# Rapport de Stage - Analyse de données textuelles musicologiques

Nadir JALALUDEEN

Juin 2024

## 1 Objectifs du stage

Ce stage vise à produire une introduction à l'analyse de données textuelles ainsi qu'à la programmation en Python, appliquée à un cas d'usage spécifique. Il s'agira d'analyser des discussions extraites d'un forum spécialisé consacré à la musique et à la musicologie. Ce forum, dont nous fournirons des détails supplémentaires plus bas, constitue un espace numérique d'échange entre experts, traitant d'un large éventail de sujets musicaux et musicologiques. Ce stage permettra de se familiariser avec ce jeu de données spécialisé et de tester diverses méthodes computationnelles pour en proposer une analyse détaillée.

Les objectifs du stage sont les suivants :

- Analyse des discussions sur le forum musiSorbonne pour l'année 2003 à l'aide de différentes méthodologies de traitement du langage naturel.

- Compréhension des dynamiques de communication du forum à l'aide du langage de programmation Python 3[1] ainsi que ses bibliothèques spécialisées (détallées ci-après).

Bien que l'année 2003 fût l'une des premières années d'existence de musiSorbonne, ayant été créé en 2002, le forum comptait déjà plus de 1000 messages sur cette période. En estimant qu'il faut environ une minute pour lire un message moyen, la lecture des 1486 messages de ce jeu de données nécessiterait plus de 20 heures. Par conséquent, l'utilisation d'outils numériques d'analyse est essentielle pour dévoiler certaines tendances et extraire des statistiques significatives sur le forum musiSorbonne.

## 2 Qu'est-ce que musiSorbonne ?

### 2.1 Présentation générale

musiSorbonne est un forum musicologique dédié aux discussions sur la musique, couvrant une gamme de sujets allant de l'analyse musicale à l'histoire de la musique. Créé par Nicolas Meeùs[7], ce forum vise à favoriser les échanges entre passionnés et experts du domaine. Avec ses 1300 abonnés en 2022, il est devenu

l'un des principaux forums musicologiques en France, offrant une plateforme précieuse pour le partage de connaissances et de réflexions approfondies sur divers aspects musicaux. Plus de détails sur le forum et ses fonctionnalités sont disponibles sur le site officiel :

http://nicolas.meeus.free.fr/musiSorbonne.html

Le forum musiSorbonne a été créé par un groupe de musicologues de l'Université Paris-Sorbonne au début des années 2000, sous l'impulsion de Nicolas Meeùs[7]. L'objectif principal était de fournir une plateforme de discussion dédiée aux chercheurs, étudiants et amateurs de musique, facilitant ainsi le partage de connaissances et la collaboration dans le domaine musicologique. Ce forum permet de débattre de divers sujets allant de l'analyse musicale à l'histoire de la musique, en favorisant un enrichissement mutuel parmi ses membres. Ce forum se présente sous la forme d'une liste de diffusion par mail, ce qui signifie que toutes les discussions se déroulent via ce moyen de communication. Il est administré par une équipe de modérateurs, souvent des professeurs ou des étudiants avancés en musicologie, qui veillent à ce que les échanges restent pertinents et respectueux.

# 3 Description et statistiques générales sur le jeu de données

## 3.1 Jeu de données

Le jeu de données utilisé pour cette analyse compile toutes les discussions du forum musiSorbonne pour l'année 2003, au format CSV. Chaque message, sous forme de courriel, inclut les informations suivantes : expéditeur, destinataire (musiSorbonne), date, objet et corps du message. Ce jeu de données a été chargé dans le notebook Jupyter[3] à l'aide de la librairie pandas[4].

Pour améliorer l'analyse, une fonction a été implémentée pour grouper les messages par discussion, ce qui peut entraîner de légères variations dans les statistiques liées aux discussions. De plus, une autre fonction a été développée pour identifier chaque expéditeur de manière unique. Toutefois, certains messages manquent d'informations sur l'expéditeur, ce qui peut introduire des erreurs dans les analyses centrées sur les individus.

## 3.2 Statistiques descriptives

En utilisant la librairie NumPy[5], on détermine les statistiques suivantes :

- Nombre total d'utilisateurs : 150
- Nombre total de messages : 1486
- Nombre total de discussions : 625
  - Si strictement plus d'un message : 248

- Nombre moyen de messages par discussion : 2.38
  - Si strictement plus d'un message : 4.47,. En se concentrant uniquement sur les discussions qui ont engendré des échanges, la moyenne augmente de manière significative, indiquant que les discussions actives sont relativement animées.
- Nombre moyen de mots par message : 351.7
  - Si premier message d'une discussion : 377.0. Les premiers messages ont tendance à être plus détaillés, probablement pour initier la discussion.
  - Si réponse à une discussion : 333.3. Les réponses sont légèrement plus courtes, ce qui est typique des échanges interactifs où les messages peuvent être plus directs et moins descriptifs.

# 4 Analyses

## 4.1 Analyse de la participation

### 4.1.1 Évolution de la participation au forum selon le temps

L'analyse des dates de publication des messages permet de dresser des graphiques Matplotlib[6] mesurant la participation en fonction d'une mesure de temps.

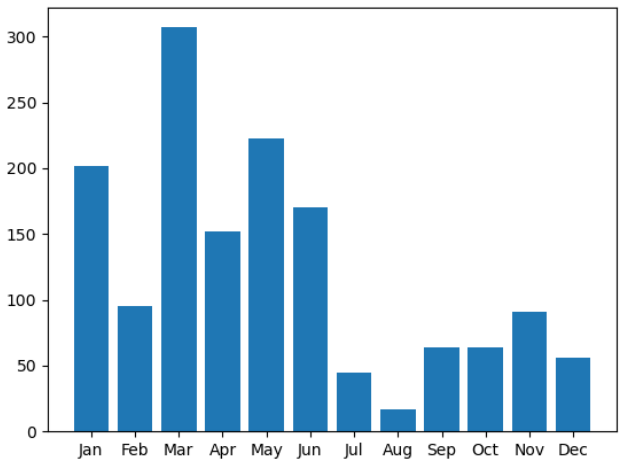


Figure 1: Nombre de messages selon le mois

- **Différence de participation entre les deux semestres** :

• Il y a une différence nette de participation entre les deux semestres de l'année 2003. La participation est significativement plus élevée entre janvier et juin, par rapport à la période de juillet à décembre. Cette tendance pourrait être influencée par des facteurs saisonniers, académiques ou personnels des participants.

• **Pic de participation en mars** :

• Le mois de mars se distingue par un pic notable de messages, indiquant une activité accrue. Ce pic pourrait correspondre à des événements spécifiques dans le domaine de la musicologie ou à des périodes de concertation académique.

• **Baisse de participation en août** :

• À l'inverse, le mois d'août affiche le plus faible nombre de messages. Cette baisse peut être attribuée aux vacances d'été, période durant laquelle les activités académiques et professionnelles ralentissent généralement, affectant ainsi la participation au forum.

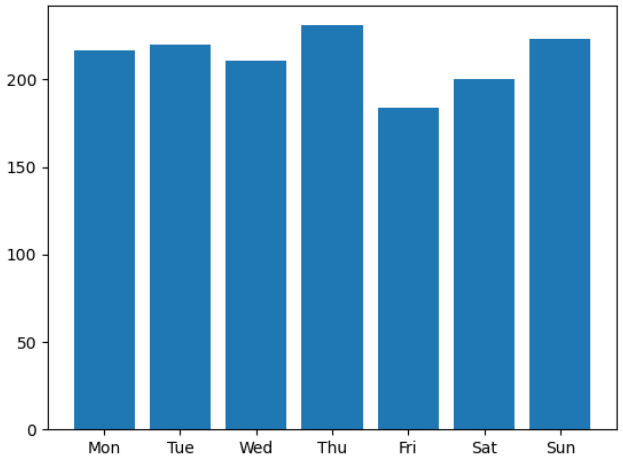


Figure 2: Nombre de messages selon le jour de la semaine

La participation reste relativement constante tout au long de la semaine. Cela suggère que les utilisateurs du forum sont engagés de manière régulière, sans variations notables liées aux jours spécifiques. Bien qu'il y ait une légère diminution de l'activité le vendredi, cette baisse n'est pas suffisamment prononcée pour être significative. Cela pourrait indiquer que les utilisateurs commencent à se désengager légèrement à l'approche du week-end, mais l'effet est minime. La constance de la participation en semaine signifie que les administrateurs peuvent planifier des activités et des interventions à tout moment sans se soucier de variations importantes de l'engagement. Cependant, ils pourraient envisager de renforcer les initiatives le vendredi pour compenser la légère baisse observée.

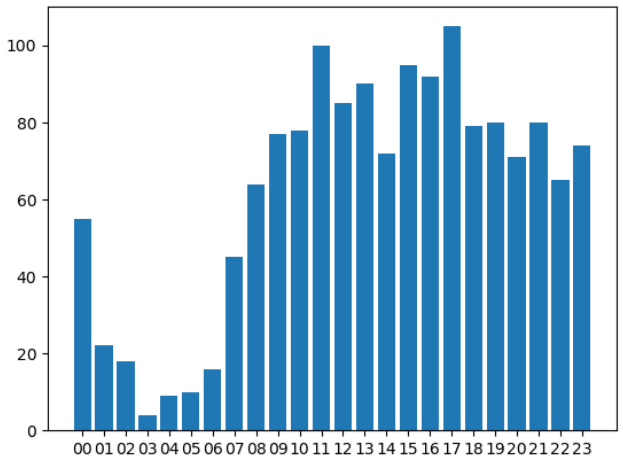


Figure 3: Nombre de messages selon l'heure

Tandis que pour l'heure, on remarque qu'il y a très peu de messages entre 1 heures et 6 heures du matin, ce qui pouvait être prévisible à cause de la nuit. Ainsi que deux pics de participation à 11 heures et 17 heures, séparé par une baisse assez nette vers 14 heures, pouvant être dû à la pause déjeuner.

#### 4.1.2 Participation selon les utilisateurs

L'analyse de la participation montre que la majorité des messages sont postés par un petit groupe d'utilisateurs très actifs, tandis que la majorité des utilisateurs contribuent moins fréquemment.

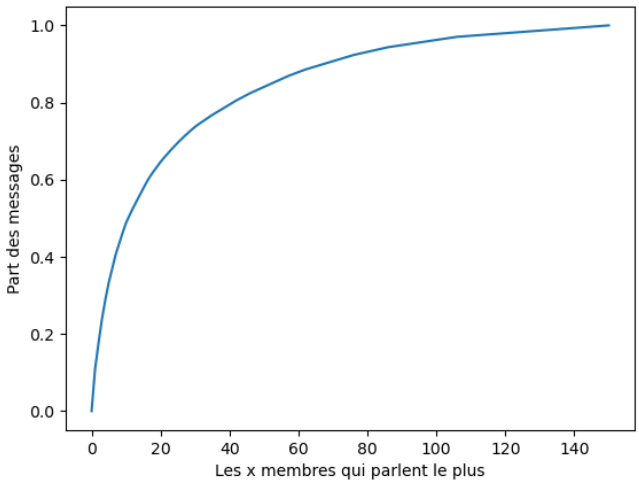


Figure 4: La part de messages publiés par les x membres qui parlent le plus

On remarque que la part de messages publiés par les x membres qui parlent le plus a l'air de suivre une courbe logarithmique.

Plus précisément,

- Nicolas Meeùs[7], l'administrateur principal de musiSorbonne, est l'auteur de 10% des messages.
- Les 11 utilisateurs qui parlent le plus sont les auteurs de 50% des messages.

On rappelle qu'il y a 150 utilisateurs au total.

## 4.2 Analyse de sentiments

Nous avons également eu l'occasion de faire de l'analyse de sentiments sur les différents messages à notre disposition, en utilisant le modèle VADER[8] proposé par la librairie NLTK[9].

### 4.2.1 Mise en place

La première étape consiste à faire du prétraitement, afin de faciliter l'analyse en réduisant le bruit. On a tout d'abord retiré les mots vides (stopwords), ce sont les mots qui apparaissent souvent et qui n'apportent aucune information tel que "il", "le", ou "et". Chacun des mots restants sont ensuite lemmatisé, c'est-à-dire réduit à sa forme la plus basique. Ainsi, "mangeait" deviendrait "mange', et "camionnettes" deviendrait "camion".

On peut ensuite appliquer le modèle d'analyse de sentiments VADER fourni par NLTK à chaque message. On va prendre la valeur "compound" du score retourné, qui est dans l'intervalle $[-1:1]$, qui est positif si le message est positif, et négatif si le message est négatif.

### 4.2.2 Glossaire

Soit $(u_k)_{k\leq n}$ une suite de n messages, $\mathbb{V}$ la fonction qui calcule la tendance sentimentale d'un message à valeurs dans [-1:1], on calcule ainsi:

- Le bilan sentimental S : $S = \sum_{k=1}^{n} \mathbb{V}(u_k)$

  Le nombre de messages a un impact sur ce score.

- La moyenne sentimentale M : $M = \frac{1}{n} \sum_{k=1}^{n} \mathbb{V}(u_k)$

- Le mouvement sentimental D : $D = \sum_{k=1}^{n-1} |\mathbb{V}(u_{k+1}) - \mathbb{V}(u_k)|$

  Le nombre de messages a un impact sur ce score.

- Le mouvement sentimental moyen T : $T = \frac{1}{n} \sum_{k=1}^{n-1} |\mathbb{V}(u_{k+1}) - \mathbb{V}(u_k)|$

  Variante du mouvement sentimental qui ne prend pas en compte le nombre de messages.

### 4.2.3 Observations

- Le bilan sentimental final de musiSorbonne 2003 est de 14, il est donc positif.

- La moyenne sentimentale de musiSorbonne 2003 est de 0.01.

- Il y a 462 messages positifs, 544 messages neutres, et 480 messages négatifs. Les messages neutres composent la majorité relative, ce qui est une bonne chose d'un point de vue académique. Néanmoins, il y a plus de messages négatifs que de messages positifs.

- Le message le plus positif est une annonce de conférence, la ICMPC8, avec une valeur de 0.99.

- Le message le plus négatif est une lettre ouverte disant que musiSorbonne n'est pas un lieu adapté à des débats concernant la guerre en Iraq, avec une valeur de -0.99.

- La discussion présentant le bilan sentimental le plus élevé est "Beethoven dans la musique contemporaine" avec une valeur de 5.

- La discussion présentant le bilan sentimental le moins élevé est "tension musicale" avec une valeur de -7.2.

- La discussion ayant été le plus mouvementée est "modes formulaires et modes scalaires" avec une valeur de 15.

- La discussion ayant été le plus mouvementée en moyenne est "josquin!" avec un mouvement moyen de 1.2.

On a également fait des analyses spécifiques à chaque utilisateurs, telle que le bilan sentimental ou le mouvement sentimental moyen, mais on ne les détaillera pas par soucis d'anonymat.

### 4.3 Analyse des communautés d'utilisateurs

Nous avons tout d'abord retranscrit les utilisateurs et leurs relations sous la forme d'un graphe pondéré non orienté, dont les points représentent les utilisateurs uniques, et les arrêtes le nombre de discussions en commun s'il y en a.

On a ensuite appliqué l'algorithme de détection de communautés de Louvain[10] proposé par la librairie NetworkX[11], pour finalement visualiser le graphe final avec la librairie PyVis[12].

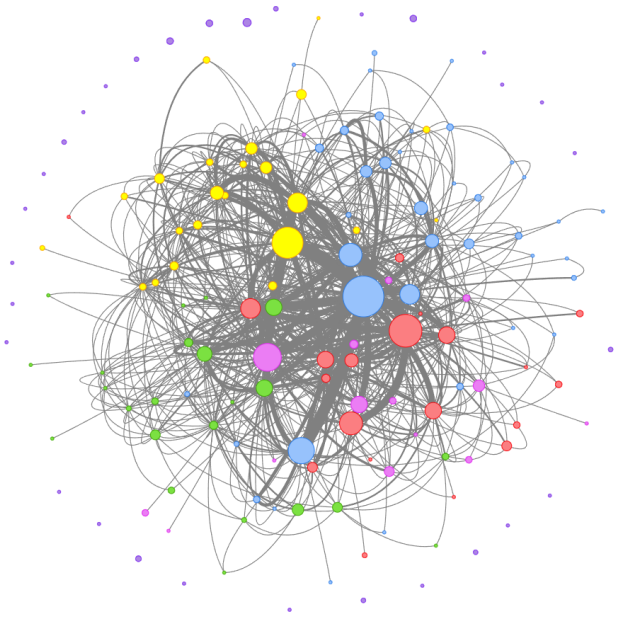

Figure 5: Visualisation du graphe relationnel entre chaque utilisateur

Sur cette figure, la taille des points est proportionnelle racinairement au nombre de messages postés par l'utilisateur représenté, tandis que la taille des arrêtes est proportionelle linéairement au nombre de discussions en commun.

On a ainsi pu identifier 6 communautés, dont une constituée d'utilisateurs n'ayant jamais interagi avec d'autres utilisateurs.

Ce graphe interactif peut être visualisé en passant par le notebook associé à ce rapport, ou à l'aide de ce fichier.

Intéressons nous ensuite à la modularité[13] de cette partition. La modularité d'une partition de graphe est une valeur comprise dans l'intervalle $[-\frac{1}{2} : 1]$ mesurant à quel point cette partition est révélateur de communautés au sein du graphe. Une valeur proche de 0 signifie que la partition est aussi efficace qu'une partition choisie aléatoirement. On peut réellement observer des communautés quand la modularité est supérieure à 0.3.

Au final, la partition finalement obtenue admet une modularité de 0.22. Ceci signifie qu'on peut exhiber des communautés, néanmoins l'impact de ces dernières sont assez faible.

Après avoir calculé un score de communauté C pour chaque discussion, correspondant à la part d'individu de la communauté C dans la discussion, multiplié par la racine du nombre de message dans la discussion. Nous avons pu identifier des semblants de tendances:

- La communauté bleue (39 membres) se distingue par l'utilisation fréquente d'un vocabulaire spécialisé et mathématique. Les membres abordent souvent des concepts musicologiques complexes, l'analyse harmonique, et l'utilisation de notations et formules mathématiques pour décrire des phénomènes musicaux. A titre d'exemple, les discussions peuvent inclure des termes comme "fréquences", "logarithmes", "harmoniques", et "intervalle d'intonation", soulignant un niveau élevé de précision et de rigueur académique.

  - 20/01/03 - Analyse mawwâl — Score : 3.61
    Présente beaucoup de fréquences en Hz, très précis au niveau des temps, usage de graphes.
  - 26/09/09 - Josquin — Score : 2.85
    Vocabulaire précis (minime, broderie, bassus), accords précis.
  - 01/04/03 - mélodie — Score : 2.67
    Accords précis, analyse en profondeur d'oeuvres.
  - 10/01/03 - Cent et Savart — Score : 2.34
    Usages de racines, de puissances et de logarithmes dans diverses bases.

- Le communauté rouge (20 membres) est caractérisée par des débats souvent animés et opposés sur une variété de sujets. Les discussions ont tendance à être plus controversées, avec une tonalité générale plus négative. Les membres de cette communauté s'engagent fréquemment dans des échanges passionnés et parfois conflictuels, abordant des thèmes qui suscitent des opinions divergentes.

  - 17/03/03 - la musique tonale contemporaine — Score : 4.56

Discussion en lien avec l'histoire de la musique présentant un bilan sentimental de -6.64.

  - 06/05/03 - valeur — Score : 2.83
    Discussion à tendance philosophique ayant des airs de conflits.
  - Parmi les 10 discussions ayant les scores de communauté le plus haut, 7 présentent un bilan sentimental négatif.

- La communauté rose (15 membres) discute surtout de l'histoire de la musique, en se focalisant également sur les autres cultures et pays.

  - 05/10/03 - Histoires de la musique — Score : 2.23
  - 10/05/03 - Chine et monde musulman — Score : 2.12
  - 31/07/03 - discussion sur les modes — Score : 2.04

- La communauté jaune (23 membres) semble particulièrement intéressée par la musique contemporaine et ses divers aspects. Les discussions se concentrent souvent sur l'enseignement de la musique, les œuvres actuelles, et les événements en présentiel tels que les concerts et les séminaires.

  - 23/06/03 - Indignation — Score : 4.46
    Débats sur l'enseignement de la musique et de l'appétence des jeunes envers la musique classique à notre époque.
  - 05/11/03 - Il est temps de réserver — Score : 2.45
    Annonce d'un concert
  - 03/06/03 - [Utilisateur] — Score : 2.00
    Deux utilisateurs se retrouvent sur musiSorbonne après avoir fait connaissance durant un séminaire.

- La communauté verte (23 membres) semble se distinguer par son rôle de questionnement et de réponse, abordant une grande variété de sujets. Les membres posent et répondent à des questions couvrant des domaines divers, allant des articles académiques et des livres aux questions technologiques. Ils recherchent souvent des clarifications, des conseils ou des informations sur des sujets spécifiques.

  - 02/04/03 - Beethoven dans la musique contemporaine — Score : 2.91
    Recherche de partitions
  - 02/04/03 - Question — Score : 2.26
  - 28/01/03 - Debussy, Golliwog — Score : 2.24
    Demande d'explication d'une oeuvre.
  - 08/09/03 - Carte son — Score : 1.73
    Questions sur les cartes son

- La communauté violette (30 membres), étant donné qu'il s'agit de la communauté composé d'utilisateurs n'ayant jamais eu d'interactions avec d'autres utilisateurs, ne publie que des messages d'annonce de concerts, de publications ou encore de colloques qui n'ont jamais eu de réponses.
  - 29/09/03 - annonce colloque
  - 05/10/03 - Rencontres Moyen-Age et Renaissance
  - 07/02/03 - prochain concert Musique en Sorbonne

## 4.4 Analyse des entités nommées

Nous avons tout d'abord recherché les entités nommés les plus communs à l'aide de la librairie spaCy[14], pour ainsi observer que les entités nommées les plus employées appartiennent à deux grandes catégories :

**1. Les lieux ou institution :**

- Paris - 422 occurrences

  On peut en déduire que de nombreuses références et annonces, de concerts par exemple, ou encore d'événements se déroulent à Paris.

- France - 167 occurrences

  On peut ici tirer une conclusions similaire, mais aussi peut-être comprendre que les thèmes, oeuvres et compositeurs discutés dans ce forum francophone sont liés à la France.

- Europe - 117 occurrences

  Ainsi, beaucoup de discussions ont un rapport avec l'Europe, le continent où se situe ce forum.

- Paris IV - 110 occurrences

  Paris IV étant l'ancien nom de la Sorbonne, il n'est pas étonnant de voir ce terme.

- Sorbonne - 107 occurrences

  Les activités académiques en lien avec ce forum se déroulent autour de la Sorbonne, université qui héberge ce forum.

**2. Les noms de compositeurs :**

- Beethoven - 137 occurrences

  A travers les annonces de concert mais aussi les discussions d'analyse musicale, plusieurs références sont en lien avec l'oeuvre de ce compositeur allemand.

- Bach - 131 occurrences

  Idem ici, aussi bien à travers les concert que via les discussions sur les sujets d'analyse ou de reconstitution de la performance historique, J.-S. Bach est naturellement cité à de nombreuses reprises dans les échanges.

- Mozart - 107 occurrences

  Pareil, Mozart est l'un des compositeurs les plus emblématiques, connu de tous.

- Berlioz - 72 occurrences

  Berlioz est le compositeur français le plus cité, et est une référence dans la composition musicale au même titre que les précédents compositeurs.

Certains résultats ont été omis afin de préserver l'anonymat des utilisateurs, et pour ne pas avoir des entités ne reflétant aucune tendance venant de signatures de couriel notamment.

Pour des analyses statistiques avancées et des visualisations, veuillez consulter le notebook Python.

# 5 Conclusion

Ce stage d'analyse de données textuelles musicologiques sur le forum musiSorbonne[2] pour l'année 2003 a permis d'explorer différentes facettes de la participation des utilisateurs, de leurs dynamiques de communication, et des thèmes de discussion abordés. Partant d'un jeu de données composé de près de 1500 messages, nous avons utilisé différentes librairies afin de proposer des axes analytiques. Après avoir chargé les 1486 messages venant de musiSorbonne à l'aide de pandas[4], nous avons tout d'abord créé une fonction afin de classer les messages par discussions, et d'identifier de manière unique chaque message.

On a ensuite fait des statistiques descriptives à l'aide de numpy[5]. On a ainsi déterminer qu'il y a 150 utilisateurs, qu'il y a 625 discussions en tout, dont 248 ayant plus d'un message. Le nombre moyen de messages par discussion est de 2.38, et cette moyenne monte a 4.47 quand on ne prend en compte que les discussions ayant plus d'un message. Les messages ont en moyenne 352 mots, et plus précisément 377 si on prend en compte les premiers messages des discussions, et 333 si on prend les autres messages.

Par la suite, on a dressé des graphiques à l'aide de matplotlib[6] afin d'analyser l'activité du forum selon l'horaire, le jour de la semaine, ainsi que selon le mois. On a ainsi déterminé que l'activité principale se déroule en journée, entre 9 heures et 23 heures, avec deux pics à 11h et 17h. De plus, on a également remarqué une inégalité marquée entre les deux semestres de l'année, l'activité étant beaucoup plus forte entre janvier et juillet, qu'entre août et décembre. Néanmoins, on n'a pas remarqué de différences notables entre les jours de la semaine.

De la même manière, on a dressé un graphique afin de représenter les disparités de publications entre les utilisateurs, pour remarquer qu'une minorité d'utilisateurs est à l'origine de la majorité des messages. Ainsi, Nicolas Meeùs[7] est l'auteur de 10% des messages, et les 11 utilisateurs ayant le plus parlé sont à l'origine de 50% des messages du forum.

On a ensuite eu l'occasion de faire de l'analyse de sentiments à l'aide du modèle VADER[8], proposé par la bibliothèque NLTK[9]. On a d'abord fait des mesures classiques, et déterminé qu'il y a 462 messages positifs, 544 messages neutres, ainsi que 480 messages négatifs. On a ensuite introduit quatre mesures de sentiments, le bilan sentimental, la moyenne sentimentale, le mouvement sentimental, ainsi que le mouvement sentimental moyen. Ainsi, on a déterminé que la moyenne sentimentale globale est de 0.01, et que le bilan sentimental final est de 14, ce qui est une bonne chose. On a ensuite déterminé les discussions présentant les mesures les plus hautes et les plus faibles.

Après cela, on a eu l'occasion de modéliser cette communauté sous la forme d'un graphe pondéré à l'aide de la bibliothèque NetworkX[11]. L'algorithme de Louvain[10] nous a permis d'exhiber 6 communautés ainsi que des tendances au sein de ces dernières. Ainsi, une communauté se distingue par l'usage de vocabulaire spécialisé et mathématique, abordant des concepts musicologiques complexes. Une autre a plutôt tendance à s'adonner à des débats animés et conflictuels sur des sujets divers. Une communauté parle surtout de l'histoire de la musique, ayant aussi de l'intérêt pour les autres cultures et pays. Une autre a tendance à parler de musique contemporaine, et la place actuelle de la musique, ainsi qu'aux événement musicaux. On a également une communauté composée d'utilisateurs posant et répondant à des questions en tout genre. Et enfin, la dernière communauté est constituée d'utilisateurs n'ayant jamais eu d'interactions avec d'autres utilisateurs, postant essentiellement des annonces sans réponses. On a pu visualiser cette communauté à l'aide de la bibliothèque pyvis[12].

Ce stage a donc permis de développer des compétences en analyse de données textuelles et en programmation, mais aussi de mettre en lumières des indicateurs sur le fonctionnement et les interactions au sein du forum. Les outils numériques se sont révélés indispensables pour traiter et analyser efficacement un volume important de données, offrant des perspectives intéressantes pour des études futures dans le domaine musicologique.

Il va de soi que cette étude prospective devra être complétée par une analyse plus large sur les 20 années d'existence du forum afin de mieux cerner les dynamique, et apprécier l'évolution des discours au sein de cette communauté scientifique.

# 6 Références

# References


[1] Python. https://www.python.org/.

[2] Meeùs, Nicolas. musiSorbonne. 31 décembre 2022, http://nicolas.meeus.free.fr/musiSorbonne.html.

[3] Project Jupyter. https://jupyter.org.

[4] pandas - Python Data Analysis Library. https://pandas.pydata.org/.

[5] NumPy -. https://numpy.org/.

[6] Matplotlib — Visualization with Python. https://matplotlib.org/.

[7] Meeùs, Nicolas. Nicolas Meeùs. 27/07/2018, http://nicolas.meeus.free.fr/.

[8] Hutto, C., et Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, n°1, mai 2014, p. 216-25. DOI.org (Crossref), https://doi.org/10.1609/icwsm.v8i1.14550.

[9] NLTK:: Natural Language Toolkit. https://www.nltk.org/.

[10] Blondel, Vincent D., et al. "Fast unfolding of communities in large networks". Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, n°10, octobre 2008, p. P10008. DOI.org (Crossref), https://doi.org/10.1088/1742-5468/2008/10/P10008.

[11] NetworkX — NetworkX documentation. https://networkx.org/.

[12] Interactive network visualizations — pyvis 0.1.3.1 documentation. https://pyvis.readthedocs.io/en/latest/.

[13] Newman, Mark. Networks. Oxford University Press, 2010. DOI.org (Crossref), https://doi.org/10.1093/acprof:oso/9780199206650.001.0001.

[14] spaCy · Industrial-Strength Natural Language Processing in Python. https://spacy.io/.