FaSEST

Economie Appliquée

ANALYSE TEXTUELLE DES OFFRES DE STAGE DANS LE DOMAINE DE LA DATA

Table des matières

Introduction

Dans un contexte où la transformation numérique s'accélère, les données occupent une place centrale dans la prise de décision au sein des entreprises. Les métiers liés à la data science, à l'analyse de données et à l'intelligence artificielle connaissent une forte croissance, et les offres de stage dans ce domaine se multiplient. Comprendre les compétences techniques et comportementales recherchées par les recruteurs devient ainsi un enjeu essentiel, tant pour les étudiants en formation que pour les institutions académiques qui adaptent leurs programmes à l'évolution du marché du travail.

L'analyse textuelle constitue un outil puissant pour explorer et structurer l'information contenue dans un grand nombre d'offres d'emploi. En extrayant les mots-clés les plus fréquents, les associations lexicales et les thématiques récurrentes, il est possible d'identifier les tendances dominantes du marché et d'obtenir une vision objective des attentes des recruteurs.

Le présent travail s'inscrit dans cette démarche. Il vise à analyser un corpus constitué de cent offres de stage dans le domaine de la Data, collectées sur la plateforme Linkedin . L'objectif principal est d'identifier les compétences techniques (hard skills) et comportementales (soft skills) les plus sollicitées, ainsi que les métiers et secteurs les plus représentés.

Pour répondre à ces questions, nous présenterons d'abord notre méthodologie de collecte et de traitement du corpus, puis nous analyserons le contenu lexical des offres pour identifier les compétences et métiers recherchés ainsi que les secteurs , avant de proposer une classification thématique permettant de structurer la diversité observée.

I. Constitution et préparation du corpus

1. Collecte et préparation des données

Dans cette étude, nous avons utilisé une base de données brute contenant 375 235 caractères. Nous avons recherché les publications d'offres sur https://www.linkedin.com/ en utilisant comme mots-clés : "stage data", et en filtrant selon la date de publication : le mois dernier. Pour chaque entreprise, nous avons sélectionné une seule offre et copié-collé son contenu afin de constituer notre base de données.

Pour préparer l'analyse lexicale des offres de stage, nous avons d'abord procédé au nettoyage et à la normalisation du texte. Le fichier contenant les offres a été lu et les étapes suivantes ont été réalisées : suppression de la ponctuation, des retours à la ligne et des caractères spéciaux, uniformisation des séparateurs d'offres, et conversion de l'ensemble du texte en minuscules.

Le texte a ensuite été découpé en offres individuelles, en filtrant les blocs trop courts (moins de 50 caractères), puis chaque offre a été séparée en titre (premier mot) et corps (reste du texte).

2. Traitement linguistique et construction des dictionnaires de mots-clés

Pour l'analyse lexicale, nous avons utilisé Stanza, un pipeline de traitement du langage naturel en français, afin d'extraire les mots, lemmes, catégories grammaticales et traits morphosyntaxiques. Les données ont été filtrées pour supprimer :

- les stopwords français courants (ex. : le, la, de, et, à, …),
- les auxiliaires et verbes génériques : être, avoir, faire, pouvoir, devoir, aller, vouloir, hésiter, venir, partir, rejoindre,
- les noms génériques : entreprise, action, propos, offre, mission,
- les adjectifs génériques : nouveau, grand, principal, possible, divers, fort,
- ainsi que tous les nombres.

Ensuite nous avons extrait les lemmes les plus fréquentes par catégorie grammaticale (noms, verbes et adjectifs) afin de constituer la base de données finale pour nos analyses.

Pour l'identification des différentes catégories (compétences techniques, qualités personnelles, métiers et secteurs),nous avons construit des dictionnaires de mots-clés issus de l'observation directe du corpus et de ce que l'on rencontre couramment dans les offres d'emploi. Ces listes ont été élaborées à partir des termes les plus récurrents, puis enrichies manuellement par le regroupement de synonymes et de variantes orthographiques (par exemple Power BI/PowerBI, machine learning/apprentissage automatique). Cette approche a permis de repérer efficacement les occurrences pertinentes et d'obtenir une vision structurée des éléments recherchés par les recruteurs.

3. Statistiques descriptives du corpus

Résumé des statistiques textuelles des offres

Statistiques	Valeur
Nombre d'offres	100
Nombre total de mots	31720
Nombre de lemmes uniques	4682
Nombre de NOUN	15596
Nombre de VERB	4364
Nombre de ADJ	3704
Nombre de PROPN	3019
Nombre de PUNCT	2439
Nombre de X	1834

Tableau 1 : Résumé des statistiques textuelles des offres

Le corpus contient 31 720 mots répartis sur 100 offres, avec 4 682 lemmes uniques. Les noms (15 596) dominent largement, suivis des verbes (4 364) et des adjectifs (3 704), ce qui reflète une forte densité descriptive. La présence de 3 019 noms propres et 1 834 mots non reconnus indique un usage fréquent de références techniques et institutionnelles.

II. Exploration lexicale des offres

1. Mots les plus fréquents par catégorie (verbes, noms, adjectifs)

Tableau 2 : Top 5 des mots les plus fréquents par catégorie dans les offres de stage en data

Mot	Catégorie	Occurrences
participer	Verbe	105
développer	Verbe	102
travailler	Verbe	98
contribuer	Verbe	78
accompagner	Verbe	77
technique	Adjectif	88
commercial	Adjectif	86
international	Adjectif	74
financier	Adjectif	62
interne	Adjectif	56
équipe	Nom	313
donnée	Nom	309
analyse	Nom	212
projet	Nom	180
client	Nom	173

L'analyse des verbes, adjectifs et noms les plus fréquents dans les offres de stage en data permet de dégager les attentes implicites des recruteurs et de mieux comprendre le profil recherché.

Verbes d'action : une posture proactive attendue Les verbes tels que « participer », « développer », « travailler », « contribuer » et « accompagner » indiquent que les stagiaires sont appelés à jouer un rôle actif dans les projets. Ils ne sont pas cantonnés à des tâches d'observation, mais sont intégrés comme contributeurs à part entière. Cela traduit une volonté de responsabiliser les stagiaires et de les impliquer dans des missions concrètes.

Adjectifs : des profils polyvalents et rigoureux Les adjectifs les plus fréquents, comme « technique », « commercial », « international », « financier » et « interne », montrent que les recruteurs recherchent des candidats capables d'évoluer dans des environnements variés. La dimension technique est essentielle, mais elle s'accompagne souvent d'une compréhension des enjeux économiques, commerciaux ou organisationnels. Cela reflète une attente de polyvalence et de sérieux.

Noms : un environnement collaboratif et orienté projet Les noms tels que « équipe », « donnée », « analyse », « projet » et « client » soulignent les éléments centraux du travail en data. Le stage s'inscrit dans une dynamique de collaboration, autour de projets concrets, avec une forte orientation vers l'analyse de données et la réponse à des besoins clients. Ces termes traduisent une approche opérationnelle et collective du travail.

Ce top lexical met en évidence les grandes lignes du profil attendu pour un stage en data : un candidat engagé, techniquement compétent, capable de comprendre les enjeux métiers et de s'intégrer dans une équipe projet. L'analyse linguistique des offres permet ainsi de mieux cerner les attentes des recruteurs et d'adapter les candidatures en conséquence.

2. Analyse des combinaisons de mots : bigrams et trigrammes

Pourquoi analyser les bigrammes et trigrammes est intéressant?

Passer à l'analyse des bigrammes (groupes de deux mots) et trigrammes (groupes de trois mots) permet d'aller bien au-delà des simples mots isolés. Alors qu'un comptage de mots uniques indique quelles compétences ou qualités sont mentionnées, il ne montre pas le contexte ou les combinaisons significatives. Par exemple :

- Le mot "Power" tout seul ne dit pas grand-chose, mais le bigramme "Power BI" indique clairement qu'on parle d'un outil spécifique très recherché.
- De même, "analyse" est vague, mais le trigramme "capacité analyse synthèse" révèle que les recruteurs valorisent la capacité à analyser et synthétiser des informations complexes.

En résumé, l'analyse en n-grammes permet de comprendre les compétences techniques exactes, le contexte d'utilisation des compétences, et même des valeurs ou exigences comportementales implicites dans les offres.

Ce que nous avons trouvé :

- Compétences techniques et outils : Les combinaisons de mots les plus fréquentes, telles que “power bi”, “tableau bord”, “machine learning” et “data science”, indiquent que les recruteurs mettent fortement l’accent sur la maîtrise des outils de Business Intelligence et sur les compétences en analyse de données et intelligence artificielle. Cela confirme la centralité de la technique dans les offres de stage.

- Profil académique ciblé : La présence récurrente des trigrammes “école ingénieur commerce” ou “école commerce université” montre que certaines entreprises ciblent explicitement des étudiants de formations spécifiques, souvent issues de grandes écoles. L’analyse révèle ainsi une préférence marquée pour ces profils, ce qui exclut de fait une partie des candidats venant de l’université, pourtant tout aussi compétents. Cela soulève une véritable question d’équité dans l’accès aux opportunités de stage.

- Environnement de travail et compétences comportementales : Des bigrammes comme “sein équipe” ou “environnement travail” et des trigrammes tels que “capacité analyse synthèse” indiquent que les recruteurs valorisent à la fois le travail collaboratif et la capacité à traiter et synthétiser l’information.

- Égalité et non-discrimination : Les trigrammes “garantir égalité chance” et “religion orientation sexuel” révèlent un souci explicite des entreprises d’inclure des mentions légales ou des engagements sur la diversité et l’égalité, ce qui peut être interprété comme un signal de l’importance des valeurs sociétales dans le recrutement.

En définitive, l’analyse des bigrammes et trigrammes révèle les attentes précises des recruteurs en matière de compétences techniques, comportementales et de valeurs. Mais elle met aussi en lumière une préférence marquée pour les profils issus de grandes écoles, au détriment des étudiants universitaires pourtant tout aussi qualifiés. Ce constat soulève une vraie question d’équité dans l’accès aux stages.

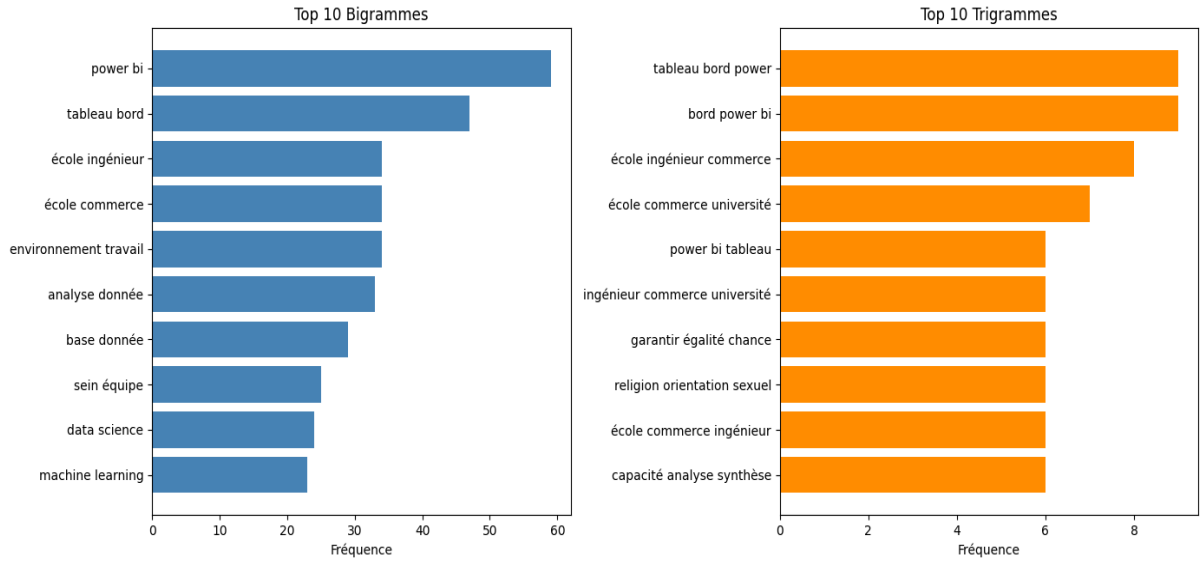

Figure 1 : Liste des dix bigrammes et des dix trigrammes les plus fréquentes

III. Cartographie des compétences et profils recherchés

1. Hard skills (compétences techniques)

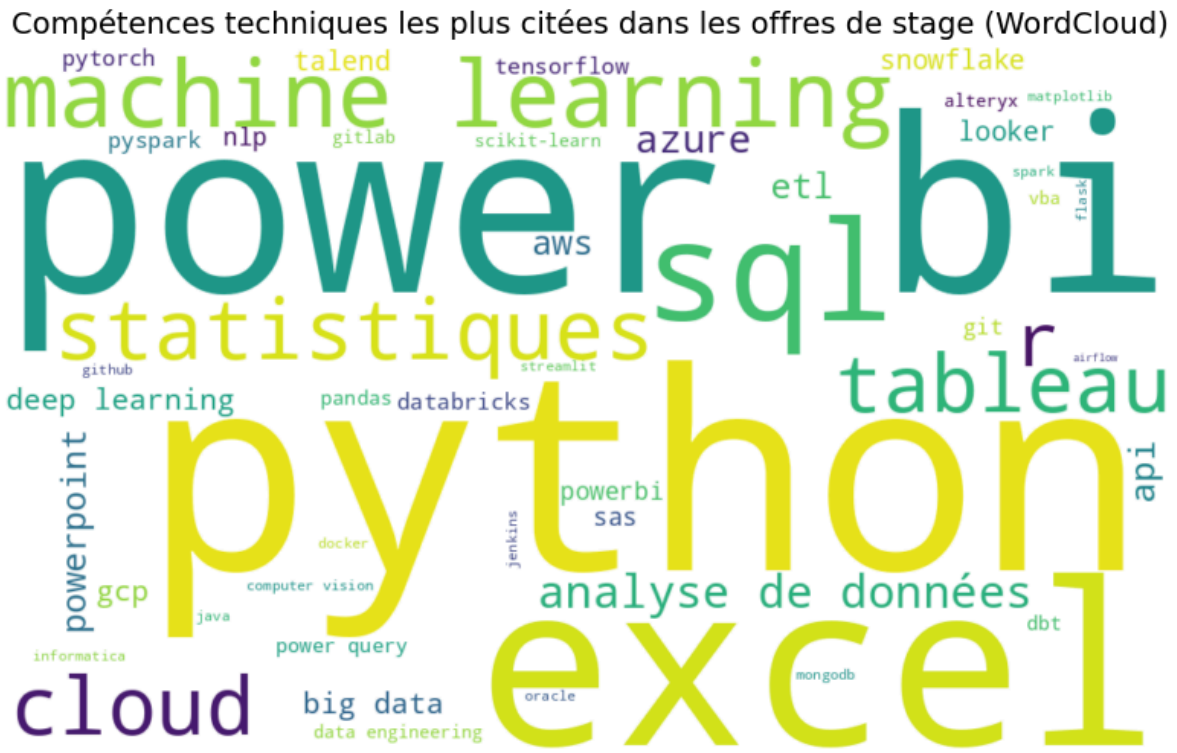

Figure 2 : Nuage de mots des compétences techniques

Ce nuage des mots montre les compétences techniques les plus citées dans les offres de stage en data :

- Les langages et outils les plus demandés sont Python, Power BI, Excel et SQL, ce qui confirme que la maîtrise des outils de programmation et d'analyse de données est essentielle.
- On retrouve également des compétences liées au Machine Learning, aux statistiques et au traitement des données (analyse de données, ETL), qui reflètent la dimension analytique et technique des stages.
- Des outils de cloud et d'infrastructure comme Azure, AWS et Cloud apparaissent aussi, soulignant l'importance des environnements cloud et Big Data.
- Enfin, des logiciels complémentaires comme Tableau ou PowerPoint montrent que la visualisation et la communication des résultats sont également valorisées.

En résumé, il met en évidence les compétences techniques clés recherchées pour un stage en data, alliant programmation, analyse, Machine Learning et outils de visualisation.

2. Soft skills (qualités personnelles)

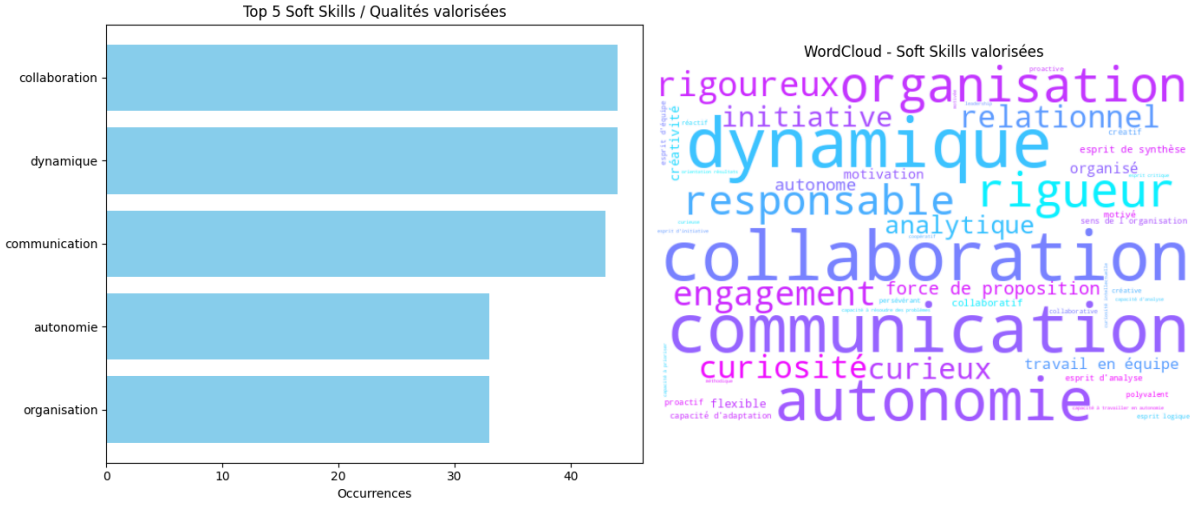

Figure 3 : Les cinq qualités personnelles les plus valorisées et leur représentation sous forme de nuage de mots

La figure 3 illustre les qualités personnelles les plus valorisées, identifiées à partir de l'analyse du corpus. Deux représentations sont proposées : un graphique en barres des cinq compétences les plus citées, et un nuage de mots offrant une vue élargie des qualités recherchées.

Les résultats montrent que les recruteurs privilégient des profils capables de travailler en équipe (collaboration), de faire preuve d'enthousiasme et de proactivité (dynamisme), de communiquer efficacement (communication), de gérer leurs tâches de manière autonome (autonomie) et d'organiser leur travail avec rigueur (organisation).

Le nuage de mots complète cette analyse en mettant en évidence d'autres qualités fréquemment mentionnées telles que rigueur, initiative, responsabilité ou encore créativité, soulignant l'importance de la fiabilité, de l'engagement personnel et de la capacité à s'adapter.

Ces éléments confirment que les soft skills constituent un critère déterminant dans l'évaluation des candidatures, au même titre que les compétences techniques. Ils traduisent une attente forte en matière de savoir-être, de coopération et de gestion autonome dans les environnements professionnels contemporains.

3. Métiers et secteurs dominants

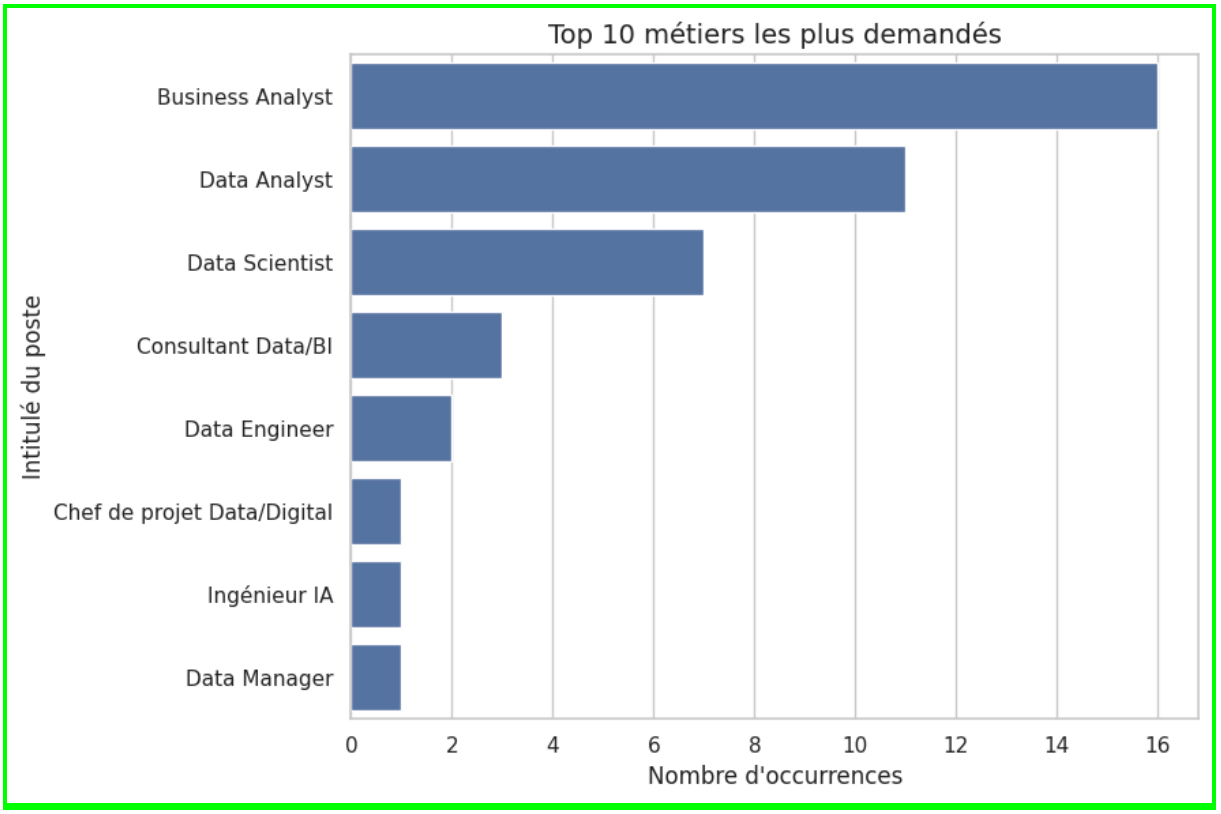

Figure 4 : Les dix métiers les plus demandés

Quels sont les profils les plus recherchés ?

L'analyse des intitulés de poste extraits du corpus révèle une forte concentration autour des métiers de l'analyse et de la data. Le métier de Business Analyst arrive en tête avec 16 occurrences, suivi par Data Analyst (11) et Data Scientist (7). Ces résultats confirment la prédominance des profils orientés vers l'analyse de données, la modélisation et l'aide à la décision dans les offres de stage.

On observe également la présence de métiers plus spécialisés comme Consultant Data/BI, Data Engineer, ou encore Chef de projet Data/Digital, bien que dans des proportions plus modestes. Cela suggère que les entreprises recherchent majoritairement des profils opérationnels capables de manipuler et interpréter les données, tout en laissant une place aux fonctions techniques ou stratégiques.

La figure 5 indique les secteurs les plus recruteurs dans les offres de stage. Nous constastons :

- Le secteur Banque & Assurance domine avec 32 offres, soit près d'un tiers du total.
- Les secteurs Retail / Commerce (19 offres) et Industrie (15 offres) suivent à distance.
- Les autres secteurs, comme Santé, Énergie et Tech, sont moins représentés, tandis que Conseil & Services et Autre comptent très peu d'offres.

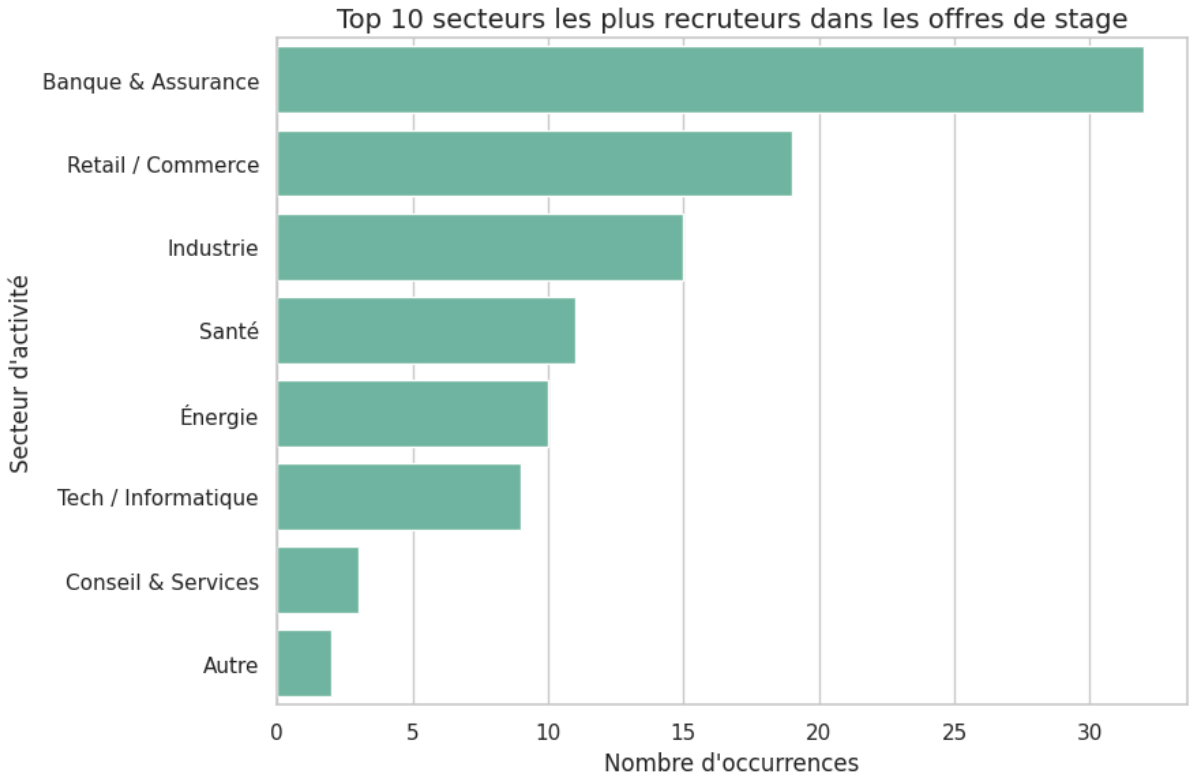

Figure 5 : Les dix secteurs les plus recruteurs

IV. Classification thématique et segmentation des offres

1. Méthode TF-IDF et K-means

Nous avons utilisé la vectorisation TF-IDF pour transformer chaque document en vecteur numérique : cela permet de pondérer les mots selon leur importance, les mots fréquents dans une offre mais rares dans l'ensemble du corpus ayant plus de poids. Avec ces vecteurs, nous avons appliqué KMeans pour regrouper les offres similaires. Le choix du nombre de clusters s'est fait avec la méthode du coude, en regardant l'inertie selon k et en vérifiant le silhouette score, ce qui a montré que k=2 était le meilleur compromis pour séparer les offres en deux groupes distincts.

Nous avons ensuite visualisé les résultats avec un graphique double : à gauche la courbe de l'inertie et à droite la répartition des offres par cluster, avec le nombre exact d'offres indiqué. Enfin, nous avons fait une analyse lexicale pour chaque cluster en calculant la moyenne des vecteurs TF-IDF et en sélectionnant les 15 mots les plus représentatifs, ce qui a permis de caractériser les clusters.

2. Résultats du clustering

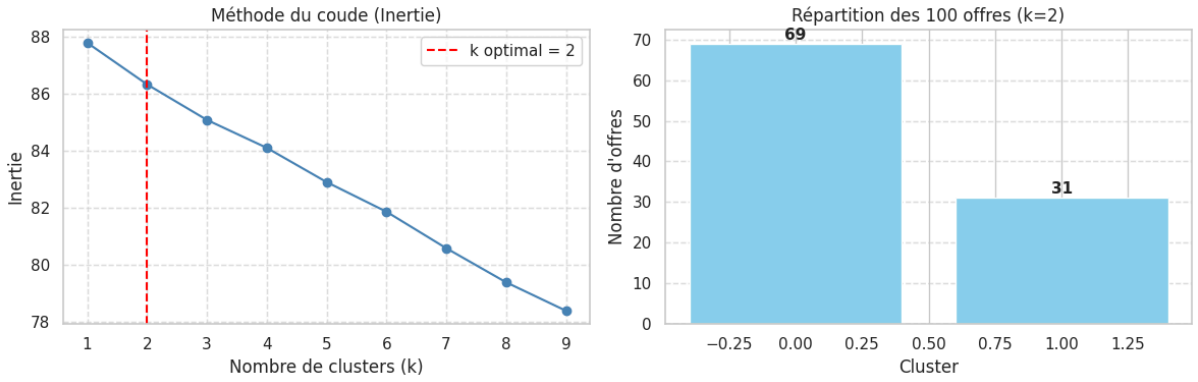

Figure 6 : Méthode du coude et la répartition des offres par clustering

L'analyse par clustering KMeans des offres de stage en data a permis de segmenter le corpus en deux clusters principaux, mettant en évidence des orientations différentes des missions proposées :

Tableau 3 : Top 15 des mots les plus fréquents par clustering

	Mot 1	Mot 2	Mot 3	Mot 4	Mot 5	Mot 6	Mot 7	Mot 8	Mot 9	Mot 10	Mot 11	Mot 12	Mot 13	Mot 14	Mot 15
Cluster 0	donnée	équipe	client	projet	modèle	solution	outil	compétence	analyse	technique	groupe	métier	développer	service	participer
Cluster 1	commercial	analyse	équipe	marque	donnée	outil	direction	performance	vente	groupe	financier	développement	environnement	international	maison

Cluster 0 – Stages techniques et projets data (69 offres)

Les mots les plus fréquents dans ce cluster sont : donnée, équipe, client, projet, modèle, solution, outil, compétence, analyse, technique, groupe, métier, développer, service, participer.

Interprétation :

- Ces offres correspondent à des stages fortement orientés technique et opérationnel, où l'étudiant sera impliqué dans la modélisation de données, l'analyse statistique et la participation à des projets concrets.
- L'accent est mis sur le travail en équipe et la contribution active à des missions structurées, typique des stages orientés data science ou data analytics classiques.

Cluster 1 – Stages orientés business et data appliquée (31 offres)

Les mots les plus fréquents dans ce cluster sont : commercial, analyse, équipe, marque, donnée, outil, direction, performance, vente, groupe, financier, développement, environnement, international, maison.

Interprétation :

- Ces offres restent des stages en data, mais avec une orientation plus business, marketing ou finance, où l'étudiant utilisera la data pour soutenir la stratégie, le suivi des performances ou l'analyse commerciale.
- Les compétences techniques sont présentes (analyse, donnée, outil), mais au service d'une application métier.

En définitive,

- La majorité des stages (Cluster 69%) sont techniques, impliquant directement des missions de data analysis, modélisation et développement.
- La partie (Cluster 31%) combine la data avec des applications business ou stratégiques.
- Le clustering montre que même dans un corpus homogène de stages en data, il existe une distinction nette entre missions purement techniques et missions appliquées au business, ce qui peut guider les étudiants selon leurs intérêts et compétences.

Conclusion

Cette étude a permis d'explorer en profondeur les tendances lexicales et thématiques des offres de stage dans le domaine de la data. À partir d'un corpus de cent annonces collectées sur LinkedIn, nous avons mis en œuvre une méthodologie rigoureuse combinant nettoyage du texte, lemmatisation, extraction des catégories grammaticales et analyse statistique des occurrences.

Les résultats ont montré que les compétences techniques les plus valorisées concernent la maîtrise des langages et outils de traitement et d'analyse de données (Python, SQL, Power BI, Excel), ainsi que des notions de Machine Learning et d'infrastructures cloud (Azure, AWS). Ces exigences traduisent le besoin croissant d'automatisation et d'exploitation intelligente des données dans les organisations.
Du côté des soft skills, les recruteurs insistent sur la rigueur, la curiosité, l'autonomie, la motivation et la capacité à travailler en équipe. Ces qualités confirment que le profil idéal du stagiaire en data doit allier savoir-faire technique et compétences relationnelles.

L'analyse des métiers montre la prédominance des postes de Business Analyst, Data Analyst et Data Scientist, essentiellement concentrés dans les secteurs Banque & Assurance, Retail et Industrie. Cette distribution illustre la place stratégique qu'occupe aujourd'hui la donnée dans la prise de décision et le pilotage des performances.

L'application de la vectorisation TF-IDF et du clustering K-means a ensuite permis de segmenter le corpus en deux grands groupes d'offres : les stages à dominante technique, centrés sur l'analyse, la modélisation et le développement de données, et les stages orientés business, où la data est utilisée pour soutenir des activités stratégiques ou commerciales.

Cependant, l'étude présente certaines limites.
D'abord, le corpus est restreint à cent offres collectées sur une période courte et sur une seule plateforme (LinkedIn), ce qui ne permet pas de généraliser entièrement les résultats à l'ensemble du marché. De plus, l'analyse s'est concentrée sur le français, alors que certaines offres contiennent des passages en anglais, parfois non traités par les outils de lemmatisation. Enfin, l'approche lexicale ne prend pas en compte les nuances contextuelles ni la polysémie de certains termes, ce qui limite la portée interprétative des résultats.

Malgré ces limites, cette recherche met en évidence l'intérêt de l'analyse textuelle appliquée aux offres d'emploi pour comprendre les attentes du marché et orienter la formation des étudiants. Elle constitue une base utile pour de futures études plus larges, intégrant d'autres langues, périodes temporelles ou sources de données.
En somme, la combinaison de l'approche linguistique et statistique offre une lecture fine des besoins en compétences dans la data et ouvre la voie à une meilleure adéquation entre la formation et les exigences professionnelles.