOTION : 2017/2018

ETUDIANTES :MA LING, SHA QIANQIAN

Git : https://gitlab-etu.fil.univ-lille1.fr/ma/PJB_SHA_MA

SOMMAIRE

Chapitre 1 Description générale du projet	3
1 Description de la problématique	3
2 Description générale de l'architecture de l'application	3
Chapitre 2 Détails des différents travaux réalisés	4
1.API Twitter	4
2.Préparation de la base d'apprentissage	5
2.1 Nettoyage des données	5
2.2 Construction de la base	5
3.Algorithme de classification	6
3.1 Mots clés	6
3.2 KNN.....	6
3.3 Bayes	7
4.Interface graphique	8
4.1 Copies d'écran	8
4.2 Manuel d'utilisation.....	12
Chapitre 3 Résultats de la classification avec le différentes méthodes et analyse	13
Chapitre 4 Conclusions.....	14

Chapitre 1 Description générale du projet

Ces dernières années, l'utilisation de réseaux sociaux comme "Twitter" s'est énormément démocratisée. De ce fait il peut être intéressant d'analyser les messages postés ("Tweets") par les utilisateurs afin d'en connaître leur nature, pour par exemple connaître leurs sentiments sur un produit, une personne, un concept, ...

1 Description de la problématique

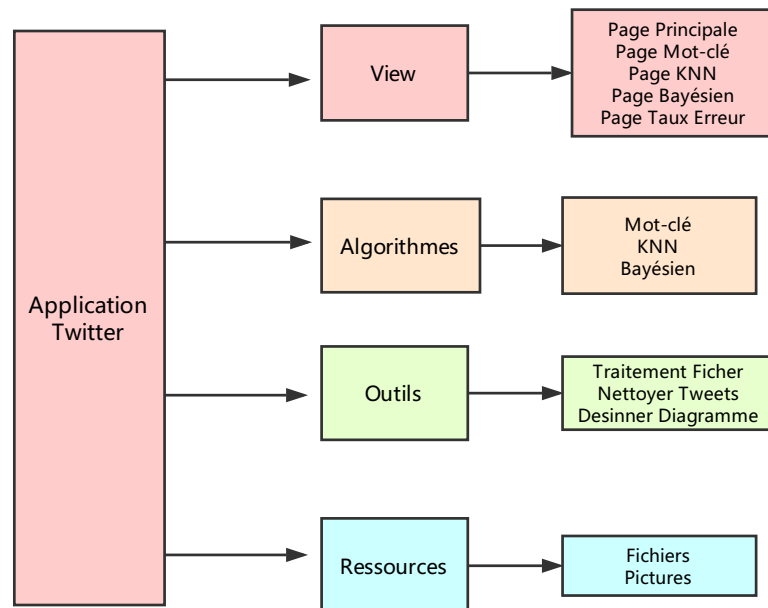
L'objectif de ce projet est d'automatiser le traitement des "Tweets" pour en tirer des informations via la création d'une application.

Pour ce faire, nous avons dans un premier temps utilisé l'API Twitter (twitter4j) pour récupérer les informations, autrement dit les tweets des utilisateurs. Dans un second temps nous avons traité les données récupérées pour faire une base d'apprentissage. Puis nous avons trié les tweets à l'aide de différents algorithmes (mots clefs, KNN et Bayésien) de classification et nous avons illustré cette classification avec des graphiques(jfreechart). Pour finir, nous avons réalisé une interface graphique à l'aide de Swing, pour rendre l'application agréable à utiliser.

2 Description générale de l'architecture de l'application

Pour réaliser notre application, nous avons choisi d'utiliser le langage Java. Nous l'avons découpé en 4 packages :

- Le package "View" qui contient les classes ayant comme rôle la création de l'interface graphique et le rendu visuel de l'application.
- Le package "Algorithmes" qui contient le code correspondant aux trois algorithmes qui permettent d'analyser la nature des tweets et d'en calculer le taux d'erreurs.
- Le package "Outils" qui contient des fonctions pour aider à la création des fonctions des autres packages ainsi que pour construire des diagrammes. Comme son nom l'indique ce package est une "boîte à outils".
- Le dossier "Ressource" qui contient les images et les fichiers utiles à la réalisation du projet.



Architecture de notre application

Chapitre 2 Détails des différents travaux réalisés

1.API Twitter

Dans ce projet, nous avons dû utiliser une interface de programmation (API) de Twitter pour obtenir les tweets par mot-clé. Pour ce faire, on s'est inscrit sur le site "<https://developer.twitter.com/>" pour obtenir le droit d'utiliser l'API. Puis on a téléchargé cet API en ligne et on l'a utilisé dans la "Library" de notre application. Dans nos codes, on a configuré les informations (API keys, API secret, access token et access token secret) et utiliser le "TwitterFactory" pour obtenir le permis d'accès aux informations de twitter. A l'aide de la classe "Query", on peut chercher les tweets récents par mot-clé, limiter la quantité de tweets d'une recherche et la langue (fr) des tweets. La classes "Status" nous permet de nous aider à obtenir toutes les informations relatives : ID du tweet, l'utilisateur, etc.

2.Préparation de la base d'apprentissage

Pour préparer la base d'apprentissage, on a récupéré 300 tweets, que l'on a classifié (positif=2, négatif=0 ou neutre=1) à la main. Puis nous avons "nettoyé" ces tweets, c'est à dire que l'on a supprimé tous les caractères inutiles (par exemple les smileys). Nous avons conservé ces tweets dans un fichier au format CSV(twitter_annotation.csv).

2.1 Nettoyage des données

Pour nettoyer les tweets, nous avons procédé en 3 étapes :

- La première pour filtrer et supprimer tous les caractères spéciaux correspondant à des smileys (ou emoji, émoticône, ...), grâce à notre fonction *"filter_contenu"* du package outils.
- La deuxième pour supprimer les liens, les re-tweets, les espaces consécutifs, les @ ainsi que les #, grâce à notre fonction *"pattern_contenu"* du package outils.
- La troisième pour supprimer les tweets doublons, en analysant lorsqu'on veut ajouter un tweet dans notre base s'il existe déjà ou non.

Une fois ces trois étapes exécutées grâce à des opérations Java sur les String, nous avons pu construire la base.

2.2 Construction de la base

Nous construisons une base dans un fichier CSV de 300 tweets après nettoyage avec différents sujet (vélo, neige, enfant, film...) pour lesquels nous avons mis l'annotation. Par exemple, dans notre page principale, si on appuie le bouton *"Ajoute-annotation"*, on peut donner manuellement les annotations pour les tweets que l'on veut ajouter dans la base des tweets.



ID	USER	DATE	CONTENU DE TOUS LES TWEETS	ANNOTATION
941118471861088	RuddyLongHa	Thu Dec 14 02	GG à qui gagne la carte d'hier!RT+Follow pour le gain du jour Les inscriptions sc	0
941118446779142	eFlechette	Thu Dec 14 02	C'est confirmé : nos bâches et nos sacs de couchage ont été détruits hier, en ple	0
941118445546016	lheb_Br	Thu Dec 14 02	rtsi ta daronn el a fini assassins creed hier soir ce matin lor de la perkiz ta été é	0
941118412511678	annedeblois	Thu Dec 14 02	Le chauffeur du parcours 18, hier, s'excusait aux gens de devoir les accueillir da	0
941118215756943	Mickacross23	Thu Dec 14 02	Comme j'ai foiré hier le kdo du nems de l'avent ! je tire 2 gagnants pour des cle	0
941118081455263	DonMateoElC	Thu Dec 14 02	Quique Guash qui annonçait hier que Griezmann ira au Barça."Si Griezmann ne s	0
941118027331981	DePanurge	Thu Dec 14 02	Hier je disais que certains médias allaient dire que si l'attentat avait eu lieu c'éta	0
941117673039192	haymella	Thu Dec 14 02	De belles couleurs hier soir sur Bergerac, avant l'arrivée de la pluie...	0
941117527257739	william99140	Thu Dec 14 02	Jurez hier o lycée ça m'a lâcher tu fais la meuf drôle sur Twitter mais vsi impro	0
941117481892139	gaelledrgSM	Thu Dec 14 02	Les meufs elles aiment trop tweeter des trucs que elles seules peuvent compren	0
941117413441048	GALATEE71	Thu Dec 14 02	Belle soirée hier pour mon 1er compte rendu de mandat. Merci aux nombreux h	0
941117394919010	TiBambino97	Thu Dec 14 02	hier Brice m'a diagnostiqué la maladie du kartel quand j'ai vu ça j'étais plié	0
941117367341518	amsa_tou	Thu Dec 14 02	Hier à 23h je dormais cause I was a bit disappointed mais tey 3h moins je suis e	0
941117147115409	MarinaPerra	Thu Dec 14 02	Guess who was in the studio with Céline last night?! Check it out on her Insta	0
941116915883347	Telokehuh	Thu Dec 14 02	New York New York - Franck Sinatrahop + mieux que hier soir et + Queen que j	0
941116818223181	Random95	Thu Dec 14 02	Vite, Mme la ministre ! ...2600 ans d'histoire sont mangés depuis hier par les tr	0
941116620604362	Taran_PigKee	Thu Dec 14 02	Victoire importante hier soir contre le PSG! Bravo à toute la team!	0
941116551629103	DavidBerard1	Thu Dec 14 02	N'oubliez pas que Dzemaili avait demandé de retourner à Bologne... demande re	0
941116512097710	Helissng	Thu Dec 14 02	Hier soir, nous avons eu le plaisir de partager notre repas de Noël en compagne	0
941116462269386	ChankleurFl	Thu Dec 14 02	Ca bouge à l'université Paul-Valéry Montpellier 3 : le CEVU a voté hier la motio	0

Exemple d'ajout d'annotations

3. Algorithme de classification

Pour analyser les tweets, nous utilisons 3 différents algorithmes de classification : la méthode dictionnaires, la méthode KNN et enfin la méthode Bayes.

Dans notre package "*algo*", il y a trois classes correspondantes aux trois algorithmes. Et sur la page principale de notre application, il y a un menu "*algo*" où on peut choisir quel algorithme de classification utiliser. Son résultat va s'afficher dans une autre fenêtre.

3.1 Mots clés

La méthode des mots clés consiste à trouver la nature d'un tweet grâce à deux fichiers de listes de mots (positifs et négatifs). On compare chaque mot des nouveaux tweets avec les deux listes de mots. Pour chaque tweet, s'il contient plus de mots positifs que négatifs, alors ce tweet est positif, et réciproquement pour les tweets négatifs. S'il contient autant de mots positifs que négatifs ou s'il ne contient pas de mots des listes, alors il est neutre.

Pour se faire, nous appelons la fonction "*analyseTheMessage()*" qui prend un tweet en paramètre et qui calcule le ratio positif/négatif du tweet par la méthode des mots clés, en comparant chaque mot de ce tweet avec nos listes de mots positifs et négatifs. Elle retourne l'annotation de ce tweet. Nous avons également créé une fonction "*algo_Mot_cle()*", qui permet d'utiliser cette méthode pour analyser les 20 tweets qui apparaissent après une recherche sur notre application. Cette fonction prend une liste de tweets en paramètre et retourne la liste des annotations. Pour afficher ces résultats dans notre application nous avons créé la classe "*Mot_cle_frame*".

3.2 KNN

L'algorithme k-nearest neighbors ou KNN (k voisins les plus proches), est un algorithme qui consiste à prendre en compte les k voisins les plus proches, de la base d'apprentissage, pour notre nouvelle entrée. On peut ainsi déterminer la valeur la plus probable pour cette entrée en fonction des k voisins. Pour notre programme, cela s'applique de la manière suivante : Pour un nouveau tweet, on calcule sa distance (nombre de mots en communs) avec tous les tweets de notre base d'apprentissage.

Pour se faire, nous appelons la fonction "*distance(t1,t2)*" pour calculer la distance entre deux tweets ($Distance(t1, t2) = (nombre_total_de_mots - nombre_de_mots_commun) / nombre_total_de_mots$), et la fonction "*Tous_les_distances(t, liste_t)*" qui prend un tweet en paramètre et calcule sa distance avec tous les tweets dans la base d'apprentissage. Puis on choisit les 5 tweets de la base les plus proches du nouveau et on annote le nouveau tweet en fonction de la majorité d'annotation (s'il y a plus de tweet positif dans notre échantillon de 5 tweets alors notre nouveau tweet est positif, même raisonnement pour négatif ou neutre). Nous avons également créé une fonction "*algo_knn(liste_nouveaux_tweets, liste_base_tweets)*", qui permet d'utiliser cette

méthode pour analyser les 20 tweets qui apparaissent après une recherche de tweets sur notre application. Cette fonction prend une liste de tweets en paramètre et retourne la liste des annotations. Pour afficher ces résultats dans notre application nous avons créé la classe *"Knn_frame"*.

3.3 Bayes

3.3.1 Représentation de présence

L'algorithme de classification bayésienne est un outil de classification qui s'exprime par : $\prod_{m \in t} p(m|c) * p(c)$ ($p(m|c) = n(m,c)/n(c)$) avec, dans notre cas, c une classe (positif, négatif ou neutre), m un mot, $n(c)$ le nombre total de tweets de la classe c , $n(m, c)$ le nombre tweets qui contient le mot m dans la classe c et $P(m|c)$ la probabilité de m sachant c et $p(c)$ le ratio de tweets de classe c dans tous les tweets de la base.

Pour connaître la classe c d'un tweet, on le divise en mot dans un premier temps, dans un second temps on cherche le nombre d'occurrence des tweets qui contiennent ce mot m . Puis on multiplie les occurrences de tous les mots par la proportionnalité de tweets de classe c de la base de tweets.

Si le mot n'apparaît jamais dans la classe c , on remplace le résultat 0 par $(p(m|n)+1)/n(c)+N$. Ensuite on multiplie le résultat avec $p(c)$, pour obtenir la probabilité que ce tweet apparaisse dans la classe c . En Comparant avec les probabilités des trois classes, on peut annoter ce tweet par la classe qui a la probabilité la plus grande.

Pour se faire, nous appelons la fonction *"AlgoUnigramme (String mot, List<Bayes> classes)"* qui calcule la probabilité de $p(m|c)$, et la fonction *"AlgoUnigramme_res(String[] mots_twitter)"* qui calcule la probabilité de $p(c|t)$. Nous avons également créé une fonction *"algoBayes(String moodle, Bayes b)"*, qui permet d'utiliser cette méthode pour analyser les 20 tweets qui apparaissent après une recherche de tweets sur notre application. Cette fonction prend une liste de tweets en paramètre et retourne la liste des annotations. Pour afficher ces résultats dans notre application nous avons créé la classe *"Bayes_frame"*.

3.3.2 Représentation de fréquence

Dans ce cas, on prend en compte le nombre de mots m dans le nouveau tweet t (n_m), on met la probabilité $P(m|c)$ à la puissance (n_m). Et on crée aussi une fonction pour nettoyer les tweets des mots de moins de trois lettres qu'on appelle *"filter_moins_de_trois_lettre(Bayes bn)"*.

3.3.3 Ensembles de mots

Parce que certaines combinaisons de mots peuvent avoir de l'importance, on utilise trois façons de comparer les mots.

● Uni-Gramme

Le uni-gramme est défini par un mot simple, par exemple la phrase : *j'aime la pomme* contient 3 uni-grammes : *j'aime*, *la*, *pomme*.

● Bi-Gramme

Le bi-gramme est considéré deux mots consécutifs, par exemple la phrase : *j'aime la pomme*, contient 2 bi-grammes: (*j'aime la*), (*la pomme*).

● Uni-Bi-Gramme

Le uni-bi-gramme est la combinaison des uni-grammes + bi-grammes, par exemple la phrase : *j'aime la pomme*. contient 5 uni-bi-grammes: *j'aime*, *la*, *pomme*, (*j'aime la*), (*la pomme*).


4.Interface graphique

Notre interface graphique se compose d'un écran principal depuis lequel on peut accéder à plusieurs autres écrans reflétant des fonctionnalités que l'on a implémenté :

4.1 Copies d'écran

Vous trouverez ci-dessous l'écran principal de notre application. Il a 3 parties :

- Menu : permettant de choisir différente fonction
- Champ de recherche : ou on peut entrer le mot-clé que l'on veut chercher.
- Tableau d'affichage : afficher des tweets séparés en 6 parties : ID, nom de d'utilisateur, date, contenu de tweets



ID	USER	DATE	CONTENU DE TOUS LES TWEETS
941479548738310	Football_life_	Fri Dec 15 02:	Pour les hors jeux et les pénalty
941479388704419	LalaanFeler	Fri Dec 15 02:	VIDEO. Quand Dijon charrie Yambéré et ses ballons dans le ruisseau
941479007895326	Daniel_Jerom	Fri Dec 15 02:	Crise du football Malien : Création d'un collectif pour réconcilier les acteurs du
941478967613280	PecheXpress	Fri Dec 15 02:	Decouvrez: Tête Plombée Football Jigs anti herbes Delalande
941478694962348	123_INFO_SP	Fri Dec 15 02:	Barça: Dembélé, le retour approche
941478451093045	Whats_upHon	Fri Dec 15 02:	Le maillot de pour sauver le club du Grau-du-RoiLes explications
941478414342606	HamzaKesk	Fri Dec 15 02:	La LFP adopte l'arbitrage vidéo à partir de la saison prochaine
941478305651359	mehdii_94	Fri Dec 15 02:	Leo Messi a livré une interview à la FIFA. La voici en intégralité ! Ses adversaire
941478241306636	OLjerome	Fri Dec 15 02:	"Aulas, 30 ans d'histoire(s)" réalisé par Étienne Pidoux. Retour sur le parcours d
941478203134275	thomas_suqu	Fri Dec 15 02:	ÉON. Boy de la Tour (présidente) : "Le métier de nos arbitres est extrêmement c
941478125375954	rokhesportm	Fri Dec 15 02:	La Pro League lancera son championnat e-sport en 2018
941478113049042	TedBlakeOff	Fri Dec 15 02:	l'organisme en charge des lois du jeu (IFAB) valide la décision de la en mars pro
941478040399417	TedBlakeOff	Fri Dec 15 02:	Platini, père de Michel Platini et ancien directeur sportif de l', est décédé ce jeu
941478013434257	TedBlakeOff	Fri Dec 15 02:	face à Amiens dimanche, l'entra neur de l'Bruno Gènesio a été suspendu un ma
941477997860806	TedBlakeOff	Fri Dec 15 02:	sanction prend effet le mardi 19 décembreBruno Gènesio sera donc bien présen
941477964016963	onlinelisting	Fri Dec 15 02:	BILLET NANTES france v SION suisse 06/12/1994 football uefa cup ticket: 12,00
941477940482670	Merlus_Addic	Fri Dec 15 02:	FC . Lemoine enfin de retour, pas Guendouzi Le Telegramme
941477939438288	onlinelisting	Fri Dec 15 02:	BILLET NANTES france v PARTIZAN BELGRADE 06/11/1985 football uefa cup tick
941477812015427	KJuliye	Fri Dec 15 02:	Didier Drogba: Il faut essayer de rassembler le football ivoirien
941477773155164	OMaanZa	Fri Dec 15 02:	T'es parano toi ptdr, tu me bloques, tu me débloques. Point de vue différent , s

Page principale

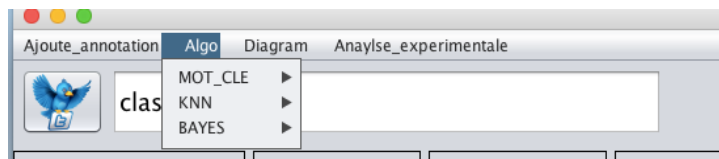
De plus, nous créons des sous-fenêtres contrôlées par les menus de la page principale.

1.Ajoute-annotation (On peut donner les annotations aux nouveaux tweets à la main)

ID	USER	DATE	CONTENU DE TOUS LES TWEETS	ANNOTATION
941502109459140	michelhersir	Fri Dec 15 03:	Browns (1e) et Giants (2e), c'est presque assuré. Après ça deux équipes à 3 victoires	1
941501060316876	Houaria47	Fri Dec 15 03:	A peine élu président des ripoublicains, Laurent Wauquiez s'empresse de donner	1
941493226241970	asnyki	Fri Dec 15 03:	: De plus, les volumes 19 et 20 et Sword Art Online sont classés respectivement	2
941492583783653	luttesdemars	Fri Dec 15 03:	Scandale : risques majeurs pr la des gens, l'Etat qui vend à 1 site du CEA ds le	null
941492362156564	Nath_Lar	Fri Dec 15 03:	Passionnés de sciences, à vos agendas! Vendredi 15 décembre de 12h30 à 13h	0
941489128792903	escobarandw	Fri Dec 15 03:	Bravo à toi, tu te classes parmi la 1ère place au classement des pigeons intersic	1
941486878984896	Jensen2018	Fri Dec 15 02:	La société est divisée en deux classes : ceux qui ont plus de d ners que d'appét	2
941485389755318	ptiyo38	Fri Dec 15 02:	C'est ce que défendent les marxistes, la lutte des classes prime. Et quand j'ai vé	null
941483632807628	couteronjp	Fri Dec 15 02:	Voici pourquoi je n'aimerai pas que l'on interdise les nouveaux outils de	null
941483017905852	wezeau1	Fri Dec 15 02:	La GRC reverra 284 dossiers d'agression sexuelle qui avaient été classés en 201	null
941481566487924	Psimoneau73	Fri Dec 15 02:	WoW quel classes ... tu va nous manqué !!! Bon succès ailleurs mais reviens vite	null
941480050083401	strauzzlane	Fri Dec 15 02:	du "schmidtisme" de gauche en somme... voilà qu'on voit poindre à gauche con	null
941478908410638	GrosCorpsSoc	Fri Dec 15 02:	Je partageais votre opinion en début d'année dernière. Force est de constater qu	null
941478398941069	CGPMFO	Fri Dec 15 02:	3 classes issues des collègues Matisse et Jean-Baptiste Clément (20e) et du lycée	null
941477430891597	In123Champa	Fri Dec 15 02:	illégaux dans des classes : sont elles cautionnées par l'état ?	null
941475038427041	Asnounjo	Fri Dec 15 02:	Voilà ce qui nous attend comme réjouissances avec les nouveaux, sachez-le!!!	null
941473016885796	handiseniors5	Fri Dec 15 02:	Pour Laurent Wauquiez, le problème de la cité, c'est "l'impossibilité d'enseign	null
941472341711818	tmo1777	Fri Dec 15 01:	oui, ça se passe comme ça à .. et ça construit des immeubles mais il n'y a plus	null
941468809679867	pol_quebec	Fri Dec 15 01:	La GRC reverra 284 dossiers d'agression sexuelle classés en 2016	null
941468787781578	z_ariu	Fri Dec 15 01:	Le prêt à taux zéro devait s'arrêter au 31/12 Dans le , les députés ont décid	null

Ajouter les annotations

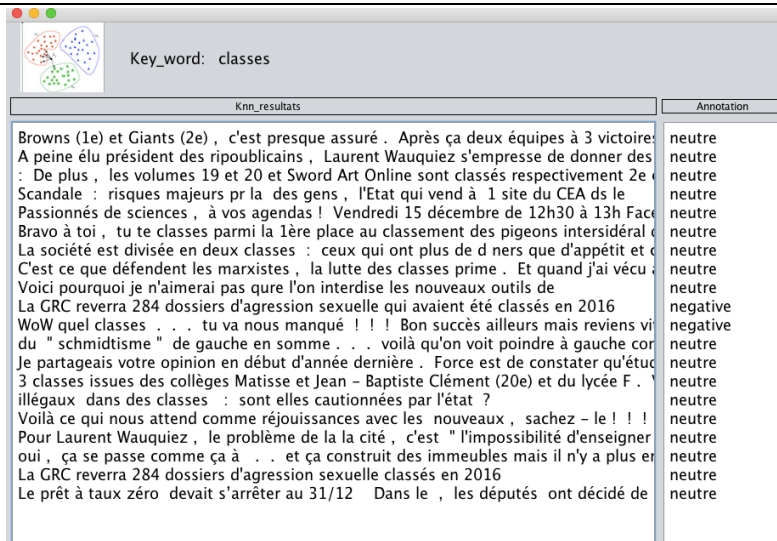
2.Algo (Utiliser différents algorithmes pour annoter les nouveaux 20 tweets)



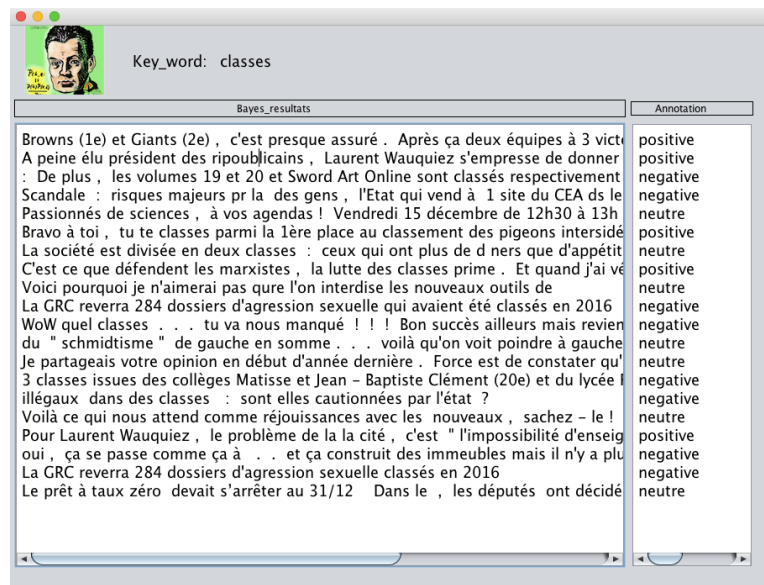
Le menu des algorithmes

Key_word: classes	
mot_cle_resultats	Annotation
Browns (1e) et Giants (2e), c'est presque assuré. Après ça deux équipes à 3 victoires	neutre
A peine élu président des ripoublicains, Laurent Wauquiez s'empresse de donner	neutre
: De plus, les volumes 19 et 20 et Sword Art Online sont classés respectivement	positive
Scandale : risques majeurs pr la des gens, l'Etat qui vend à 1 site du CEA ds le	neutre
Passionnés de sciences, à vos agendas! Vendredi 15 décembre de 12h30 à 13h	neutre
Bravo à toi, tu te classes parmi la 1ère place au classement des pigeons intersic	positive
La société est divisée en deux classes : ceux qui ont plus de d ners que d'appét	positive
C'est ce que défendent les marxistes, la lutte des classes prime. Et quand j'ai vé	positive
Voici pourquoi je n'aimerai pas que l'on interdise les nouveaux outils de	neutre
La GRC reverra 284 dossiers d'agression sexuelle qui avaient été classés en 201	neutre
WoW quel classes ... tu va nous manqué !!! Bon succès ailleurs mais reviens vite	neutre
du "schmidtisme" de gauche en somme... voilà qu'on voit poindre à gauche con	neutre
Je partageais votre opinion en début d'année dernière. Force est de constater qu	neutre
3 classes issues des collègues Matisse et Jean - Baptiste Clément (20e) et du l	neutre
illégaux dans des classes : sont elles cautionnées par l'état ?	neutre
Voilà ce qui nous attend comme réjouissances avec les nouveaux, sachez-le!!!	neutre
Pour Laurent Wauquiez, le problème de la cité, c'est "l'impossibilité d'enseign	neutre
oui, ça se passe comme ça à .. et ça construit des immeubles mais il n'y a plus	positive
La GRC reverra 284 dossiers d'agression sexuelle classés en 2016	neutre
Le prêt à taux zéro devait s'arrêter au 31/12 Dans le , les députés ont d	positive

Mot-clé

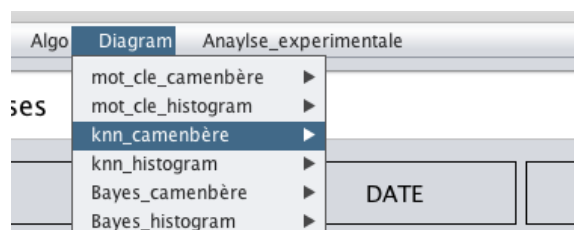


Knn

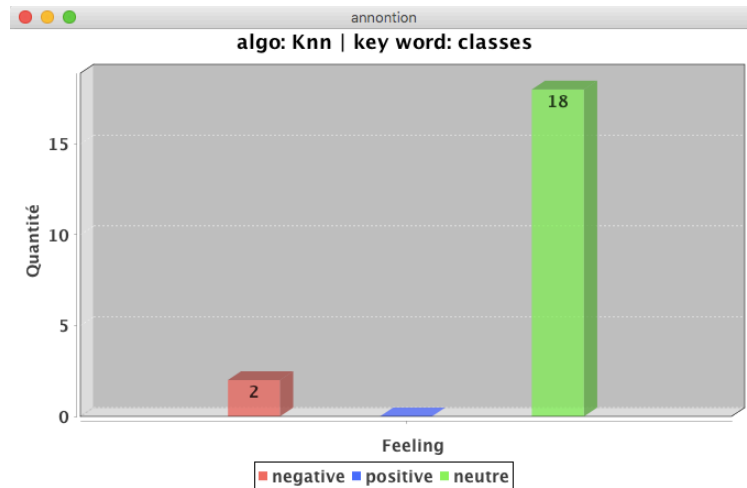


Bayes

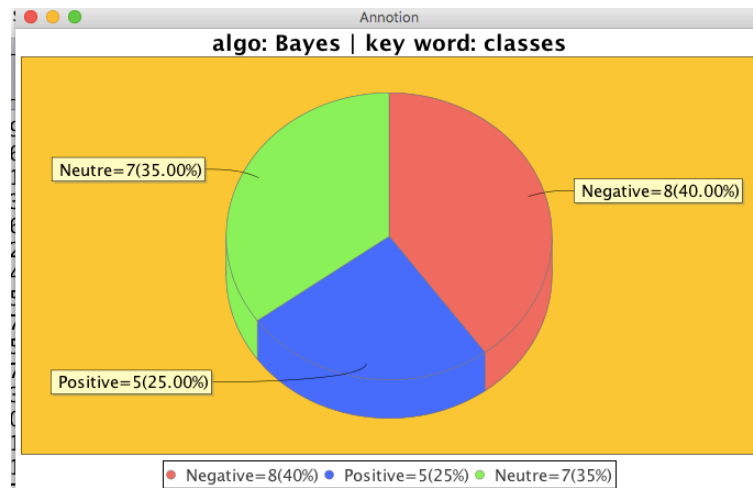
3. Diagram (Permet de mettre en valeur les résultats sous forme de diagrammes)



Le menu de diagrammes



L'histogramme

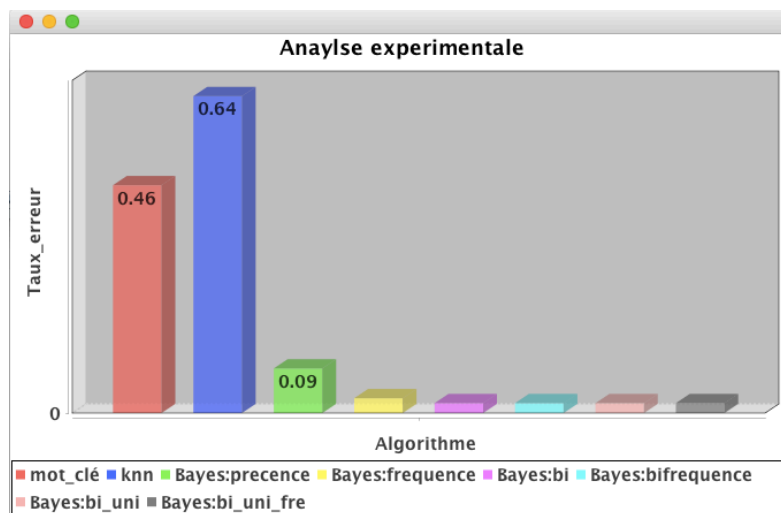


Le camembert

4. Analyse_experimentale (Permet d'afficher les taux d'erreurs de tous les algorithmes)

Algorithme	Taux_erreur	Nombre_erreur
mot_clé	0.46	137
knn	0.64	191
Bayes:precence	0.09	27
Bayes:frequence	0.03	8
Bayes:bi	0.02	6
Bayes:bifrequence	0.02	6
Bayes:bi_uni	0.02	5
Bayes:bi_uni_fre	0.02	6

Le taux d'erreurs de tous les algorithmes



Le diagramme d'erreurs

4.2 Manuel d'utilisation

4.2.1 Exécuter l'application :

Pour exécuter l'application il suffit de double cliquer sur "*pjb_twitter.jar*". L'environnement de java doit être java 7 ou supérieur.

4.2.2 Chercher un mot clef :

Entrez un mot clef dans la barre de recherche, et puis cliquez sur le bouton avec le petit oiseau pour rechercher les tweets. Le résultat des tweets correspondant sera affiché.

4.2.3 Faire une annotation manuelle et l'ajouter dans la base des tweets :

Cliquer sur l'onglet "Ajoute_annotation", dans la barre de menu, pour annoter manuellement les tweets et cliquer sur "OK" pour valider. Les tweets avec annotations manuels seront sauvegardés dans le fichier "twitter_annotation.csv".

4.2.4 Faire une annotation automatique via différents algorithmes :

Cliquer sur l'onglet "Algo ", dans la barre de menu, pour ouvrir le menu déroulant permettant de sélectionner l'algorithme avec lequel on souhaite faire l'annotation automatique. Une fois choisi, le programme ouvre une nouvelle fenêtre affichant les résultats

4.2.5 Faire un diagramme automatique via différents algorithmes :

Cliquer sur l'onglet "Diagram ", dans la barre de menu, pour ouvrir le menu déroulant permettant de sélectionner l'algorithmme avec lequel on souhaite faire dessiner son digramme ainsi que le style de diagramme (camembert, histogramme).

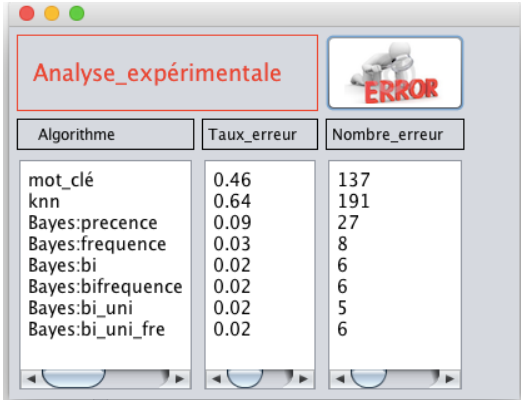
Une fois choisi, le programme ouvre une nouvelle fenêtre affichant le diagramme.

4.2.6 Calculer les taux d'erreurs de tous les algorithmes :

Cliquer sur l'onglet "Analyse expérimentale ", dans la barre de menu, pour ouvrir le menu analyse expérimentale permettant de voir les taux d'erreurs des différents algorithmes.

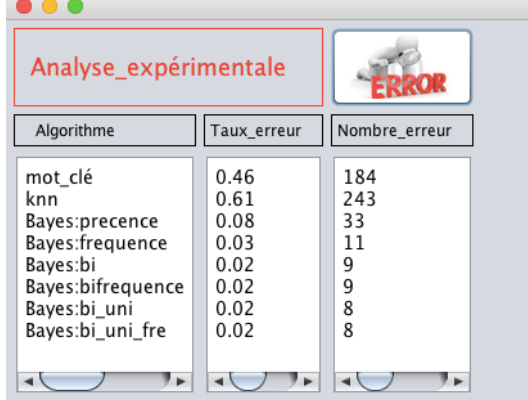
Dans ce menu double cliqué sur le bouton "error" pour afficher les statistiques sous forme de diagramme.

Chapitre 3 Résultats de la classification avec le différentes méthodes et analyse



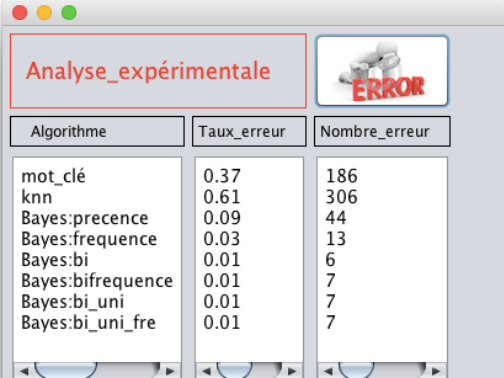
Algorithme	Taux_erreur	Nombre_erreur
mot_clé	0.46	137
knn	0.64	191
Bayes:precence	0.09	27
Bayes:frequence	0.03	8
Bayes:bi	0.02	6
Bayes:bifrequence	0.02	6
Bayes:bi_uni	0.02	5
Bayes:bi_uni_fre	0.02	6

300 tweets



Algorithme	Taux_erreur	Nombre_erreur
mot_clé	0.46	184
knn	0.61	243
Bayes:precence	0.08	33
Bayes:frequence	0.03	11
Bayes:bi	0.02	9
Bayes:bifrequence	0.02	9
Bayes:bi_uni	0.02	8
Bayes:bi_uni_fre	0.02	8

400 tweets



Algorithme	Taux_erreur	Nombre_erreur
mot_clé	0.37	186
knn	0.61	306
Bayes:precence	0.09	44
Bayes:frequence	0.03	13
Bayes:bi	0.01	6
Bayes:bifrequence	0.01	7
Bayes:bi_uni	0.01	7
Bayes:bi_uni_fre	0.01	7

500 tweets

Dans notre base de tweets, il y a 3 situations ,300 tweets,400 tweets et 500 tweets. On calcule le taux d'erreurs par l'analyse expérimentale, d'après nos résultats, pour :

La méthode mot clé, il y a un taux d'erreurs d'environ 40%.

La méthode Knn, il y a un taux d'erreurs d'environ 60%.

Les méthodes bayésiennes sur la présence en unigramme a un taux d'erreurs d'environ 9%.

Les méthodes bayésiennes sur la fréquence en unigramme a un taux d'erreurs d'environ 3%.

Les méthodes bayésiennes sur la présence en bigrammes a un taux d'erreurs d'environ 2%.

Les méthodes bayésiennes sur la fréquence en bigrammes a un taux d'erreurs d'environ 2%.

Les méthodes bayésiennes sur la présence en uni+bi a un taux d'erreurs d'environ 2%.

Les méthodes bayésiennes sur la fréquence en uni+bi a 5 erreurs, le taux d'erreur est 2%.

Grâce à ces résultats, on peut en déduire que les méthodes de Bayes sont les plus efficaces et parmi les méthodes de Bayes, les méthodes en uni+bi sont les meilleurs.

Chapitre 4 Conclusions

Ce projet nous a permis de concevoir une application fonctionnelle, qui permet d'analyser des tweets en fonction d'un mot clé, de les trier automatiquement par leur nature (positive, négative, neutre) via différents algorithmes (knn, bayes, mots clés), de représenter les résultats sous forme de graphes et de calculer les taux d'erreurs des différents algorithmes pour vérifier la fiabilité des résultats.

Dans ce projet, nous avons appris à utiliser des APIs pour obtenir leurs fonctionnalités et nous en servir pour notre projet. De plus nous avons appris de nouveaux algorithmes de classification. Nous avons utilisé NetBean et swing pour la première fois, nous permettant de créer des interfaces graphiques facilement. La méthode de travail était agréable, en effet nous avons apprécié l'autonomie de travail tout en aillant des professeurs à l'écoute pour nous guider quand nous en avions besoin.

Cependant nous avons rencontré de nombreux problèmes, au niveau du design de l'application, en effet nous n'étions jamais vraiment satisfaites de notre interface. De plus nous avons eu des difficultés à retranscrire les algorithmes de classification sous forme de code.

Pour finir, je pense que l'on pourrait améliorer notre application, en permettant d'analyser tous les tweets postés par un utilisateur, pour analyser son avis général. Cela permettrait aussi de connaître l'évolution de ses pensées sur différents sujets.