UNIVERSITÉ RENNES 2

Projet sur la fouille de données textuelles (*Text Mining*)

Wenger vs Mourinho : qu'en pense Twitter ?

Pierre BUREAU
Guillaume LE FLOCH
Année 2017-2018

Encadrante : Mme Fabienne MOREAU

Table des matières

1 Introduction et problématique

Tout amateur de football, et plus particulièrement du championnat anglais de **Premier League**, est au courant de la rivalité qui existe entre **Arsène Wenger** et **José Mourinho**. En effet, l'histoire entre nos deux protagonistes est riche en péripéties, faisons un bref rappel des faits.

Lorsque José Mourinho est nommé manager du club londonien de **Chelsea** à l'été 2004, Arsène Wenger sort d'une saison incroyable avec le club rival d'**Arsenal**, puisque les *Gunners* ont remporté le championnat anglais sans avoir perdu le moindre match parmi les 38 journées sur lesquelles se déroule la Premier League. Arsenal termine donc champion avec **90 points**, c'est-à-dire 11 points de plus que le second... Chelsea !

On comprend ainsi qu'Arsène Wenger était à ce moment-là l'homme à abattre, et José Mourinho, qui sortait d'un doublé historique (Championnat du Portugal-Ligue des Champions) avec le **FC Porto** était l'élu pour accomplir cette tâche. Tout est fait pour opposer ces deux-là : l'entraîneur français est connu pour sa classe, pour donner leur chance aux jeunes joueurs, pour faire jouer son équipe avec un style « léché » et offensif, quand son homologue portugais est décrit comme un grand tacticien, un entraîneur vicieux, au pragmatisme légendaire, qui s'appuie plutôt sur une défense de fer et des joueurs expérimentés. Mais ce n'est pas tout, José Mourinho est également un maître dans l'art de la provocation, pour déstabiliser l'adversaire. Il va ainsi lancer sans cesse des piques à Arsène Wenger ou Arsenal dans les médias, en conférence de presse avant ou après les matchs, ce qui va définitivement rendre la relation entre les deux managers tendue et mener à des altercations sur le bord du terrain lors des oppositions entre Arsenal et Chelsea, qui étaient déjà rivaux à la base. Chelsea terminera champion en 2005 et 2006, avant que Mourinho ne soit évincé. Il fera son retour à Chelsea en 2013, ce qui va raviver les tensions dans notre couple favori, remportera à nouveau la Premier League en 2015, puis sera à nouveau limogé 6 mois plus tard après une première partie de saison catastrophique. En 2016, il signe à Manchester United, autre rival historique d'Arsenal. L'histoire se poursuit donc encore aujourd'hui en 2018.

La situation actuelle est la suivante : Arsène Wenger est toujours en poste à Arsenal (depuis 1996) mais n'a plus remporté le championnat depuis cette fameuse saison 2003-2004. José Mourinho a terminé 6^e du championnat après avoir dépensé une fortune sur le marché des transferts pour bâtir son équipe. Cette saison, Manchester United et Arsenal n'arrivent pas à tenir la cadence infernale du Leader de Premier league, Manchester City, et sont à la traîne en championnat. Cela fait 14 ans qu'Arsenal n'a pas remporté le championnat, l'équipe commet tout le temps les mêmes erreurs, tandis que pour Mourinho son équipe est ennuyeuse à regarder jouer (à cause de sa tactique du « *park the bus* » qui consiste à défendre très bas en utilisant tous ses joueurs de champ et donc à « garer le bus » devant son but pour empêcher l'adversaire de marquer), il n'a plus autant de succès qu'avant, c'est un « chequebook-manager » (dans le sens où il ne bâtit son succès qu'en dépensant des sommes astronomiques pour attirer des grands joueurs, par opposition à Wenger qui développe des jeunes talents n'ayant quasiment rien coûté).

En bref, aujourd'hui ces deux managers sont très décriés, beaucoup de fans d'Arsenal et de Manchester United semblent vouloir le limogeage de leur manager respectif, en tout cas si l'on s'en fie à ce que l'on peut voir dans les journaux anglais ou sur les réseaux sociaux.

Nous avons donc décidé de nous pencher sur le cas de ces deux personnalités du monde du football et de vérifier si ces hypothèses sont vraies, en analysant les tweets dans lesquels ils sont mentionnés. Notre problématique sera la suivante : **Quelle est l'opinion des utilisateurs de Twitter vis-à-vis d'Arsène Wenger et de José Mourinho ?**

Cette problématique est assez large, et pour cette raison, notre étude va s'articuler en plusieurs étapes pour y répondre :

- Une analyse préliminaire des données afin de procéder à une sélection de tweets pertinents
- Une analyse globale des tweets concernant chaque manager, afin de résumer les opinions
- Une analyse plus détaillée pour tenter d'extraire de vraies tendances aux niveau des **hashtags** et de la **valeur sentimentale** des tweets

2 La récupération des données

Une fois la problématique choisie, l'étape suivante de cette étude a été de construire notre base de données. Comme nous souhaitions analyser l'opinion des utilisateurs de **Twitter**, nous avons procédé au *scraping* des tweets grâce à la librairie **tweepy** en Python, qui possède une API permettant de communiquer directement avec l'interface du célèbre réseau social.

Il est important de noter que l'extraction des tweets ne peut pas s'opérer sur des tweets anciens (c'est la réglementation de Twitter). Ce sont donc tous les tweets qui sont envoyés après le lancement du programme qui sont enregistrés et écrits dans un fichier au format *json*. Ensuite, vous pouvez laisser votre programme tourner indéfiniment du moment que la connexion internet n'est pas coupée. Cependant, il existe des limitations (*cf documentation complète dans les sources*) comme pour toute API. Pour notre cas particulier, les tweets qui ont été extraits contenaient soit le mot « wenger » soit « mourinho » (le code correspondant se trouve dans le script *twitter_streaming.py*). En ce qui nous concerne, l'extraction s'est faite en 4 fois pour plusieurs raisons :

- Il fallait un volume de données conséquent afin d'avoir suffisamment de tweets à analyser après avoir effectué un nettoyage complet
- Nous ne pouvions pas laisser tourner le programme assez longtemps pour emmagasiner assez de tweets en une seule fois
- L'opinion des fans peut être biaisée à la suite d'un match perdu ou gagné par Arsenal et/ou Manchester United, nous avons donc souhaité capter des tweets à différents moments

Au final, nous nous retrouvons avec **1.60 Go** de données brutes à analyser, après agrégation des tweets extraits entre le **22 décembre 2017** et le **03 janvier 2018** (le code pour agréger les données correspond au script *data_concat.py*). Ces données sont regroupées dans un fichier texte appelé *twitter_data.txt*, qui est donc composé de tweets. Ces derniers sont au format *json*, c'est à dire que ce sont des **dictionnaires** en Python. Ils renferment énormément d'informations et sont de la forme suivante :


```
{
    created_at: "Wed Jan 03 08:17:09 +0000 2018",
    id: 948468259682168832,
    id_str: "948468259682168832",
    text: "Stewart Robson is with us now talking Arsenal v Chelsea, and Wenger v Mike Dean.\n\nGet your questions and comments i… https://t.co/uSAc2zEIZ7",
    source: "<a href=\"https://about.twitter.com/products/tweetdeck\" rel=\"nofollow\">TweetDeck</a>",
    truncated: true,
    in_reply_to_status_id: null,
    in_reply_to_status_id_str: null,
    in_reply_to_user_id: null,
    in_reply_to_user_id_str: null,
    in_reply_to_screen_name: null,
  ▾ user: {
        id: 401488459,
        id_str: "401488459",
        name: "Off The Ball",
        screen_name: "offtheball",
        location: "🎙 - 📻 - 🎧 - 📹",
        url: http://www.offtheball.com,
        description: "📱#OTBAM weekdays 7:45-9am 📻 #OffTheBall Mon-Fri from 7pm, Sat-Sun from 1pm. 🌪 sport@offtheball.com 📷 offtheballnt",
        translator_type: "none",
        protected: false,
        verified: true,
        followers_count: 82571,
        friends_count: 944,
        listed_count: 427,
        favourites_count: 1263,
        statuses_count: 33882,
        created_at: "Sun Oct 30 16:32:42 +0000 2011",
        utc_offset: 0,
        time_zone: "Casablanca",
        geo_enabled: true,
        lang: "en",
        contributors_enabled: false,
        is_translator: false,
        profile_background_color: "C0DEED",
        profile_background_image_url: http://abs.twimg.com/images/themes/theme1/bg.png,
        profile_background_image_url_https: https://abs.twimg.com/images/themes/theme1/bg.png,
        profile_background_tile: false,
        profile_link_color: "1DA1F2",
        profile_sidebar_border_color: "C0DEED",
        profile_sidebar_fill_color: "DDEEF6",
        profile_text_color: "333333",
        profile_use_background_image: true,
        profile_image_url: http://pbs.twimg.com/profile_images/906172613533339649/6CR1ue1x_normal.jpg,
        profile_image_url_https: https://pbs.twimg.com/profile_images/906172613533339649/6CR1ue1x_normal.jpg,
        profile_banner_url: https://pbs.twimg.com/profile_banners/401488459/1513780128,
        default_profile: true,
        default_profile_image: false,
```

Dans notre étude, nous allons nous servir principalement de la clé 'text' qui renferme le contenu des (désormais) 280 caractères maximum à disposition de l'utilisateur pour s'exprimer. La clé 'language' nous servira également dans l'analyse préliminaire pour sélectionner les tweets qui seront dans notre base finale à exploiter.

3 Analyse préliminaire et nettoyage des données

Extraire tous les tweets qui contiennent une chaîne de caractère donnée n'est pas compliqué, en revanche on peut se douter à l'avance que tout ne sera pas pertinent et que l'on va rencontrer certains problèmes. C'est pour cette raison qu'il nous fallait une base conséquente au départ, car elle va se retrouver largement affinée par la suite.

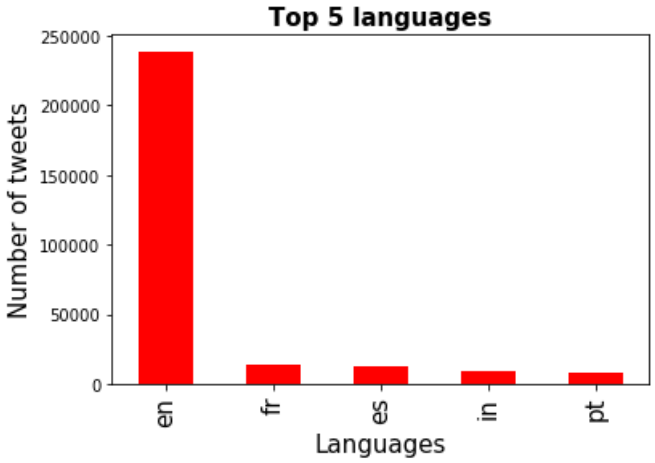
Un problème que nous n'avons pas eu, mais que nous aurions pu rencontrer concerne les différents sens du mot cible qui nous sert à effectuer l'extraction des tweets. Par exemple, si nous avions voulu nous pencher sur le cas de l'actuel manager de Chelsea, **Antonio Conte** et que nous avions seulement utilisé le mot clé « conte » pour y faire référence, nous nous serions retrouvés avec des tweets parlant de personnes d'origine africaine portant le même nom, ou bien encore par exemple d'un « conte de Noël ». Et si vous renseignez « antonio conte » (l'API ne tient pas compte de la casse) vous ne pourrez extraire que les tweets contenant la chaîne complète, ce qui réduira donc largement le champs des tweets accessibles.

En ce qui nous concerne, il n'existe pas d'ambiguïté sur Mourinho et Wenger, ce sont des noms propres et aucune autre personnalité ne porte le même. En revanche, lorsque les fans ou journalistes parlent d'eux ils utilisent parfois « Mou » ou « José » pour parler de Mourinho et « AW », « the Boss » ou encore « Big Wengz » concernant Arsène Wenger. Nous n'avons donc pas pu capter tout ce qui se disait sur eux.

Néanmoins, nous avons réussi à extraire un total de **304 512** tweets. Un premier nettoyage a consisté à enlever les tweets pour lesquels il n'y avait pas de texte (ce qui correspond à une liste vide associée à la clé 'text' dans le dictionnaire python). En effet, un autre problème lié à la fouille de texte est que l'on ne peut pas exploiter les images, les GIFs ou encore les vidéos qui sont pourtant très porteurs de sens. Après avoir procédé à cette opération, il nous reste tout de même **304 488** tweets.

Un autre problème que l'on peut rencontrer avec les tweets est la présence de **doublons** ou de tweets qui sont quasiment des doublons. Cela concerne les fameux **retweets** : quand une personne va retweeter un tweet, cela va apparaitre sous la forme « RT : blablabla ». Pour régler ce problème, nous avons donc décidé de supprimer tous les tweets commençant par « RT », on ne garde ainsi que le tweet original.
Le problème des « quasi-doublons » était quant à lui dû aux liens (vers un article de presse par exemple) : on retrouve à chaque fois le même contenu textuel, mais puisque le lien est différent, la méthode *drop_duplicates()* de la librairie **Pandas** en Python ne détecte pas le doublon. La solution que nous avons choisie consiste tout simplement à supprimer au préalable tous les liens dans les tweets à l'aide d'une expression régulière.

Un autre aspect à prendre en compte est la langue dans laquelle est écrit le tweet. En effet, Arsenal et Manchester United sont des clubs de football suivis dans le monde entier, des fans de tous les pays tweetent à propos de ces 2 entités et de leurs managers. Nous avons donc décidé de regarder un *Top 5* des langues dans lesquels nos tweets sont écrits.



Les résultats sont sans appel, une écrasante majorité des tweets sont rédigés en anglais, ce qui n'est pas surprenant puisque la Premier League est beaucoup suivie au Royaume-Uni, en Amérique du Nord, en Inde et au Nigeria. Le français représente la deuxième langue, probablement grâce à Arsène Wenger et à la communauté francophone d'Arsenal. Cependant la quantité de tweets est insuffisante, puisque les doublons n'ont pas encore été enlevés. Nous allons donc conserver uniquement les tweets en anglais.

Après avoir effectué ces différentes opérations, nous allons stocker les tweets concernant José Mourinho dans un dataframe *mourinho.csv*, idem pour Arsène Wenger (*wenger.csv*). Au final, nous avons 26087 tweets pour José Mourinho et 28244 tweets pour Arsène Wenger, à analyser.

4 Traitement global des tweets

Lorsqu'on parle de traitement global des tweets (par opposition au traitement détaillé que l'on verra par la suite), nous faisons référence à une première approche assez basique qui consiste à regarder la fréquence des mots dans notre corpus. Pour ce faire, nous avons utilisé les modules de la librairie Python **NLTK** (Natural Language ToolKit). Le code est disponible dans le script *traitement_nltk.py*). Le but est de regarder les mots les plus fréquents afin d'obtenir un premier résumé des tweets concernant chaque manager, pour créer un **nuage de mots** à l'aide de l'outil *wordle.net*. Nous trouvons que cette représentation est visuelle et permet d'obtenir une première tendance en ce qui concerne les thèmes liés à nos protagonistes. Pour parvenir à une représentation « propre », il faut effectuer quelques traitements au préalable :

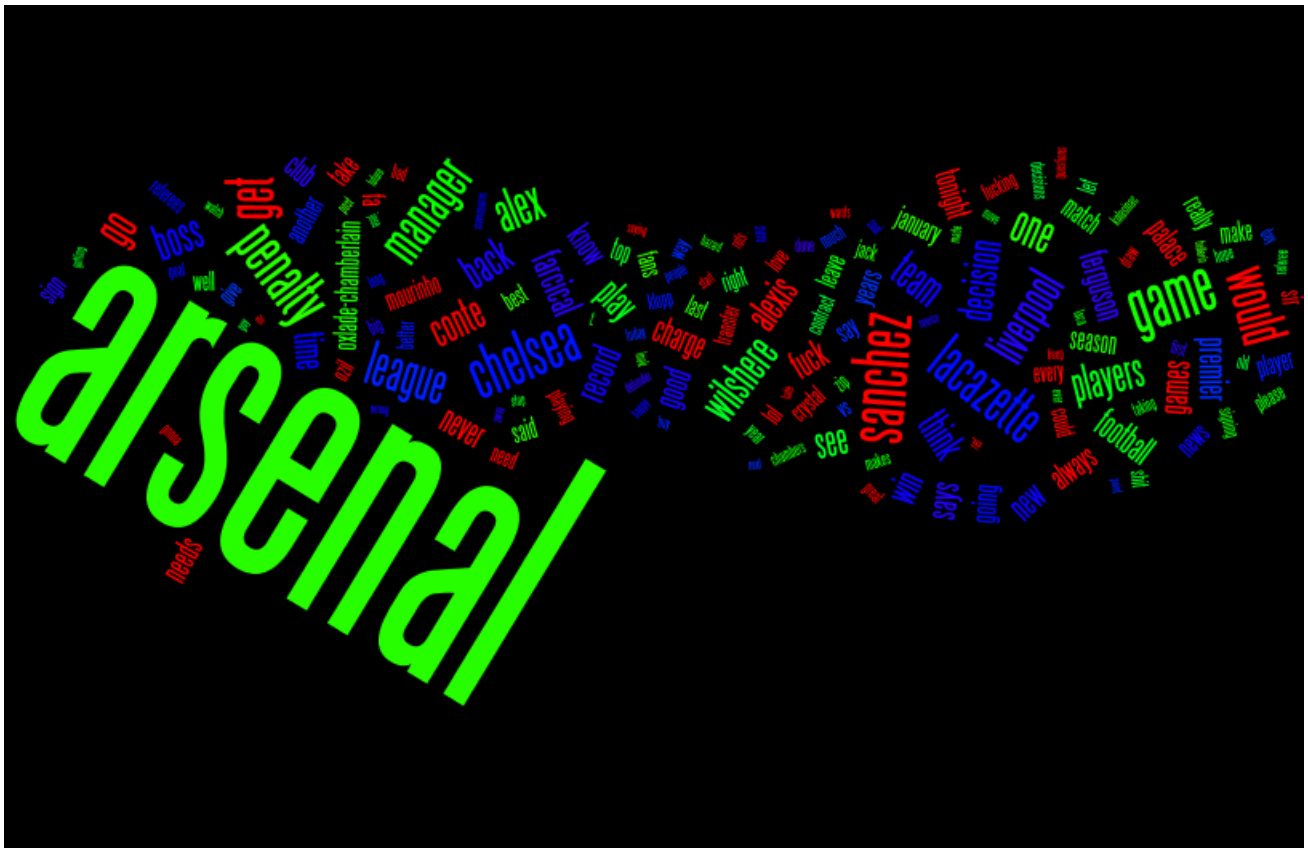
- Enlever les « RT » dans les tweets car ils ne nous sont d'aucune utilité
- Enlever les noms d'utilisateurs (qui sont précédés d'un « @ ») car nous jugeons qu'ils pourraient polluer l'analyse
- Enlever tous les éléments de la ponctuation et les accents à l'aide d'expressions régulières
- Enlever les mots vides et en ajouter certains qui ne figurent pas dans la liste implémentée par le module *nltk*, puis enlever les hashtags puisqu'ils seront traités séparément à la fin de cette étude
- Supprimer des caractères « parasites » propres à l'extraction des tweets et que les expressions régulières n'ont pas pu enlever comme '...' ou encore les emojis : pour cela nous avons implémenté la fonction **remove_useless** qui ne conserve que les chaînes de caractères contenant seulement des chiffres, des lettres et/ou des symboles monétaires (qui ont leur importance dans notre analyse)
- Regrouper tous les mots en un seul paragraphe (un pour chaque manager) afin de former notre propre corpus
- Enlever les mots « wenger » et « mourinho » qui vont être les plus représentés, puisque leur fréquence sera supérieure ou égale au nombre de tweets, car ils risquent d'écraser les autres mots

Une fois ce travail effectué, on peut se pencher sur l'analyse des corpus « propres » que nous avons formé.

4.1 Une analyse simple de la fréquence des mots

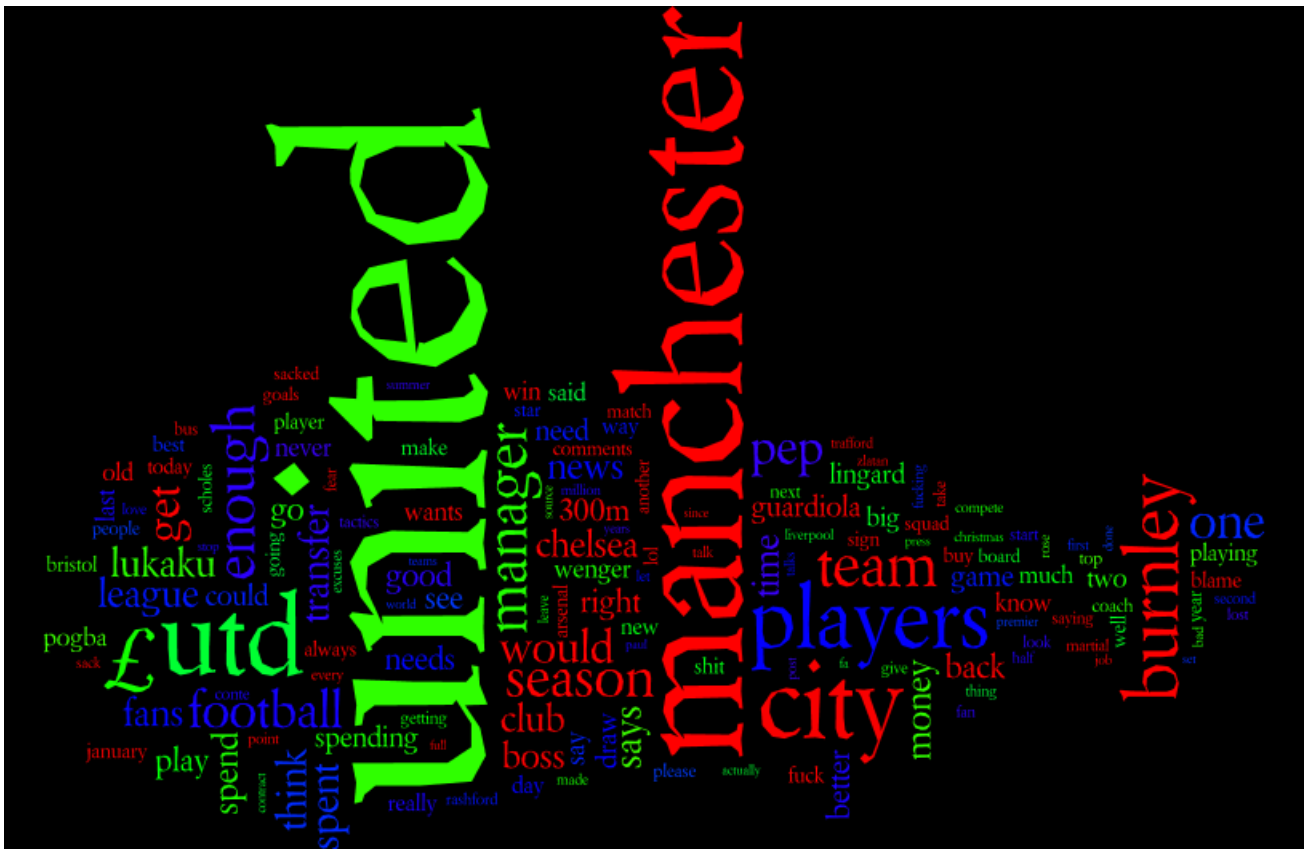
Ce qui ressort de cette première analyse s'est retrouvé assez influencé par la période de scraping des tweets. Il est donc logique de trouver dans nos résultats le nom des adversaires d'Arsenal et de Manchester United, ainsi que les noms de certains joueurs ayant fait l'actualité dans ces matchs. A côté de ces aspects, on notera tout de même certain points intéressants que nous allons détailler tout de suite avec les nuages de mots.

En ce qui concerne Arsène Wenger, le résultat visuel est à la page suivante.



Arsenal a affronté Liverpool, Chelsea et Crystal Palace, on retrouve donc les noms de tous ces clubs ainsi que les noms des managers rivaux de Liverpool (Klopp) et Chelsea (Conte). Au niveau des joueurs, **Alex Oxlade-Chamberlain** qui a joué sous les couleurs d'Arsenal pendant 6 saisons et qui a été transféré à Liverpool l'été dernier a fait parler de lui pour son retour du côté de l'Emirates Stadium. Les autres joueurs associés à Arsène Wenger dans les tweets sont des joueurs d'Arsenal : Jack Wilshere, Alexis Sanchez et Alexandre Lacazette.
Ensuite, on retrouve des termes tels que « farcical », « decision », « penalty », « FA » ou encore « charge » qui font référence aux critique d'Arsène Wenger à l'encontre des arbitres et à la suspension qui lui a été infligée par la Fédération Anglaise.
Il a également été question de Sir Alex Ferguson, l'illustre manager resté pendant 26 ans à la tête de Manchester United et rival en son temps d'Arsène Wenger. Le français a égalé son homologue écossais en terme de matchs en Premier League (810), durant notre période d'extraction des tweets.
Enfin, on remarque la présence de termes comme « needs », « go » qui peuvent suggérer des phrases du style « Wenger needs to go », lorsque les fans veulent qu'il s'en aille. Pour vérifier cela nous ferons par la suite une analyse de **bigrammes**, **trigrammes** et même **quadrigrammes** qui pourront peut-être apporter plus d'informations à ce niveau-là. Comme il est question également de la période des transferts (« transfer », « january ») cela pourrait également concerner des joueurs.

Dernier point intéressant : il est question de José Mourinho, souvent lorsque les gens tweetent à propos de Wenger, ils tweetent en même temps à propos de Mourinho. Nous allons vérifier tout de suite si cela fonctionne également dans l'autre sens en se penchant sur le nuage de mots issu des tweets sur le manager portugais.



On retrouve donc cette réciprocité puisque Wenger est dans le nuage de mots. Avant cela, on comprend bien que le sujet tourne autour de Manchester United et de son « boss ». Il est également question des adversaires de United, à savoir Bristol City et Burnley. Les twittos ont également beaucoup comparé Mourinho à son plus grand rival ; Pep Guardiola le manager de Manchester City. Ils ont parlé de joueurs de United tels que Lukaku, Pogba, Ibrahimovic ou encore Lingard. Il est intéressant de noter la fréquence élevée d'apparition du symbole monétaire de la **livre sterling**, des termes « spend », « money » et « 300m », quand on se rappelle de la réputation de « chequebook manager » du technicien portugais. Il a en effet affirmé en conférence de presse après des comparaisons avec le Manchester city de Guardiola qui caracole en tête du classement que les 300 millions de pounds qu'il avait dépensé pour construire son équipe n'étaient pas encore suffisants. Les fans sur Twitter s'en sont donc donnés à cœur joie.
On notera également qu'il était question de son ancien club de Chelsea, et du marché des transferts avec des mots tels que « january », « buy », « sign », ...
Enfin, on note la présence de « park » et « bus », qui font référence à cette fameuse tactique employée par l'équipe de José Mourinho pour défendre, et le mot « sacked » qui signifie tout simplement « limogé ». Serait-ce un signe de lassitude des supporters vis-à-vis de Mourinho ?

4.2 Une tentative d'amélioration

Le premier traitement était « brut », dans le sens où nous avons regardé les mots tels quels. Une deuxième approche va consister à rajouter deux étapes supplémentaires en faisant une **racinisation** et une **lemmatisation** des mots. En effet, avant ces étapes on fait une distinction par exemple entre les mots « wins » et « win » tandis qu'ils désignent le même concept. De même avec « spend » et « spending ».
En rajoutant ces étapes, on peut donc régler ces problèmes, mais on en apporte d'autres : certaines fois ce n'est pas pertinent car quand on racinise le mot « league » par exemple il devient « leagu »... Regardons donc ce que donnent ces traitements au niveau des résultats.

La différence n'est pas flagrante, et dans notre analyse elle n'apporte rien de significatif et aurait même tendance à dégrader la qualité des résultats avant de procéder à une normalisation des mots. Nous allons donc nous pencher sur l'analyse des **ngrams** pour tenter d'enrichir l'analyse textuelle.

4.3 Complément d'analyse : les ngrams

En ce qui concerne les ngrams, les traitements appliqués seront les mêmes que précédemment, à la différence près que l'on va conserver les hashtags ici (le code se trouve dans le script *n_grams.py*). On regarde la fréquence des ngrams, voici ce que l'on trouve pour Arsène Wenger

5 Traitement détaillé des tweets

6 Bilan

7 Sources

- **Réglementation de Twitter**, *https://developer.twitter.com/en/developer-terms/agreement-and-policy*

- **Scraping de tweets**, *https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/*

- **Analyse de sentiment**, *http://www.nltk.org/howto/sentiment.html*

- **Elements des cours et des TP de Text Mining vus pendant le semestre**