

Rapport PJE

Analyse de Comportements avec Twitter



Promotion : 2015/2016

Lien GITHUB : <https://github.com/m1TP/pje>

Christopher DESHAIES

Master Informatique Groupe 2

Sommaire

Analyse de Comportements avec Twitter	1
1. Description générale du projet	3
1.1. Description de la problématique.....	3
1.2. Description générale de l'architecture de votre application	3
2. Détails des différents travaux réalisés	5
2.1. API Twitter	5
2.2. Base d'apprentissage.....	5
2.3. Algorithme de classification	7
2.4. Interface graphique	9
3. Résultats	17
4. Conclusions.....	18

1. Description générale du projet

1.1. Description de la problématique

Le but général de ce PJE était de réaliser une application qui porte sur l'analyse de comportements sur Twitter, en particulier d'analyser les sentiments des différents tweets qui auront été récupérés.

Le traitement automatique des tweets est au cœur de ce projet. Les messages qui circulent sur des sites de type réseaux sociaux sont des mines d'informations qui peuvent être utilisés pour réaliser des études sur les sentiments généraux des utilisateurs vis-à-vis d'une idée, concept, personne politique, etc.

Donc ce projet a pour but la réalisation d'une application qui :

- Utilisera l'API Twitter pour récolter de l'information sur twitter.
- Une interface utilisateur réalisée en Java.
- Apprentissage de modèles et outils algorithmique d'analyse pour traiter les tweets

Le projet aura été réalisé durant 24h de TD et 24h de TP entouré par une équipe d'intervenants : Bilel Derbel responsable du projet, Laetitia Jourdan et Arnaud Liefoghe intervenants.

1.2. Description générale de l'architecture de votre application

L'application a été réalisée en Java. La réalisation de l'application a été faite dans un environnement JavaSE-1.7. La librairie Twitter4j a été utilisée pour la réalisation de l'application.

Il comprend 5 dossiers : « bin », « db », « img », « src ». Dans le dossier « db » on y retrouve des fichiers csv contenant les données sur les tweets sauvegardés en base de données, dans le dossier « img » on y retrouve les images qui ont été ajoutées à l'interface de l'application et le dossier « src » on y trouve le code de l'application.

Le code de l'application est divisé en 8 packages : « control », « csv », « graph », « main », « methode », « model », « util » et « view ». Dans le package « control » on y retrouve des ActionListener nécessaires au fonctionnement des interfaces, dans le package « csv » on y retrouve le code nécessaire au traitement des fichiers csv dans le dossier « db », le package « graph » contient le code nécessaire à l'édition de graph pour analyser les tweets, « main » contenant la classe de lancement de l'application, « methode » contenant toutes les méthodes de lecture et classification de tweets, « model » contenant la structure de base d'un tweet, « util » contenant des classes nécessaires aux méthodes de classification puis dans le package « view » on y retrouve les différentes interfaces.



Packaging de l'application

2. Détails des différents travaux réalisés

2.1. API Twitter

L'API Twitter permet d'accéder à la base de données de twitter et de récupérer/poster plusieurs informations. Durant ce projet, j'ai principalement utilisé les classes SEARCH et REST de l'API.

Pour pouvoir utiliser l'application, il a fallu pouvoir s'authentifier en utilisant de protocole OAuth. Une fois la création d'un compte Twitter effectuer, et avoir déclaré dans l'espace développeurs du site la création d'une application.

La librairie utiliser dans ce projet a été la librairie Java twiter4j. La documentation associée à la librairie est accessible sur : <http://twitter4j.org/>

La compréhension de l'utilisation de l'API Twitter ne fut pas très compliquée. Il a fallu tenir compte pour le développement de certaines restrictions par l'application Twitter comme un problème de jetons disponibles pour interroger Twitter ou encore qu'il ne soit pas possible d'obtenir plus de 100 tweets avec une seule requête.

Cette librairie a permis la réalisation de requêtes Java entre notre application et twitter principalement pour récupérer diverses informations concernant des tweets afin d'élaborer nos bases de données.

2.2. Base d'apprentissage

Pour la préparation de la base d'apprentissage, j'ai commencé par construire un premier fichier CSV avec ses informations sauvegardés :

- L'id du tweet
- Le pseudonyme du titulaire du tweet
- Le texte
- La date de publication
- Le sujet de la recherche
- L'annotation

Chaque bases d'apprentissage sont enregistrés par sujet sauf pour le fichier « all.csv » qui regroupe tous les tweets des différents sujets.

Pour la réalisation d'une base d'apprentissage, il a fallu effectuer un nettoyage de la base de donnée pour enlever certains caractères spéciaux comme : @, #, RT, les liens URL, retour à la

ligne ou entre multiples espaces effectuer grâce à des fonctions utilisant des expressions régulières grâce aux classes Java : Pattern et Matcher.

Il a s'agit également d'enlever les tweets qui comportent des émoticônes à la fois positifs et négatifs en même temps, d'enlever les tweets redondants en base (principalement utile à partir de la classification KNN) et récupérer principalement des tweets français.

```

5 666728007441055000;LyonOnline;La Marseillaise a été reprise en chœur par les spectateurs présents au stade de Wembley , mardi à Londres . . . URL # ;Tue Nov 17 22:21:58 CET
6 666724125986148000;Antonia Renna;L onobgs london # ;Tue Nov 17 22:06:33 CET 2015;london;1
7 666715046127472000;Ryvers;Quand t'arrive à Londres t'entend "" ça va péter distance de sécurité "" ;Tue Nov 17 21:30:28 CET 2015;london;0
8 66670784059836000;Aurélien Btd;Minutes de silence à Londres ( match ) ;Tue Nov 17 21:02:01 CET 2015;london;1
9 666707246961790000;querricr;très belle marseillaise à l'unison au stade de Londres # ;Tue Nov 17 20:59:29 CET 2015;london;2
10 666696970464067000;sevil;Oh putain , je crois que je peux pas partir à Londres . . ;Tue Nov 17 20:18:39 CET 2015;london;0
11 666695390536802000;????????;Je part à Londres faire les magasins ta oublier ?? ducou sa me soûl je peut pas y aller ;Tue Nov 17 20:12:22 CET 2015;london;0
12 66669400123946000;l'ermite pas net ??;Londres c'est le feu ??? ;Tue Nov 17 20:06:51 CET 2015;london;2
13 666687104710254000;6kiceTea;ma cousine de Londres elle vient d'connaitre niska grace a son gars mtn elle ecoute que ça jui mooooorte ??? ;Tue Nov 17 19:39:26 CET 2015;london;2
14 666686902804836000;????;Ananas à Londres ? . . ;Tue Nov 17 19:38:38 CET 2015;london;1
15 666686325903462000;????;J'ai envie d'aller vivre à Londres . . . Finalement enft là-bas c'était le feu ! Juste les gens me manquait fort ;Tue Nov 17 19:36:21 CET 2015;london;0
16 666683984634585000;TISMÉ PUR SANG;Prochainement à tllmt vrai : Adèle 15 ans décide de tt quitter pour vivre l'amour à Londres avec un rappeur français URL ;Tue Nov 17 19:27:02
17 666671401840082000;kranttshepard;@ j ai libéré tout london plus envie pour 1 instant de mon côté ;Tue Nov 17 18:37:03 CET 2015;london;0
18 666668808657092000;gauthier;en avril je pars 4jours à Londres et à Torquay avec Marine , Clara , Yasmine , @ et tout? ;Tue Nov 17 18:26:44 CET 2015;london;1
19 666666106891968000;FRL;""D'après une histoire vraie"" de # 8 danseurs réinterprétant les # de guerres turques URL # ;Tue Nov 17 18:16:00 CET 2015;london;1
20 6666658672580419000;Edouar.S;J'aime une vidéo @ de @ Assassins'a Creed Syndicate : Tous les Secrets De Londres + n° de boîte ;Tue Nov 17 17:46:28 CET 2015;london;2
21 671298479755522000;- 19 FMR?;@ Ouais ouais Londres Nous voilà ?? ;Mon Nov 30 13:03:24 CET 2015;london;2
22 671293987312025000;?;il est pas à Londres ? ;Mon Nov 30 12:45:33 CET 2015;london;0
23 671258534961676000;Olivier Pope;Quelqu'un est déjà allé au Roundhouse à Londres ? ;Mon Nov 30 10:24:40 CET 2015;london;1
24 671230192797185000;4soliderLikeMeFaps ;8 sans oublier notre Manager @ notre FR depuis Londres @ notre Boss @ et God Almighty . B blessed ;Mon Nov 30 09:51:31 CET 2015;london;1
25 6712177930276745000;Shooba;8 Direction Londres URL ;Mon Nov 30 07:42:32 CET 2015;london;1
26 671315493710528000;Jack Bauer;8 putain je suis pas le seul à dire ça mes ces pas un truc de noir tu devrait venir à Londres loool ;Mon Nov 30 14:11:00 CET 2015;london;0
27 671362746345259000;m+m?;J'espère pouvoir partir à Londres dans 2semaines ??? ;Mon Nov 30 17:18:46 CET 2015;london;2
28 67134898878664000;Xadyn;8 euh . . . à Londres en Angleterre plutôt non ? 8-o ;Mon Nov 30 16:24:06 CET 2015;london;1

```

Figure : Exemple de Tweet en base

Les différents tweets contenu dans les bases de données sont annotés soit positif (2), soit neutre (1), négatif(0) ou indéfinie (-1) selon la méthode de classification choisi.

Les données des divers tweets en base d'apprentissage seront classifiées et analysées grâce à différente méthode :

- Manuellement
- Avec des mots clés
- KNN
- Bayésienne

2.3. Algorithme de classification

2.3.1. Mots clefs

A partir d'une liste de tweets récupéré non annoté suite à une recherche, on a la possibilité d'utiliser une première méthode simpliste de vérifications des divers mots des tweets pour vérifier si le mot fait partie du corpus positif ou négatif et ainsi si un tweet contient plus de mot positif que négatif alors il sera classifié positif et inversement compris. Si le nombre de mot positif et négatif sont équivalents alors on considère que le tweet sera classifié neutre.

On a utilisé la grammaire sur la polarité dans notre fichier csv : 0 pour négatif, 1 pour neutre, 2 pour positif. Suite à l'annotation des tweets grâce au mot-clé, il n'y aura pas en base de tweet indiqué à -1 (indéterminé).

Principaux intérêts de la méthode de classifications :

- Méthode simple
- Non dépendant des autres tweets

Principaux inconvénients de la méthode :

- Demande un ajustement de la liste des corpus négatif et positif pour un meilleur résultat
- Classification naïve

2.3.2. KNN

La méthode de classifications KNN (k-nearest neighbor) ou des K plus proches voisins permet de classer des tweets selon une base d'apprentissage existante et d'évaluer la qualité de la base d'apprentissage. Pour la réalisation de cette méthode il a fallu réaliser un algorithme Java qui traduit l'algorithme suivant :

```
Données : x le tweet à étiqueter, k le nombre de voisins
pour i allant de 1 à k
    mettre le point i dans proches_voisins
fin pour
pour i allant de k+1 à N
    si la distance entre i et x est inférieure à la distance d'un des
    points de proches_voisins à x
        supprimer de proches_voisins le point le plus éloigné de x
        mettre dans proches_voisins le point i
    fin si
fin pour
proches_voisins contient les k plus proches voisins de x
vote(proches_voisins) donnent la classe majoritaire des voisins
```

Figure 1 : Algorithme KNN

Pour évaluer un tweet, par défaut on prend 1/3 de la base d'apprentissage pour la recherche des plus proches voisins avec la possibilité via l'interface de pouvoir paramétrer la recherche (définir manuellement combien de k voisins on va prendre en compte pour la classification).

L'évaluation de la qualité de la base consiste à évaluer notre base d'apprentissage en évaluant chaque tweets en base en vérifiant la polarité des tweets en fonction des K plus proches voisins.

Et pour la vérification de la qualité de la base on effectuera un croisement des classifications estimés par KNN avec les classifications réel déterminer par la recherche de mots clés.

L'algorithme de vérification demande une base d'apprentissage contenant beaucoup de tweet pour commencer à être fiable (environ 1000 tweet pour commencer à être fiable).

2.3.3. Bayes

La classification naïve bayésienne est un type de classification Bayésienne probabilité simple. Pour cette classification, on s'est rapproché du modèle de classification de textes.

Il suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques. Pour chaque mot d'un tweet on va estimer la probabilité de le trouver dans des tweets de même classe (même polarité) puis on estime la polarité du tweet en fonction des probabilités obtenu par chaque mots dans le tweet s'il a plus de chance d'être positif, négatif ou neutre.

Des paramètres ont été ajoutés pour la méthode de classification comme la possibilité de ne pas prendre en compte les mots de moins de 3 lettres ou de prendre en compte les bi grammes ou trigrammes dans l'algorithme de classification.

L'évaluation de la base d'apprentissage consiste à une réévaluation de la base d'apprentissage via la classification Bayésienne afin de déterminé le taux d'erreur de l'algorithme.

2.4. Interface graphique

2.4.1. Copie d'écran

Figure 2 : Interface principale au démarrage

Tweets obtenu

Options		
Alias	Tweet	Annotation
Djibril R	On a donc perdu contre :ParisMarseilleLyonMonacoCaenLille(Et Angers dans 1 semaine)#EquipeDeBabτουςFragiles	Neutre
Actualité 24	Ligue 1 : Lille confirme, Montpellier respire, Guingamp coule. https://t.co/qRgFDRban	Neutre
Flood lord	Régionales 2015: Une victoire de Marine Le Pen serait un tremplin pour 2017 https://t.co/VmhWPh5yyd https://t.co/uNzXAaGbg9 #infloodwetr...	Neutre
Flood lord	Régionales 2015: Xavier Bertrand espère faire barrage à Marine Le Pen https://t.co/okVWz7jrBk https://t.co/yojGa78xb8 #infloodwetrust	Neutre
Jane	A Lille je lui donne 1 semaine après il se fait tabasser c'est sûr https://t.co/BE2raSbMwr	Neutre
20 Minutes	Régionales 2015: Une victoire de Marine Le Pen serait un tremplin pour 2017 https://t.co/WYH5VYMiMB https://t.co/KcJmZUpkDr	Neutre
20 Minutes	Régionales 2015: Xavier Bertrand espère faire barrage à Marine Le Pen https://t.co/Kol09J4BJT https://t.co/Bjvuwk7cQ2	Neutre
Médias Fran...	Dans #7a8Les tortues Golfina au MexiqueLa mort de Julie MartinLes boutiques du Vieux Lille https://t.co/Ct6lZbwI5H	Neutre
PlaceToBe	L'élection #MissFrance 2016 au Zénith de #Lille. Retour sur une très belle ville. https://t.co/ij8jBtkDaU https://t.co/pKXqOZ1yIC	Positif
Arnaud GI	Devoir du citoyen accompli! #nord #electionsregionales #Regionales2015 #Lille #regions https://t.co/gW15eDANIY	Neutre

SAUVEGARDER TWEETS

Figure 3 : Recherche de Tweet en fonction de mot-clé et nombre de tweet souhaité

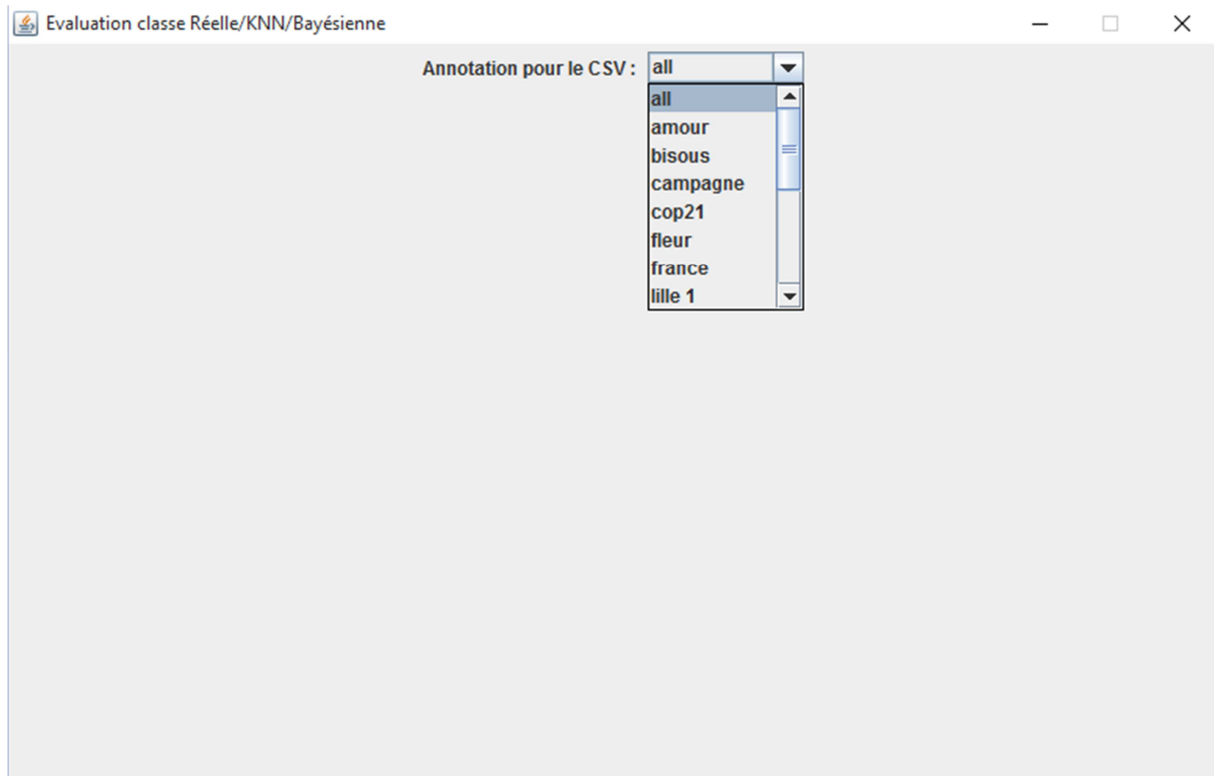


Figure 4 : Ecran de sélection de csv

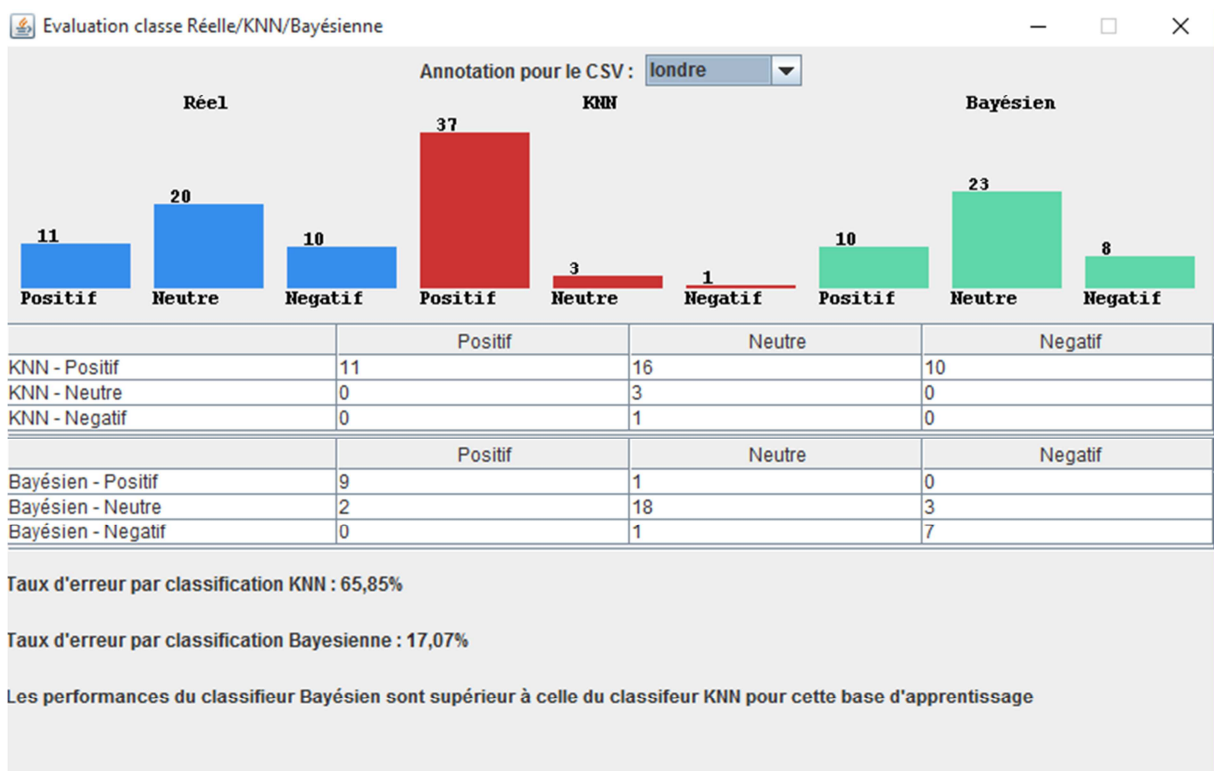


Figure 5 : Ecran d'analyse du csv

2.4.2. Manuel d'utilisation

L'interface est très simple d'utilisation. Une fois l'interface lancée vous arrivez sur l'interface principale.

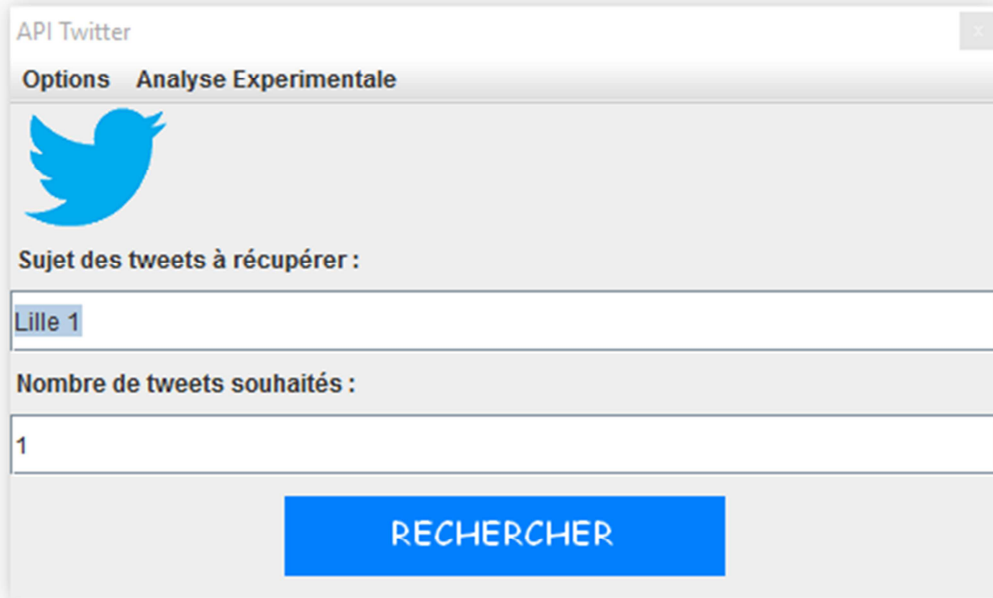


Figure 6 : Ecran principale

A partir de cette interface vous aurez accès aux principales fonctions de l'application :

- Rechercher des Tweets
- Sauvegarder les résultats de la recherche
- Effectuer une analyse via KNN ou Bayésien de la base d'apprentissage

Divers cas d'erreurs d'utilisation ont été pris en compte dans la programmation de l'interface comme :

- Recherche avec aucun mot-clé
- Recherche avec un indice de nombre de tweet souhaité inférieur à 1 ou null
- Impossibilité d'utiliser une méthode de classification
- Paramètres saisis incorrects.

Donc pour effectuer une recherche il suffit de renseigner dans la zone indiquée le mot-clé de la recherche et le nombre de tweets souhaités.

Il vous suffit par la suite d'appuyer sur le bouton de « Recherche » pour récupérer les tweets et les afficher dans un tableau.

Tweets obtenu

Options		
Alias	Tweet	Annotation
DjibriL.R	On a donc perdu contre :ParisMarseilleLyonMonacoCaenLille(Et Angers dans 1 semaine)#EquipeDeBabousFragiles	Neutre
Actualité 24	Ligue 1 : Lille confirme, Montpellier respire, Guingamp coule. https://t.co/qRgFDRban	Neutre
Flood lord	Régionales 2015: Une victoire de Marine Le Pen serait un tremplin pour 2017 https://t.co/VmhWPh5yyd https://t.co/uNzXAaGbg9 #infloodwetru...	Neutre
Flood lord	Régionales 2015: Xavier Bertrand espère faire barrage à Marine Le Pen https://t.co/okVWz7jrBk https://t.co/yojGa78xb8 #infloodwetru...	Neutre
Jane	A Lille je lui donne 1 semaine après il se fait tabasser c'est sûr https://t.co/BE2raSbMwr	Neutre
20 Minutes	Régionales 2015: Une victoire de Marine Le Pen serait un tremplin pour 2017 https://t.co/WYH5VYMIMB https://t.co/KcJmZUpkDr	Neutre
20 Minutes	Régionales 2015: Xavier Bertrand espère faire barrage à Marine Le Pen https://t.co/Kol09J4BjT https://t.co/BJjuwk7cQ2	Neutre
Médias Fran...	Dans #7a8Les tortues Golfina au MexiqueLa mort de Julie MartinLes boutiques du Vieux Lille https://t.co/Ct6lZbw1SH	Neutre
PlaceToBe	L'élection #MissFrance 2016 au Zénith de #Lille. Retour sur une très belle ville. https://t.co/ij8jBtKDaU https://t.co/pKXqQZ1yIC	Positif
Arnaud Gi	Devoir du citoyen accompli! #nord #electionsregionales #Regionales2015 #Lille #regions https://t.co/gW15eDANtY	Neutre

SAUVEGARDER TWEETS

Figure 7 : Ecran des Tweets récupérés

Une fois les tweets récupérer vous avez la possibilité soit de l'évaluer vous-même en cliquant sur la zone annotation est choisir entre :

- Négatif
- Positif
- Neutre

Ou/Et de faire « Sauvegarder tweets » et si le tweet était annoté manuellement alors il sera sauvegarder comme tel après nettoyage sinon une classification du tweet sera effectuer par recherche de mots-clefs.

API Twitter

Options Analyse Experimentale

☐ Annotation manuelle

☒ Annotation automatique

☐ Annotation par KNN

☐ Annotation par Bayesienne

Configuration

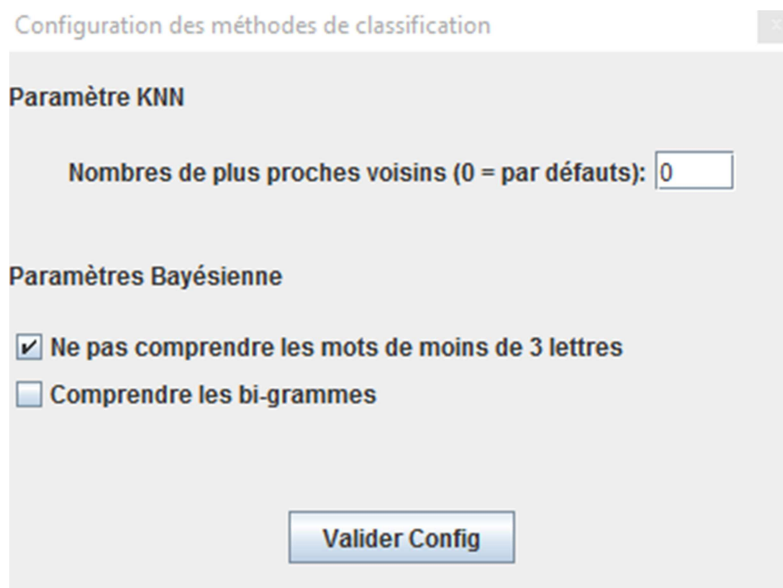
Exit

1

RECHERCHER

Figure 8 : Les options possibles avec l'interface

Dans le menu, l'onglet « option » permet de modifier le choix de l'annotation des tweets, configurer les méthodes de classification ou quitter l'interface.



Configuration des méthodes de classification

Paramètre KNN

Nombres de plus proches voisins (0 = par défauts):

Paramètres Bayésienne

☒ Ne pas comprendre les mots de moins de 3 lettres

☐ Comprendre les bi-grammes

Valider Config

Figure 9 : Ecran des options

Sur l'interface de paramétrage des méthodes de classification vous avez la possibilité de modifier la méthode KNN en indiquant le nombre de plus proche voisin souhaité ou encore de validé des options pour la méthode bayésienne.



API Twitter

Options **Analyse Experimentale**

Reelle/KNN

Reelle/Bayesienne

Reelle/KNN/Bayesienne

Sujet des tweets à récupérer :

Nombre de tweets souhaités :

RECHERCHER

Figure 10 : Analyse Experimentale

L'interface principale permet de lancer des analyse expérimentale sur nos base de données, pour cela il suffit de cliquer sur « Analyse Experimentale » dans le menu comme ci-dessus et de choisir l'analyse à faire :

- Reelle/KNN
- Reelle/Bayésienne
- Reelle/KNN/Bayésienne

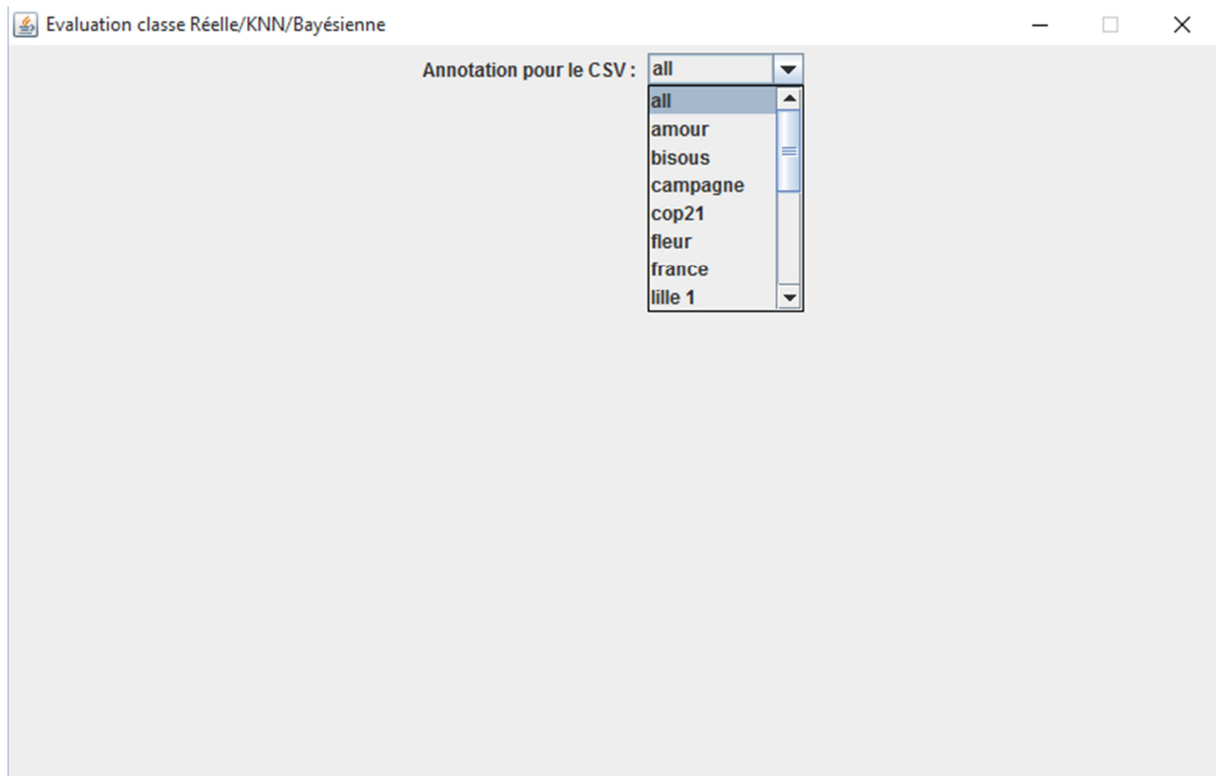


Figure 11 : Ecran de sélection des csv pour analyse expérimentale

Vous arrivez sur l'interface ci-dessus qui permet de choisir sur quelle base d'apprentissage on souhaite effectuer l'analyse expérimentale.

Vous pouvez modifier la base d'apprentissage de l'analyse dès que vous le souhaitez.

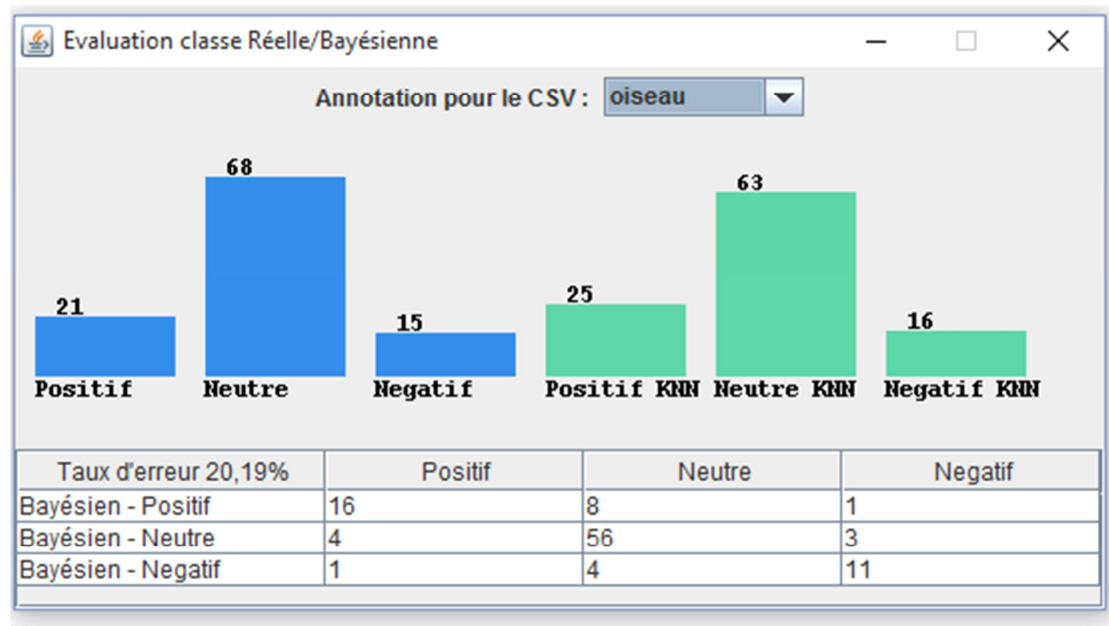


Figure 12 : Evaluation classe Réelle/Bayésienne

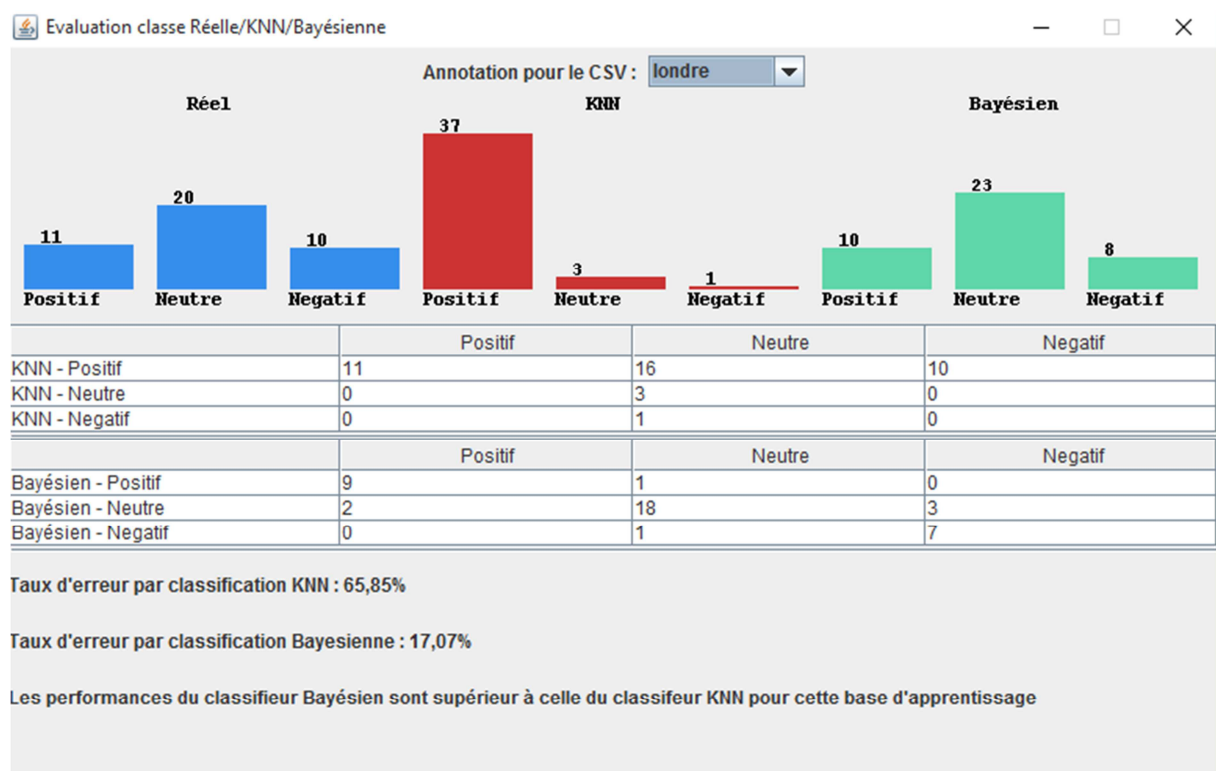


Figure 13 : Evaluation classe Réelle/KNN/Bayésienne

Ci-dessus vous avez un exemple d'analyse expérimentale pour « Reelle/KNN/Bayésienne ».

Il y a un rendu visuel de la répartition des tweets selon les méthodes de classification. Des tableaux qui indiquent le nombre de tweet ayant leurs polarités modifiées et en quoi (exemple après vérification du tweet via l'algorithme Bayésien : Positif vers Neutre).

L'affichage comprend également les taux d'erreur de classification selon les algorithmes utilisés et il permet d'interpréter les résultats.

3. Résultats

Durant ce projet j'ai récupéré environ plus de 1500 tweets. Les différentes méthodes de classification (par mot-clé, KNN et Bayésien) sont fonctionnelles.

La vérification de la qualité de la base d'apprentissage a rapidement souligné que beaucoup de tweet voit leur polarité modifier en fonction de l'algorithme effectué.

Des problèmes pour la classification des tweets par KNN est rapidement apparu, si la base de donnée est majoritairement polarisée positive, neutre ou négative, l'algorithme va classer les tweets restant sur l'annotation majoritaire. Le problème varie en fonction de la qualité de base, quantité de données en base mais aussi des paramètres indiqués pour l'algorithme.

La classification des tweets avec Bayésien semble dans ce cas beaucoup plus fiable avec des bases d'apprentissage contenant moins de 200 tweets que la méthode KNN.

Le taux d'erreur calculé grâce à l'interface pour chaque base d'apprentissage permet d'établir si la polarité de la base est fiable. Plus le taux est bas plus la base serait annotée correctement.

4. Conclusions

Ce projet a été riche d'enseignement et a été très intéressant. Actuellement l'application permet d'effectuer des recherches selon des mots-clés ainsi que d'annoter les tweets en fonction de divers méthode de classification (par mots-clés, par KNN, par Bayésien). Elle permet également de réaliser une analyse par sujet ou au global de la base d'apprentissage grâce à des tableaux et diagramme. Les taux d'erreur permettent aussi d'avoir une idée sur la fiabilité des bases de données et de déterminer quel sont les méthodes d'annotation les plus fiables.

Si j'avais la possibilité d'aller plus loin dans le projet j'aurais aimé ajouté un diagramme indiquant l'évolution de la polarité en fonction de la date de publication (du temps) des tweets obtenu pour voir l'évolution des avis de chacun sur divers sujet comme pour la politique ou autre.