

Louvain School of Management

Étude de la mise en avant des produits durables dans les e-boutiques engagées

Rapport Web Mining (MLSMM2153)

Auteurs : FERON Clément, NICELLI Romain et SIMOENS Mathias

Professeur : VANDE KERCKHOVE Corentin et COURTAIN Sylvain

Année académique : 2025-2026

LSM Master 1 : Ingénieur de gestion à finalité spécialisée - Business Analytics

Table des matières

1	Introduction	3
2	Collecte des données (Scraping)	3
2.1	Sélection des sources	3
2.2	Outils et environnements techniques	4
2.3	Stratégie de couverture	4
2.4	Nettoyage et préparation	5
2.5	Volume des données	6
3	Text Mining	6
3.1	Prétraitement et normalisation du corpus	6
3.2	Vectorisation TF-IDF	7
3.3	Clustering	8
3.4	Détermination du nombre de clusters	8
3.5	Analyse des résultats et biais inter-sites	8
3.6	Visualisation par réduction de dimension	9
3.7	Mesure de la distance sémantique : la « Similarité Cosinus »	9
3.8	Analyse sémantique orientée par catégories lexicales	9
3.9	Définition des axes lexicaux	10
3.10	Métrique et comparaison inter-sites	10
3.11	Vers une classification par dominance	10
3.12	Calcul de similarité entre rapports ESG et descriptions produits	11
4	Link Analysis	11
4.1	Méthodologie	12
4.2	Choix des métriques et justification	12
4.3	Analyse des résultats obtenus avec Patagonia	12
4.3.1	Analyse de la centralité de degré (Degree Centrality)	13
4.3.2	Analyse des plus courts chemins (Shortest Path)	13
4.3.3	Analyse de la centralité d'intermédiarité (Betweenness Centrality)	13
5	Conclusion	13
Annexes		15
Annexe 1 – Structure des fichiers Excel	15	
Annexe 2 – Répartition hommes/femmes	16	
Annexe 3 – Matrices TF-IDF	17	
Annexe 4 – Nuages de mots	18	
Annexe 5 – Méthode du coude	25	

Annexe 6 – Clustering	27
Annexe 7 – PCA	29
Annexe 8 – Statistiques de similarité	31
Annexe 9 – Dictionnaires lexicaux	33
Annexe 10 – Distribution lexicale	35
Annexe 11 – Similarité rapports ESG	36
Annexe 12 – Degree Centrality	38
Annexe 13 – Shortest Path	39
Annexe 14 – Betweenness Centrality	40
Annexe 15 – Gephi	41

1 Introduction

La montée en puissance des enjeux environnementaux et sociaux a profondément transformé les pratiques de consommation, en particulier dans le domaine du commerce en ligne. Face à une demande croissante pour des produits respectueux de l'environnement, de nombreuses marques se positionnent aujourd'hui comme éthiques, responsables ou durables, déployant un effort communicationnel important pour valoriser leurs engagements écologiques directement au sein des pages produits.

Ce projet, réalisé dans le cadre du cours de *Web Mining (MLSMM2153)*, mobilise une approche *data-driven* pour analyser la communication durable des e-boutiques engagées. Il s'appuie sur les trois piliers méthodologiques du cours magistral : la collecte automatisée de données (*Scraping*), l'analyse textuelle (*Text Mining*) et l'analyse de graphes (*Link Analysis*).

L'objectif de ce projet est d'analyser la manière dont les principaux acteurs du e-commerce durable mettent en valeur leurs produits responsables à travers leurs descriptions en ligne, ainsi que d'identifier les éventuelles convergences avec les messages et pratiques présentés dans leurs rapports ESG.

Voici nos questions de recherche :

- Quels sont les marqueurs sémantiques dominants mobilisés par les marques éthiques pour valoriser leurs produits durables ?
- Quelle est la distance sémantique entre le discours institutionnel de la marque et les descriptions individuelles de produits ?
- - Peut-on, à partir des similarités textuelles entre les descriptions de produits, construire un graphe permettant d'identifier des relations et des regroupements entre produits similaires ? Quels produits occupent des positions centrales ?

Voici le lien vers notre répertoire GitHub : [Web Mining 2025-2026 Groupe 6](#) (Dossier .main)

2 Collecte des données (Scraping)

2.1 Sélection des sources

Notre corpus repose sur l'analyse de trois e-boutiques internationales reconnues pour leur positionnement éthique dans le secteur de la mode durable : [Patagonia](#) (États-Unis), [Ecoalf](#) (Espagne) et [Armedangels](#) (Allemagne).

Ce choix repose sur trois critères majeurs :

- Premièrement, les trois sites présentent une architecture comparable permettant une comparaison pertinente (à savoir des pages produits détaillées et des rapports sur leurs efforts en matière de développement durable). De plus, au vu de leurs catalogues importants, cela nous garantissait un volume de données suffisant.

- Deuxièmement, ces marques sont 3 des acteurs les plus reconnus de la mode éthique mais leurs positionnements se distinguent les uns des autres. Tandis que Patagonia incarne un activisme environnemental radical avec une transparence poussée (certification B Corp), Ecoalf, lui, se concentre plus sur l'économie circulaire en valorisant l'innovation dans le recyclage des matériaux. Enfin, Armedangels privilégie une approche "slow fashion" axée sur les certifications sociales et la traçabilité complète. Une comparaison de ces 3 manières de penser l'éco-responsabilité constitue donc un grand intérêt.
- La troisième et dernière raison est le fait que tous ces sites proposent des versions anglophones complètes, garantissant une cohérence linguistique pour les analyses ultérieures.

Toutefois, ces sites intègrent des protections techniques contre le scraping automatisé (détection de bots). Ceci nous a donc forcé à utiliser des outils de navigation simulée comme *Playwright* pour contourner ces restrictions.

2.2 Outils et environnements techniques

Afin de simuler un comportement utilisateur réel et contourner les mécanismes de détection de bots mis en place par les boutiques analysées nos scripts python s'appuient sur la bibliothèque *Playwright*, utilisée via son interface asynchrone. Cet outil permet d'automatiser un navigateur, de déclencher le chargement complet des pages et d'interagir avec les éléments dynamiques tel que le défilement. Son utilisation est justifiée par les mécanismes de protection contre le scraping automatisé présents sur les sites analysés, ainsi que par la structure dynamique des pages de catégories et de produits.

La gestion asynchrone du processus de collecte repose sur la bibliothèque *asyncio*, permettant d'optimiser le temps d'exécution des scripts et de gérer efficacement l'ouverture successive de multiples pages web. Cette approche améliore les performances globales de collecte tout en conservant un contrôle précis sur le déroulement des différentes étapes.

Une fois le code HTML entièrement chargé par le navigateur automatisé, la bibliothèque *BeautifulSoup* (module bs4) est utilisée pour l'analyse et le parsing du contenu. Elle permet de parcourir la structure semi-structurée des documents HTML et d'extraire les éléments textuels pertinents à savoir le nom du produit, son « breadcrumb » associé et sa description complète. Le « breadcrumb » ou fil d'Ariane en français reprend la série de liens hiérarchiques depuis la page d'accueil jusqu'au produit (*ex : Men's > Shop by Category > Fleece > Jackets > Nom du produit*), fournissant la liste des catégories associées.

Enfin, nous avons eu recours à la bibliothèque *pandas* pour structurer, stocker et manipuler les données extraites. Elle permet de transformer les informations collectées en tableaux structurés facilitant l'export vers des fichiers Excel et la préparation du corpus pour les étapes ultérieures d'analyse textuelle.

2.3 Stratégie de couverture

Dans ce projet, la collecte repose sur une stratégie contrôlée en deux étapes, combinant l'exploration large d'une page de catégorie et le traitement exhaustif d'une liste d'URL produits. Dans un premier

temps, deux scripts dédiés aux sections hommes et femmes chargent intégralement les pages de catégorie correspondantes. Le script déclenche un défilement automatique de la page et passe à la page suivante jusqu'à ce qu'aucun nouveau produit ne soit ajouté. Une fois le chargement terminé, tous les liens vers les pages produits sont sauvegardés dans des fichiers texte. Cette approche correspond à une couverture de type breadth-first limité à un seul niveau de profondeur.

Dans un second temps, un script de scraping détaillé lit la liste d'URL produits contenue dans le fichier texte et parcourt chaque lien. Pour chaque URL, le script télécharge la page produit puis extrait les informations qui vont nous servir (nom, « breadcrumb » et description). Les enregistrements obtenus sont ensuite stockés dans un fichier Excel (voir annexe 1). Pour chaque boutique, nous avons isolé des champs spécifiques. Pour Armedangels, nous avons ciblé la section « Overview » (lorsqu'elle est disponible), ainsi que les blocs « Material and Care » et « Transparency ». Pour Patagonia, l'extraction s'est concentrée sur les sections « Intro », « Specs and Features » et « Material and Care instruction ». Chez Ecoalf, nous avons privilégié les champs « Products Details » et « Sustainability Report ». Bien que les sections portent des intitulés différents selon les sites, elles remplissent une fonction discursive équivalente à savoir la description des caractéristiques produits et la mise en avant des engagements en matière de durabilité. Ce choix de section permet d'obtenir des données sémantiquement similaires et comparables entre elles.

Le stockage intermédiaire des liens dans un fichier texte avant l'extraction détaillée offre une sécurité. Il permet de tester et de modifier les traitements ultérieurs sans avoir à relancer la phase de navigation complète ce qui représente un gain de temps considérable.

En parallèle, un troisième script récupère des rapports en matière d'éco-responsabilité publiés par chaque marque. Ces documents PDF sont téléchargés, convertis en texte brut et stockés dans des fichiers texte distincts par marque, complétant le corpus nécessaire à notre analyse.

NB : Les codes python de notre partie scraping étant similaires entre nos 3 marques, nous avons décidé de commenter uniquement ceux consacrés à l'entreprise Patagonia.

2.4 Nettoyage et préparation

Certains produits entre hommes et femmes étant similaires, lorsque nous fusionnons les 2 fichiers Excel obtenus pour chaque marque, nous obtenons des doublons (*fichiers « all_site_products.xlsx »*). Dès lors, nous avons créé un fichier python pour à la fois les supprimer mais également pour nettoyer une première fois les descriptions :

- Standardisation de la casse : L'intégralité du corpus a été convertie en minuscules. Cette étape est cruciale pour éviter la duplication des entrées dans notre matrice de termes (par exemple, pour que "Coton" et "coton" soient traités comme une entité sémantique unique).
- Filtrage du bruit : Nous avons supprimé les caractères spéciaux, la ponctuation et les valeurs numériques. Le choix de collecter des données issues de sites anglophones s'avère ici stratégique.

La langue anglaise, utilisant le code *ASCII* standard sans accentuation complexe (contrairement au français), minimise les risques d'erreurs d'encodage et facilite drastiquement le nettoyage syntaxique.

- Nettoyage des espaces : Les espaces multiples consécutifs sont normalisés en un seul espace via des expressions régulières et les espaces en début/fin de chaîne sont supprimés.
- Suppressions des doublons et lignes vides : Les produits dupliqués sont identifiés sur la base du champ "name" et seule la première occurrence est conservée. Les descriptions dont la longueur nettoyée est inférieure à 3 caractères sont également éliminées.
- Standardisation des colonnes : Les noms de colonnes sont uniformisés en minuscules et débarrassés des espaces inutiles.

Tout cela nous permet d'obtenir les fichiers « site_dataset.xlsx » pour chaque marque.

2.5 Volume des données

Au total, 2197 produits uniques ont été collectés, constituant un corpus suffisamment large pour une analyse automatisée.

3 Text Mining

Suite à la phase de nettoyage de la collecte, nous disposons donc d'un corpus composé de 2197 descriptions de produits (*répartis entre Patagonia, Ecoalfe et Armedangels*) structurées sous format tabulaire dans un fichier Excel ainsi que 3 rapports ESG convertis en fichiers texte. Bien que nettoyées des balises HTML et autres artefacts de scraping, les descriptions textuelles des produits demeurent des données non structurées, « bruitées » et de haute dimensionnalité. Pour rendre ces données exploitable dans une optique analytique, nous avons mis en place un pipeline de transformation complète.

Cette section présente la chaîne de traitement du langage naturel (*NLP*) appliquée au corpus. Nous détaillons successivement le prétraitement et la normalisation des textes (*tokenisation, lemmatisation, filtrage des stopwords*), la vectorisation par pondération *TF-IDF* puis les trois approches analytiques déployées : l'analyse fréquentielle par nuages de mots, le clustering k-Means (*avec détermination empirique du nombre optimal de clusters et visualisation par réduction de dimension PCA*). Pour finir, l'analyse sémantique par dictionnaires lexicaux pour quantifier les stratégies de communication des marques. La métrique de similarité cosinus, que nous avons retenue pour mesurer la proximité sémantique entre produits, leurs descriptions et les discours institutionnels, est également justifiée.

3.1 Prétraitement et normalisation du corpus

L'objectif de cette phase est de réduire la dimensionnalité du vocabulaire en ne conservant que les tokens porteurs de sens. Notre approche, implémentée en Python via la bibliothèque *NLTK*, procède selon les étapes suivantes :

- Élimination des mots vides : Afin de réduire le bruit sémantique, nous avons filtré les mots très fréquents mais peu informatifs (*articles*, *prépositions*, *pronoms*). Notre méthodologie de filtrage est hybride et itérative. Nous avons d'abord utilisé la liste standard de stopwords anglais de NLTK, puis généré une liste complémentaire via Intelligence Artificielle générative pour identifier des termes contextuellement non pertinents dans le domaine de la mode (*exemple* : "click", "shipping", "details"). Enfin, suite à plusieurs itérations d'analyse, nous avons manuellement enrichi cette liste en identifiant des termes récurrents polluant les résultats sans apport sémantique.
- Normalisation (*Lemmatisation vs Racinisation*) : Pour consolider le vocabulaire, notre architecture de code permet une flexibilité dans le choix de la normalisation. L'utilisateur peut opter pour la méthode la plus adaptée aux résultats observés (par exemple, la lemmatisation pour conserver la racine lexicographique valide, ou la racinisation pour une réduction plus agressive). Nous avons privilégié la lemmatisation pour nos analyses finales afin de garantir l'intégrité du vocabulaire technique lié à la durabilité. Contrairement à la racinisation qui tronque les mots (*ex* : "recycl"), la lemmatisation préserve des termes réels (*ex* : "recycled"), assurant une visibilité professionnelle des nuages de mots et une précision sémantique pour le clustering.
- Tokenisation et N-grams : La segmentation du texte a été effectuée via le tokenizer *NLTK* avec une granularité configurable. En effet, notre architecture Python permet de sélectionner la granularité d'analyse entre unigrams (*analyse mot par mot*), bigrams et trigrams (*paires ou triplets de mots*). L'approche par bigrams s'est révélée particulièrement pertinente pour capturer les expressions composées fréquentes dans la mode durable (*ex* : "organic cotton", "recycled polyester") qui perdraient leur sens si elles étaient séparées. En revanche, pour la construction du graphe dans Gephi, nous avons opté pour les unigrams afin de simplifier la représentation visuelle et limiter la complexité du réseau.

3.2 Vectorisation TF-IDF

Une fois le texte nettoyé et tokenisé, il est nécessaire de le transformer en une représentation numérique vectorielle. Plutôt que d'utiliser une simple matrice de fréquence (*Bag-of-Words*) qui donnerait trop de poids aux mots génériques, nous avons opté pour la méthode *TF-IDF* (*Term Frequency-Inverse Document Frequency*). Cette métrique statistique permet d'évaluer l'importance d'un terme contenu dans une description par rapport à l'ensemble du corpus.

Dans le contexte de la mode durable, le vocabulaire est relativement restreint et technique. Des mots comme "t-shirt" ou "wear" apparaissent dans presque tous les documents et seraient jugés importants par une simple analyse de fréquence. À l'inverse, le TF-IDF pénalise ces termes omniprésents, via le logarithme de l'inverse de leur fréquence documentaire, tout en valorisant les termes plus rares et discriminants comme "hemp", "upcycled" ou "tencel". Cela permet de faire émerger la spécificité de chaque produit au sein d'un corpus thématiquement homogène. Nos matrices *TD-IDF* sont disponibles à l'annexe 3.

En complément de la vectorisation *TF-IDF*, nous avons généré des nuages de mots (voir annexe 4)

reposant sur le décompte brut des occurrences après nettoyage. Cette visualisation offre un aperçu immédiat du vocabulaire dominant à l'échelle de chaque marque avant toute pénalisation des termes génériques, permettant d'identifier les champs lexicaux privilégiés dans la communication des produits.

Nous avons délibérément choisi de concentrer nos analyses sur les unigrams et les bigrams, car ils offrent le meilleur équilibre entre fréquence et richesse sémantique. Les unigrams permettent de capter les thèmes globaux tandis que les bigrams identifient les matières et labels composés. A l'inverse, les trigrams se sont révélés peu pertinents en raison de leur grande spécificité et tendent à fragmenter l'information nuisant à la lisibilité des nuage de mots. Une seule exception, pour Patagonia, serait l'expression “*without added pfas*”. En dehors de ce cas, les trigrams n'apportent aucune valeur ajoutée.

3.3 Clustering

Afin d'explorer la structure sous-jacente de notre corpus sans recourir à des étiquettes prédéfinies, nous avons opté pour une approche d'apprentissage non supervisé. L'objectif est de segmenter les produits en groupes homogènes maximisant la similarité intra-cluster ainsi que la distinction inter-clusters.

Pour ce faire, nous avons utilisé l'algorithme *k-Means*, implémenté via la bibliothèque scikit-learn. Cet algorithme partitionne les données en minimisant l'inertie intra-classe (la somme des carrés des distances entre chaque point et le centroïde de son cluster).

3.4 Détermination du nombre de clusters

L'une des contraintes majeures du *k-Means* est la nécessité de définir a priori le nombre de groupes k. Pour pallier cela de manière empirique, nous avons implémenté la méthode du coude (Elbow method) (voir annexe 5).

Cette méthode consiste à exécuter l'algorithme pour une plage de valeurs de k (*par exemple de 1 à 15*) et à tracer l'évolution de l'inertie globale. Sur le graphique généré par *Matplotlib*, nous recherchons le point d'infexion (*le "coude"*) où l'ajout d'un nouveau cluster n'entraîne plus de baisse significative de l'inertie, indiquant ainsi le nombre optimal de groupes pour modéliser nos données.

Dans le cas de Patagonia, nous avons fixé notre nombre de clusters à 7, sur base de l'analyse conjointe de la méthode du coude ainsi que du score de silhouette. L'examen de la courbe d'inertie met en évidence un point d'infexion autour de cette valeur et par ailleurs, le score de silhouette augmente de manière significative jusqu'à k = 7 puis n'évolue que faiblement (voir annexe 6).

3.5 Analyse des résultats et biais inter-sites

Lors de l'application de cette méthode sur l'ensemble du corpus fusionné, nous avons observé un phénomène particulier lié à la nature hétérogène de nos sources. La méthode du coude tendait à suggérer un partitionnement en 3 clusters principaux. L'analyse des mots-clés de ces clusters a révélé que ce regrou-

nement ne correspondait pas à des catégories de produits (*ex : pantalons vs t-shirts*), mais plutôt à la provenance des données (*Patagonia, Ecoalf ou Armedangels*). (Voir annexe 5 et 7 “tous les produits”).

Cette observation met en évidence que la variance stylistique et lexicale entre les différents sites web est supérieure à la variance sémantique entre les types de produits. Les descriptions sont trop "marquées" par le vocabulaire spécifique de chaque marque. Par conséquent, nous avons conclu que le clustering est plus pertinent lorsqu'il est utilisé pour des analyses intra-sites, ou doit être interprété en gardant à l'esprit que les regroupements globaux reflètent d'abord une signature rédactionnelle avant de refléter une typologie de vêtements.

3.6 Visualisation par réduction de dimension

Les vecteurs TF-IDF évoluant dans un espace de très haute dimension (*autant de dimensions que de mots dans le vocabulaire*), ils ne sont pas directement visualisables. Pour interpréter graphiquement les résultats de notre clustering, nous avons appliqué une Analyse en Composantes Principales (*PCA*). (voir annexe 7)

Cette technique nous a permis de projeter les données sur un plan 2D en conservant la variance maximale. Le graphique de dispersion (*scatter plot*) résultant, généré via Matplotlib, permet ainsi d'observer la séparation spatiale des groupes et d'identifier visuellement les éventuels chevauchements ou les produits atypiques (*outliers*).

3.7 Mesure de la distance sémantique : la « Similarité Cosinus »

Pour quantifier la proximité sémantique entre produits à partir de leurs vecteurs TF-IDF, nous utilisons la similarité cosinus. Contrairement à la distance euclidienne, sensible à la longueur des documents, cette métrique mesure le cosinus de l'angle entre deux vecteurs dans l'espace multidimensionnel ce qui neutralise ainsi l'impact de l'hétérogénéité des longueurs textuelles.

Ce choix se justifie doublement par sa robustesse face à l'hétérogénéité des données et son efficacité computationnelle. D'une part, la similarité cosinus permet d'identifier la proximité sémantique entre un texte court et un texte long décrivant un même type de produit, évitant le biais volumétrique des descriptions plus élaborées. D'autre part, elle s'avère particulièrement performante pour traiter les espaces vectoriels de haute dimension générés par la matrice TF-IDF, gérant efficacement la nature éparse des données où la majorité des valeurs sont nulles. Nos statistiques sur cette distance se trouvent à l'annexe 8.

3.8 Analyse sémantique orientée par catégories lexicales

Afin de diversifier notre approche analytique et dépasser la simple exploration par clustering vue précédemment, nous avons mis en place une seconde méthode d'analyse textuelle dite orientée par dictionnaires. Contrairement au clustering qui laisse les thématiques émerger d'elles-mêmes, cette approche

vise à quantifier la présence de concepts pré-identifiés comme stratégiques dans le domaine de la mode durable.

3.9 Définition des axes lexicaux

Nous avons constitué trois lexiques distincts, voir annexe 9, correspondant aux trois dimensions fondamentales de la communication produit dans ce secteur :

- Axe Durabilité et ESG (*Environnement, Social et Gouvernance*) : Ce lexique regroupe les termes liés à l'éthique, l'écologie et la responsabilité sociale (*ex : sustainable, ethical, fair trade, carbon footprint, eco-friendly*).
- Axe Matériaux et Composition : Ce lexique se concentre sur les matières premières et les intrants (*ex : organic cotton, hemp, recycled polyester, tencel, linen*).
- Axe Technique et Visuel : Ce lexique capture les aspects descriptifs, fonctionnels et esthétiques du vêtement (*ex : fit, pockets, zipper, color, cut, size, durable*).

3.10 Métrique et comparaison inter-sites

Pour chaque description de produit, notre algorithme calcule la fréquence d'apparition des termes appartenant à chacune de ces trois catégories. Afin de neutraliser le biais lié à la longueur des textes (*certaines sites étant plus « verbeux » que d'autres*), nous normalisons ces fréquences pour obtenir un pourcentage de couverture lexicale.

Cette méthodologie nous permet de réaliser une analyse comparative inter-sites robuste.

En agrégeant ces scores par site web, nous pouvons identifier la stratégie marketing dominante de chaque entreprise sur base des scores détaillés en annexe 10.

- Patagonia privilégie une communication axée sur la transparence des matériaux, mettant en avant la composition brute.
- Ecoalf insiste sur la dimension ESG, cherchant à vendre une éthique ou une valeur morale plutôt qu'un simple produit.
- Enfin, Armedangels reste ancré sur une description fonctionnelle, utilisant la durabilité comme un argument secondaire.

3.11 Vers une classification par dominance

Au-delà de l'analyse descriptive, cette quantification ouvre la voie à une méthode de classification automatique supervisée. En définissant des seuils d'appartenance, il est possible de catégoriser chaque produit selon sa "dominante sémantique". Par exemple, un produit dont plus de X pourcentages du vocabulaire descriptif relève de la catégorie ESG pourrait être classé automatiquement comme "Produit à forte valeur éthique", indépendamment de sa catégorie vestimentaire réelle (*pantalon, pull*). Cette classification permettrait de structurer l'offre non plus par type de vêtement, mais par type d'argumentaire

de vente.

3.12 Calcul de similarité entre rapports ESG et descriptions produits

Dans cette dernière section de notre partie Text Mining, nous avons comparé, pour chaque marque, la distribution lexicale de son rapport au vocabulaire utilisé sur ses pages produits. Concrètement, nous avons d'abord calculé, séparément pour les descriptions produits et pour le rapport ESG, la fréquence de chaque token en unigram. Puis nous avons normalisé ces fréquences par le nombre total de mots de chaque document de façon à neutraliser le biais lié à la longueur des textes. Les deux tableaux de fréquences normalisées ont ensuite été fusionnés sur la colonne « token » afin d'obtenir, pour chaque mot, sa fréquence relative dans les descriptions produits et dans le rapport. Pour chaque token commun ou spécifique à l'un des deux documents, nous avons calculé une « part commune » définie comme le minimum des deux fréquences normalisées. La somme de ces parts communes fournit un score global de similarité compris entre 0 et 1, interprétable comme la proportion de vocabulaire effectivement partagé entre le discours institutionnel et le discours produit. En complément, le tri décroissant de cette « part commune » permet d'identifier les mots qui contribuent le plus à cette proximité lexicale, mettant en évidence les notions de durabilité réellement cohérentes entre les engagements affichés dans le rapport et la manière dont les produits sont décrits sur le site. Les résultats obtenus sont disponibles dans l'annexe 11. L'application de cette mesure de similarité met en avant Ecoalf qui présente le score de similarité le plus élevé (20,23%), traduisant une forte continuité entre le vocabulaire institutionnel de ses rapports ESG et celui mobilisé dans ses descriptions produits, notamment autour des notions de recyclage, économie circulaire et de durabilité. Patagonia affiche un niveau de cohérence également élevé (18,94%), reposant principalement sur des concepts ESG transversaux liés aux matériaux et aux pratiques environnementales. A l'inverse, Armedangels présente un score de similarité plus bas, révélant un discours produit davantage orienté vers une approche descriptive et technique de la fabrication. Ces résultats suggèrent que la durabilité occupe un rôle plus ou moins central dans l'argumentaire commercial des marques, selon leur stratégie de communication.

4 Link Analysis

Dans le cadre de l'exploration des relations sémantiques intra-site, nous avons adopté une approche de Link Analysis (*analyse de liens*) basée sur la théorie des graphes. Cette méthode vise à modéliser et visualiser la structure des similarités entre les produits afin d'en extraire des connaissances exploitables pour la navigation ou la recommandation. Notre méthodologie repose sur la construction d'un graphe : $G = (V, E)$ où V représente les produits et E les liens pondérés par la similarité cosinus.

Ces arêtes sont pondérées par le score de similarité cosinus (*Cosine Similarity*), calculé sur la base des descriptions textuelles vectorisées (TF-IDF). Pour garantir la robustesse de nos résultats, nous avons déployé une stratégie d'analyse en deux temps, combinant visualisation et calcul algorithmique.

4.1 Méthodologie

Approche 1 : Dans un premier temps, notre pipeline de Text Mining exporte la matrice de similarité sous forme de fichiers structurés (`« nodes.csv » et « edges.csv »`). Cette étape matérialise la proximité sémantique sous forme de liens topologiques, permettant l'importation des données dans le logiciel Gephi. Cet outil nous permet de visualiser la macrostructure du réseau et d'appliquer des algorithmes de spatialisation vectorielle pour identifier des clusters visuels. Pour cela, nous utilisons un algorithme « Force Atlas » (*Fruchterman-Reingold*) qui simule un système physique où les nœuds se repoussent comme des particules chargées, tandis que les arêtes agissent comme des ressorts attirant les nœuds connectés. Plus deux produits sont proches visuellement, plus leur similarité textuelle est forte. Nous utilisons l'attribut « cluster » créé dans la fonction « `export_to_gephi` » dans le fichier « `text_mining_main` » afin de visualiser les différents clusters dans Gephi. En suivant ces étapes, nous obtenons un graphe qui reflète relativement bien les différentes catégories du site internet.

Approche 2 : En complément, nous avons développé un code python dédié « `link_analysis_main.py` ». Cette approche, plus flexible que l'interface rigide de Gephi, nous permet d'automatiser le calcul des métriques et de vérifier manuellement les chemins sémantiques. Afin de pouvoir utiliser les codes réalisés au cours, nous transformons la matrice de similarité en matrice d'adjacence en utilisant un seuil similaire à celle utilisée dans l'exportation vers Gephi.

4.2 Choix des métriques et justification

Conformément aux concepts du cours, nous avons retenu trois mesures de centralité pertinentes :

- Degré de centralité : Elle mesure le nombre de connexions directes d'un produit. Dans notre contexte, un produit avec un fort degré est un produit générique, partageant un vocabulaire commun avec une grande partie du catalogue.
- Plus court chemin : Cette mesure calcule la distance minimale entre deux produits. Elle est cruciale pour comprendre la « distance sémantique » : combien d'étapes conceptuelles sont nécessaires pour passer d'un vêtement de sport à un accessoire d'hiver ?
- Centralité d'intermédiarité : Elle identifie les nœuds qui agissent comme des « ponts » entre des groupes de produits disparates.
- PageRank : conçu pour les graphes orientés avec notion d'autorité (*pages web, citations*), l'algorithme n'est pas applicable à notre graphe de similarité textuelle symétrique.

4.3 Analyse des résultats obtenus avec Patagonia

L'analyse a été menée sur le catalogue complet extrait du site Patagonia. Voici les interprétations des trois métriques clés.

4.3.1 Analyse de la centralité de degré (Degree Centrality)

Nous observons une moyenne de connexions de 35,002 par produit. Cela indique un maillage sémantique relativement dense : chaque produit est textuellement similaire à environ 35 autres produits du catalogue. (Voir annexe 12).

Interprétation : Les produits les plus centraux sont majoritairement des équipements techniques de pluie (« *Men's Granite Crest Rain Jacket* », « *Men's Endless Run Shorts* »). Cela s'explique par l'ADN de la marque Patagonia : ces produits utilisent un vocabulaire technique très standardisé, des termes omniprésents dans l'ensemble du corpus. Ce sont les « piliers sémantiques » du site.

4.3.2 Analyse des plus courts chemins (Shortest Path)

Pour tester la pertinence de la navigation sémantique, nous avons calculé le plus court chemin entre deux produits a priori éloignés : de « *Women's Long-Sleeved Rugby Top* » à « *PowSlayer Beanie* ». (Voir annexe 13).

Interprétation : Le résultat valide notre modèle. Le chemin ne saute pas aléatoirement. Il opère une transition logique : il part de *Women's Long-Sleeved Rugby Top* -> *Cotton Down Jacket* -> *Natural Blend Retro Cardigan* -> *Brodeo Beanie* -> *PowSlayer Beanie*. Le Shortest Path est donc bien de 4.

4.3.3 Analyse de la centralité d'intermédiarité (Betweenness Centrality)

Le calcul de cette métrique s'est révélé nettement plus coûteux en temps de calcul que les précédents. Les résultats de cette métrique ne sont pas similaires à ceux sur Gephi car Gephi utilise les poids de la matrice de similarité pour quantifier les relations entre les produits or nous utilisons, dans notre approche sur python, une matrice d'adjacence. (Voir annexe 14).

Interprétation : Contrairement à la centralité de degré dominée par l'équipement de pluie technique, les produits à forte intermédiarité sont des vêtements « hybrides ». Ils font le pont entre plusieurs catégories qui n'ont pas forcément de ressemblance. « *Fieldsmith Hip Pack 5L* », qui a la betweenness la plus élevée, fait le lien entre les sacs à dos et des pantalons/short alors que les deux catégories n'ont rien à voir.

5 Conclusion

Pour terminer ce rapport, voici une synthèse des réponses à nos 3 questions de recherche développées dans l'introduction.

Notre analyse a permis d'identifier trois stratégies sémantiques radicalement divergentes dans la valorisation des produits durables. Chez Patagonia, citons les mots : recycled polyester/nylon, fair trade, DWR et l'absence de PFAS. L'analyse par dictionnaires révèle que cette marque surinvestit l'axe « Technical » (15,23%), complété par des termes ESG (12,01%), ce qui intègre la durabilité comme propriété

intrinsèque de la performance du vêtement. Ecoalf adopte une logique inverse, privilégiant un lexique ESG dense (27,07% des occurrences) où recycled, supply chain, certification et traceability forment un récit explicite d'économie circulaire et de responsabilité. Armedangels, enfin, privilégie une description fonctionnelle (13,24% « Technical » avec leur mot phare production par exemple contre seulement 7,43% « ESG ») où la durabilité est surtout portée par des certifications normatives (GOTS, GRS, PETA) et par un langage de slow fashion du quotidien. Les mots animal (right) et vegan sont également très présents. Ces trois approches révèlent que le discours durable n'est pas homogène : il peut être technique, narratif ou normatif selon la stratégie de marque.

La cohérence entre discours institutionnel et descriptions produits varie fortement selon les marques. Ecoalf présente le score de similarité lexicale le plus élevé (20,23%), avec recycled, water, environmental et sustainability qui circulent aussi bien dans le rapport ESG que dans les fiches produits. Patagonia affiche une cohérence intermédiaire (18,94%), fondée sur organic, recycled et water, traduisant une certaine intégration des engagements vers le niveau produit. Armedangels, en revanche, présente un score significativement plus bas (10,68%), où les termes institutionnels restent cloisonnés dans les rapports sans vraiment se retrouver dans les descriptions individuelles. Cette fragmentation lexicale suggère une division fonctionnelle des discours : les engagements normalisés dans un registre corporate, les fiches produits ancrées dans une description plus utilitaire.

Le graphe de similarité construit à partir des vecteurs TF-IDF (similarité cosinus moyenne : 0,3834) démontre effectivement qu'il est possible de reconstituer une structure sémantique porteuse à partir du vocabulaire seul. Le clustering $k=7$ identifie des ensembles cohérents : cluster 0 groupe les équipements techniques (membrane, shell, waterproof), cluster 5 concentre les pièces fleece/insulation avec DWR/P-FAS. La centralité de degré révèle des produits « piliers » massivement connectés incarnant le cœur standardisé de la durabilité, tandis que la betweenness centrality met en avant des produits hybrides reliant des univers distincts. Ces analyses confirment que le vocabulaire de la durabilité structure implicitement l'architecture du catalogue et ouvre des perspectives opérationnelles réelles : les produits à forte intermédialité constituent des leviers naturels pour la recommandation sémantique, capable de faire circuler les promesses de durabilité à travers des segments variés.

Notre rapport s'inscrit dans une approche purement scolaire et statistique, analysant exclusivement la dimension lexicale des communications durables sans prétendre aux ressemblances réelles entre les marques. Les trois piliers méthodologiques du cours se sont révélés complémentaires pour déconstruire le discours marketing à travers un prisme sémantique mais leurs conclusions demeurent limitées à cette dimension unique. Une validation robuste nécessiterait de dépasser le Text Mining pour valider les informations que nous mettons en avant ici, en étudiant par exemple la perception des consommateurs pour chacune de ces marques.

Annexes

Annexe 1 – Structure des fichiers Excel

name	category	description	url	Source_file
73 Savine Upstate Hoodie	FeatureFeatured > Shop by CategoryShop by Category > Sweatshirts & Hoodies	The 73 Savine is a plus-sized pullover hoodie with a brim made with Bureau's fully traceable Netfleec® 100% recycled fishing nets. This full-panel, organza-like hoodie is perfect for layering over your favorite t-shirt or tank top.	https://eu.patagonia.com/gb/en/product/73-savine-upstate-hoodie/396781.html	patagonia_gb_men
Airflow™ Upstate Hoodie	FeatureFeatured > Shop by CategoryShop by Category > Hats & Accessories	Featuring a brim built with Bureau's full traceable Netfleec® 100% recycled fishing nets, this full-panel, organza-like hoodie is perfect for layering over your favorite t-shirt or tank top.	https://eu.patagonia.com/gb/en/product/airflow-upstate-hoodie/379365.html	patagonia_gb_men
Atom Sling Bag 8L	PacksPacks > Packs & GearBags & Luggage > BackpacksBackpacks	On a bike, in boats or catching the bus, this take-it-everywhere 8-liter sling pack has enough storage to keep all your gear organized. It's made with Bureau's fully traceable Netfleec® 100% recycled fishing nets.	https://eu.patagonia.com/gb/en/product/atom-sling-bag-8-liters/48262.html	patagonia_gb_men
Atom Tote Pack 20L	PacksPacks > Packs & GearBags & Luggage > BackpacksBackpacks	Atom Tote The Atom Tote Pack 20L is for those who want a smaller-fitting pack with removable tech storage that's capable of holding a laptop. It's made with Bureau's fully traceable Netfleec® 100% recycled fishing nets.	https://eu.patagonia.com/gb/en/product/atom-tote-backpack-20-liters/48125.html	patagonia_gb_men
Balacava	FeatureFeatured > Shop by CategoryShop by Category > Hats & Accessories	Simplicity gets a technical twist with the Balacava, built from warm, quick-drying Capilene® Thermal Weight fabric.	https://eu.patagonia.com/gb/en/product/balacava/2501.html	patagonia_gb_men
Balancer	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/balancer/2502.html	patagonia_gb_men
Better Sweater™ Fleece Gloves	FeatureFeatured > Shop by CategoryShop by Category > Hats & Accessories	These warm, cozy mittens convert to fingerless gloves and are designed for multipurpose sport or casual activities.	https://eu.patagonia.com/gb/en/product/better-sweater-fleece-convertible-gloves-mittens/32.html	patagonia_gb_men
Black Hole® Pack 25L	PacksPacks > Packs & GearBags & Luggage > BackpacksBackpacks	Black Hole® This burly 25-liter daypack has just the right amount of space to haul your daily essentials. It delivers a sleek look and a sturdy feel.	https://eu.patagonia.com/gb/en/product/black-hole-pack-25-liters/49298.html	patagonia_gb_men
Black Hole® Pack 35L	PacksPacks > Packs & GearBags & Luggage > BackpacksBackpacks	Black Hole® Our midsize 35-liter workhorse pack in the Black Hole® collection is perfect for organizing your gear and getting you where you need to go.	https://eu.patagonia.com/gb/en/product/black-hole-pack-35-liters/49300.html	patagonia_gb_men
Black Hole® Work Pack 30L	PacksPacks > Packs & GearBags & Luggage > BackpacksBackpacks	Black Hole® Work Our midsize 30-liter workhorse pack in the Black Hole® collection is perfect for organizing your gear and getting you where you need to go.	https://eu.patagonia.com/gb/en/product/black-hole-work-pack-30-liters/49302.html	patagonia_gb_men
Boardshort Label Funfair Cap	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/boardshort-label-funfair-cap/38378.html	patagonia_gb_men
Broadcaster Hat	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/broadcaster-high-crown-trucker-hat/35395.html	patagonia_gb_men
Brooks® Cap	WomenWomen's > Shop by CategoryShop by Category > Hats & Accessories	Go off the beaten path with a smooth, soft cap for easy wearing, and hollow-core seams and a diamond-grid base.	https://eu.patagonia.com/gb/en/product/brooks-cap/34549.html	patagonia_gb_men
Captain's Midweight Liner Gloves	WomenWomen's > Shop by CategoryShop by Category > Hats & Accessories	WomenWomen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/captains-midweight-liner-gloves/34549.html	patagonia_gb_men
Chouinard® Crest Upstate Hoodie	MenMen's > Shop by CategoryShop by Category > TopsTops > Sweatsuits & Trousers	The Chouinard® Crest Upstate Hoodie is made with 100% recycled knit fleece that has a brushed interior and a ribbed exterior.	https://eu.patagonia.com/gb/en/product/chouinard-crest-upstate-fleece-hoodie/39707.html	patagonia_gb_men
Corduroy Cap	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/corduroy-cap/33835.html	patagonia_gb_men
Cottonwood™ Jersey	FeatureFeatured > Shop by CategoryShop by Category > Shirts	MenMen's > Shop by CategoryShop by Category > Shirts	https://eu.patagonia.com/gb/en/product/cottonwood-jersey/22765.html	patagonia_gb_men
Daily Crewneck Sweatshirt	FeatureFeatured > Shop by CategoryShop by Category > Sweatshirts & Hoodies	Embrace the basics in our cozy, long-lasting Daily Crewneck Sweatshirt. Whether you're relaxing at home or getting outside.	https://eu.patagonia.com/gb/en/product/daily-crew-neck-sweatshirt/22765.html	patagonia_gb_men
Daily Hoody Sweatshirt	FeatureFeatured > Shop by CategoryShop by Category > Sweatshirts & Hoodies	Embrace the basics in our cozy, long-lasting Daily Hoody Sweatshirt. Whether you're relaxing at home or getting outside.	https://eu.patagonia.com/gb/en/product/daily-hoody-sweatshirt/22770.html	patagonia_gb_men
Dash™ Tee	MenMen's > Shop by CategoryShop by Category > TopsTops > Short-Sleeved	A lightweight tee made of 100% Cotton in a Convention fit, which supports natural working toward better health.	https://eu.patagonia.com/gb/en/product/dash-tee/42185.html	patagonia_gb_men
DASH Light Shorts	SportSports > SnowSnow MenMen's Snow & Backcountry Touring	MenMen's Snow & Backcountry Touring	https://eu.patagonia.com/gb/en/product/dash-light-insulated-shorts/85355.html	patagonia_gb_men
Descentional Snow Pant 3L	PacksPacks > Packs & GearBags & Luggage > Technical PackTechnical Packs	Designed for a quick and easy dose of warmth on cold days, our ultralight DASH Light Shorts have dual side zips.	https://eu.patagonia.com/gb/en/product/dash-light-insulated-shorts/85378.html	patagonia_gb_men
Dispenser Roll-Top Pack 40L	PacksPacks > Packs & GearBags & Luggage > Technical PackTechnical Packs	For long days or multi-day treks in the backcountry, the Dispenser Roll-Top Snow Pant is a no-nonsense pack with a bonded interior and a bonded exterior.	https://eu.patagonia.com/gb/en/product/dispenser-roll-top-snow-pant/40-liter/482.html	patagonia_gb_men
Duckbill Cap	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/duckbill-cap/28818.html	patagonia_gb_men
Duckbill Running Trucker Hat	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/duckbill-running-trucker-hat/28738.html	patagonia_gb_men
Dusk™ Tee	MenMen's > Shop by CategoryShop by Category > TopsTops > Short-Sleeved	MenMen's > Shop by CategoryShop by Category > TopsTops > Short-Sleeved	https://eu.patagonia.com/gb/en/product/dusk-tee/42186.html	patagonia_gb_men
Feldsmith Hip Pack 3L	PacksPacks > Packs & GearBags & Luggage > FieldSmith Hip Pack	The jack-of-all-trades Fieldsmith Hip Pack keeps essentials at the ready for wherever the day takes you. Built in.	https://eu.patagonia.com/gb/en/product/feldsmith-hip-pack-5-liters/48480.html	patagonia_gb_men
Feldsmith Lid Pack 3L	PacksPacks > Packs & GearBags & Luggage > FieldSmith Lid Pack	The Fieldsmith Lid Pack has a water-resistant eXtreME.	https://eu.patagonia.com/gb/en/product/feldsmith-lid-pack-28-liters/48484.html	patagonia_gb_men
Feldsmith Linked Backpack 3L	PacksPacks > Packs & GearBags & Luggage > FieldSmith Linked Backpack	Embrace the basics with a Fieldsmith Linked Backpack. Made with Bureau's fully traceable Netfleec® 100% recycled fishing nets.	https://eu.patagonia.com/gb/en/product/feldsmith-linked-backpack-3-liters/48485.html	patagonia_gb_men
Fisherman's Rolled Beanie	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/fishermans-rolled-beanie/29105.html	patagonia_gb_men
Fitz Roy Icon Uprisal Crewneck Sweater	FeatureFeatured > Shop by CategoryShop by Category > Sweatshirts & Hoodies	MenMen's > Shop by CategoryShop by Category > Sweatshirts & Hoodies	https://eu.patagonia.com/gb/en/product/fitz-royc-icon-uprisal-fleece-crewneck-sweatshirt/396.html	patagonia_gb_men
Fitz Roy Icon Uprisal Crewneck Sweater	FeatureFeatured > Shop by CategoryShop by Category > Sweatshirts & Hoodies	MenMen's > Shop by CategoryShop by Category > Sweatshirts & Hoodies	https://eu.patagonia.com/gb/en/product/fitz-royc-icon-uprisal-fleece-crewneck-sweatshirt/396.html	patagonia_gb_men
Fitz Roy Icon Uprisal Crewneck Sweater	FeatureFeatured > Shop by CategoryShop by Category > Sweatshirts & Hoodies	MenMen's > Shop by CategoryShop by Category > Sweatshirts & Hoodies	https://eu.patagonia.com/gb/en/product/fitz-royc-icon-uprisal-fleece-crewneck-sweatshirt/396.html	patagonia_gb_men
Fitz Roy Trouser Trouser Hat	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/fitz-royc-trouser-trouser-hat/38286.html	patagonia_gb_men
Fly Catcher Hat	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	MenMen's > Shop by CategoryShop by Category > Hats & Accessories	https://eu.patagonia.com/gb/en/product/flycatcher-hat/33475.html	patagonia_gb_men
Foot Tractor Aluminum Bar ReplacementSports > Fly FishingFly Fishing Poles & Gear	MenMen's > Shop by CategoryShop by Category > Fly FishingFly Fishing Poles & Gear	If you need this kit, it means you're fishing a lot. It contains everything you need to repair the bars and rods.	https://eu.patagonia.com/gb/en/product/foot-tractor-aluminum-bar-replacement/39220.html	patagonia_gb_men
Foot Tractor Wading Boots - Felt Sole	MenMen's > Shop by CategoryShop by Category > Waders & Boots	WadersMen's Waders & Boots	https://eu.patagonia.com/gb/en/product/foot-tractor-wading-boots-felt/79345.html	patagonia_gb_men
Foot Tractor Wading Boots - Sticky	MenMen's > Shop by CategoryShop by Category > Waders & Boots	WadersMen's Waders & Boots	https://eu.patagonia.com/gb/en/product/foot-tractor-wading-boots-sticky-rubber/79170.html	patagonia_gb_men
Foot Wading Boots - Grip Studs™	SportsSports > Fly FishingFly Fishing Poles & Gear	Specifically designed for Foot Wading Boots built by Freewell, the Grip Studs™ Traction Kit bolsters grip in slick	https://eu.patagonia.com/gb/en/product/foot-wading-boots-grip-studs-traction-kit/81765.html	patagonia_gb_men

FIGURE A1 – Structure des fichiers Excel utilisés "patagonia_all_products.xlsx"

Annexe 2 – Répartition hommes/femmes

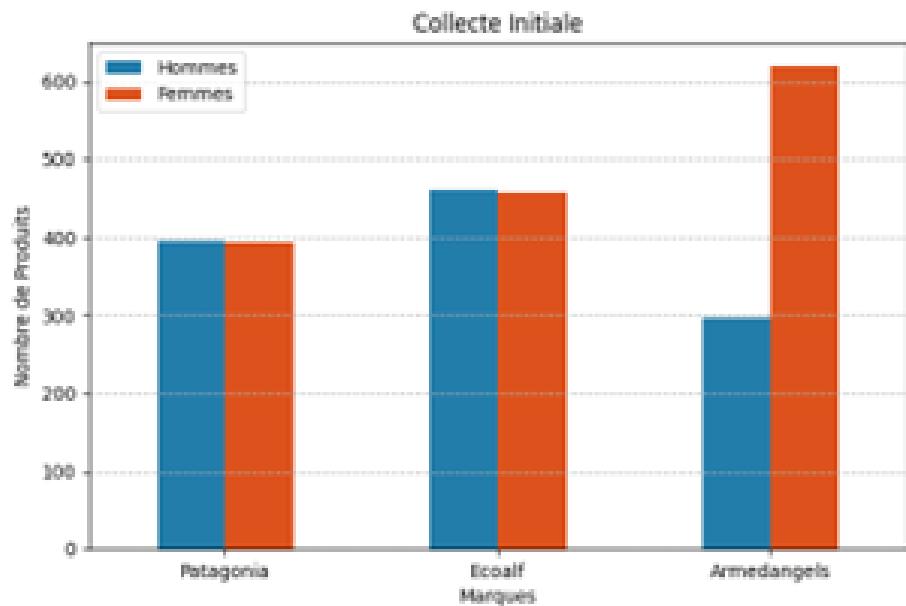


FIGURE A2 – Répartition H/F - Graphique 1 - avant nettoyage des doublons

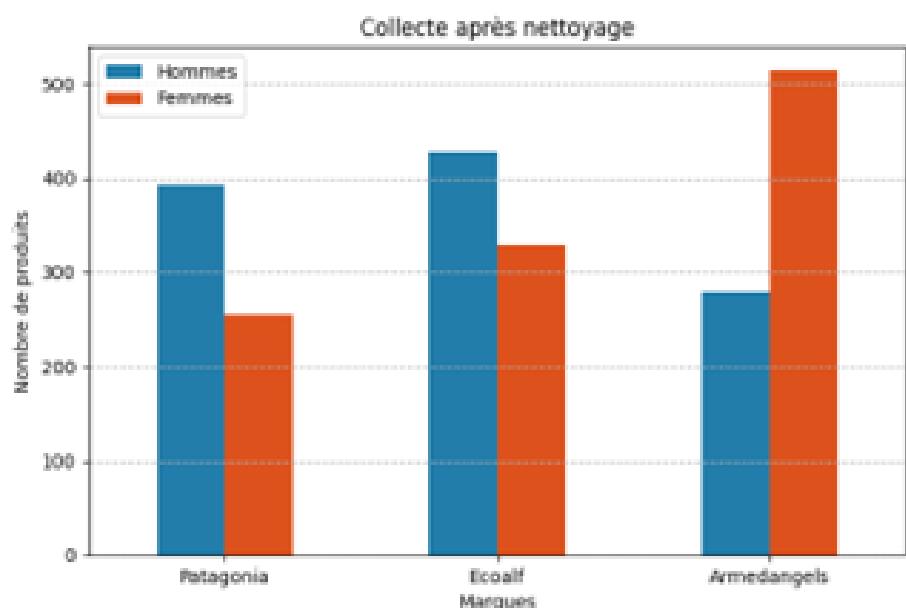


FIGURE A3 – Répartition H/F - Graphique 2 - après nettoyage des doublons

Annexe 3 – Matrices TF-IDF

```

> TF (Relative Frequency):
access added adjustable airmesh allows another anything approved
0.000 0.000 0.25 0.000 0.000 0.000 0.00 0.000
0.000 0.000 0.75 0.000 0.000 0.000 0.25 0.000
0.143 0.286 0.00 0.143 0.000 0.143 0.00 0.143
0.200 0.400 0.00 0.200 0.000 0.000 0.00 0.200
0.000 0.000 0.00 0.000 0.667 0.000 0.00 0.333

> IDF (Logarithm):
access added adjustable airmesh allows another anything approved
1.984 0.559 1.111 3.907 1.668 5.374 2.861 0.559

> TF-IDF Final:
access added adjustable airmesh allows another anything approved
0.000 0.000 0.043 0.000 0.000 0.000 0.000 0.000
0.000 0.000 0.109 0.000 0.000 0.000 0.094 0.000
0.052 0.029 0.000 0.103 0.000 0.141 0.000 0.015
0.053 0.030 0.000 0.104 0.000 0.000 0.000 0.015
0.000 0.000 0.000 0.000 0.134 0.000 0.000 0.022

```

FIGURE A4 – Matrice TF-IDF - Patagonia unigram

```

> TF (Relative Frequency):
access breathable added pfas adjustable drawstring ... allows full anything fair approved fair
0 0.00 0.0 0.333 ... 0.0 0.00 0.00 0.00
1 0.00 0.0 0.000 ... 0.0 0.25 0.00 0.00
2 0.00 0.5 0.000 ... 0.0 0.00 0.00 0.25
3 0.25 0.5 0.000 ... 0.0 0.00 0.00 0.25
4 0.00 0.0 0.000 ... 0.5 0.00 0.00 0.50

[5 rows x 8 columns]

> IDF (Logarithm):
access breathable added pfas adjustable drawstring ... allows full anything fair approved fair
0 5.374 0.729 4.526 ... 4.17 5.374 0.603

[1 rows x 8 columns]

> TF-IDF Final:
access breathable added pfas adjustable drawstring ... allows full anything fair approved fair
0 0.00 0.000 0.115 ... 0.000 0.000 0.000
1 0.00 0.000 0.000 ... 0.000 0.141 0.000
2 0.00 0.045 0.000 ... 0.000 0.000 0.019
3 0.11 0.030 0.000 ... 0.000 0.000 0.012
4 0.00 0.000 0.000 ... 0.147 0.000 0.021

[5 rows x 8 columns]

```

FIGURE A5 – Matrice TF-IDF - Patagonia bigram

Annexe 4 – Nuages de mots

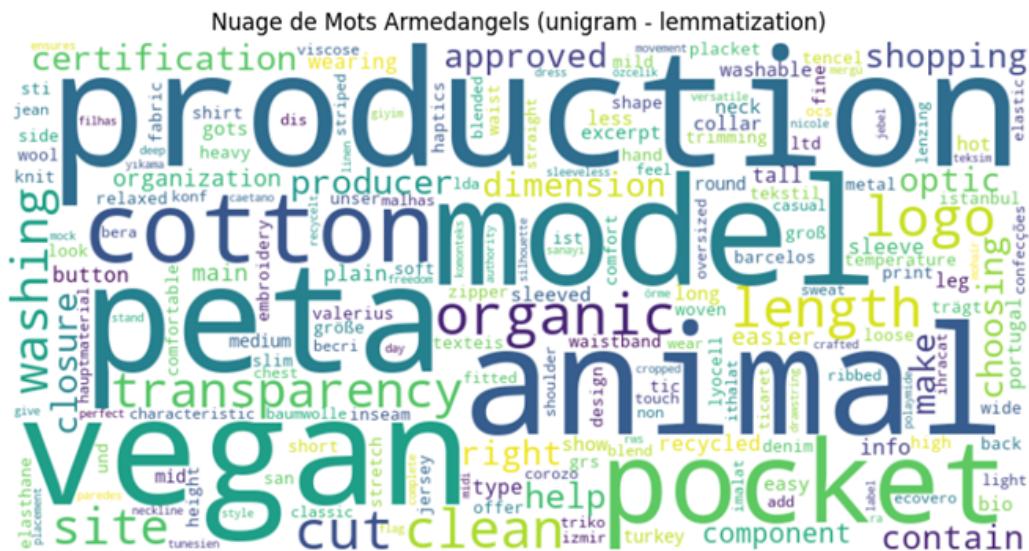


FIGURE A6 – Nuage de mots 1 (Armedangels - Unigram - Lemmantisation)

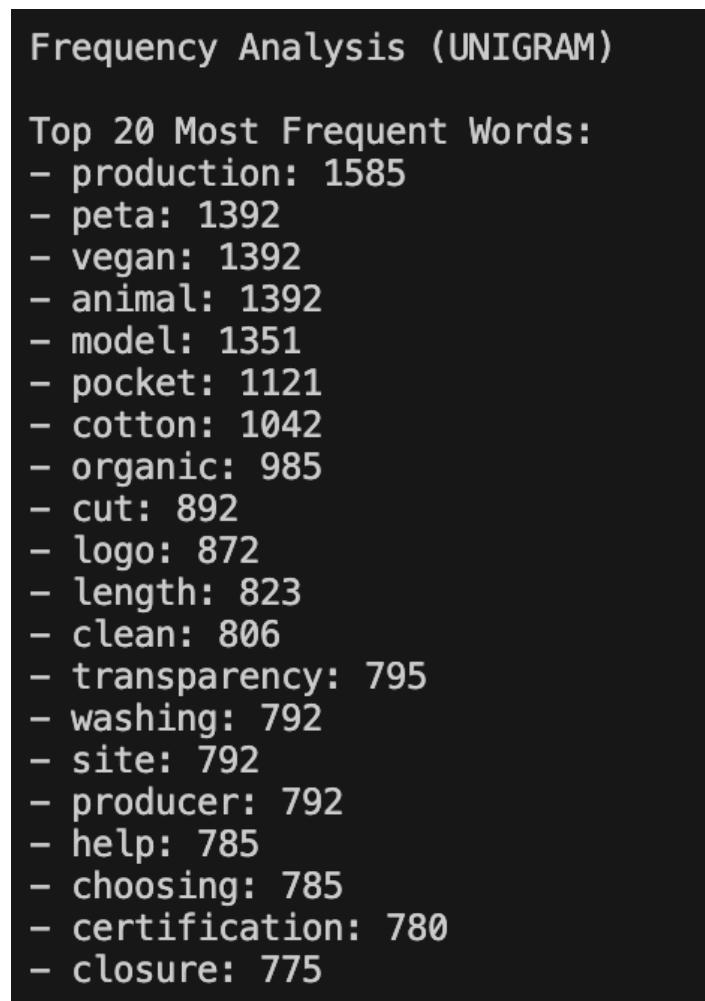


FIGURE A7 – Top 20 des mots les plus fréquents (*Armedangels - Unigram - Lemmantisation*)

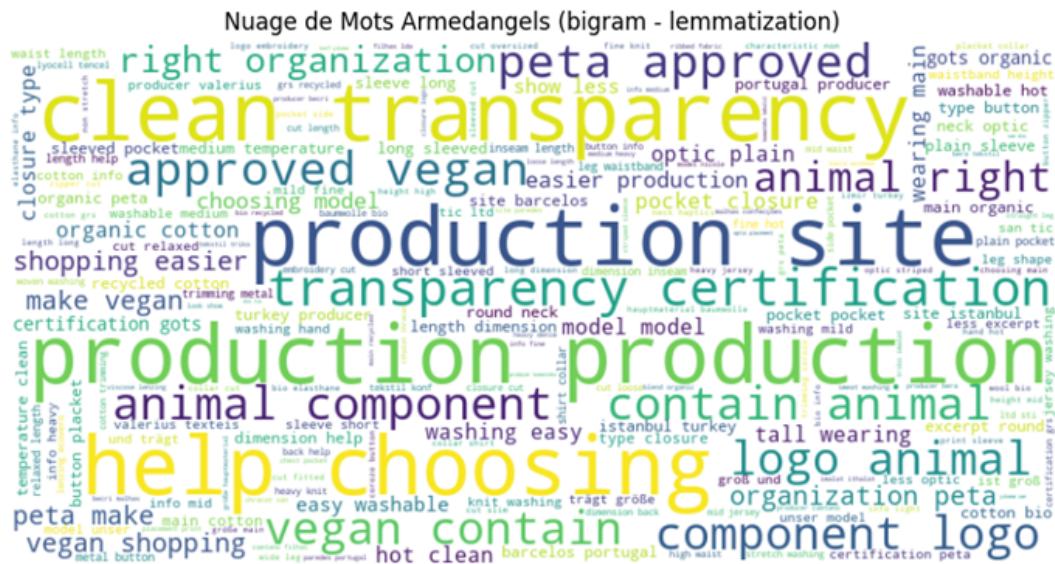


FIGURE A8 – Nuage de mots 2 (Armedangels - Bigram - Lemmantisation)

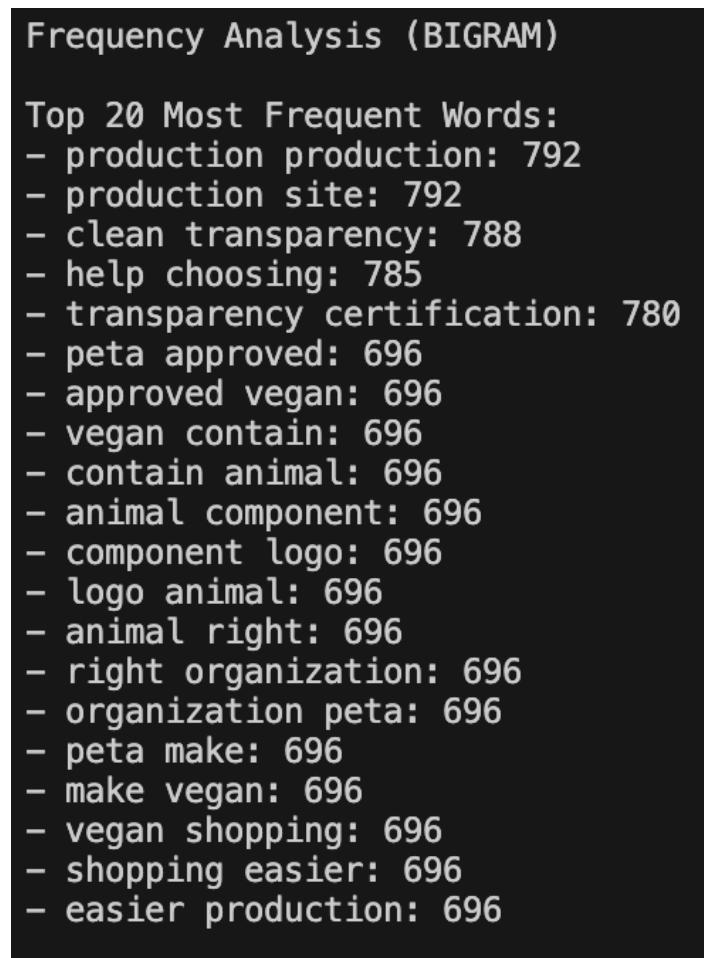


FIGURE A9 – Top 20 des mots les plus fréquents (Armedangels - Bigram - Lemmantisation)

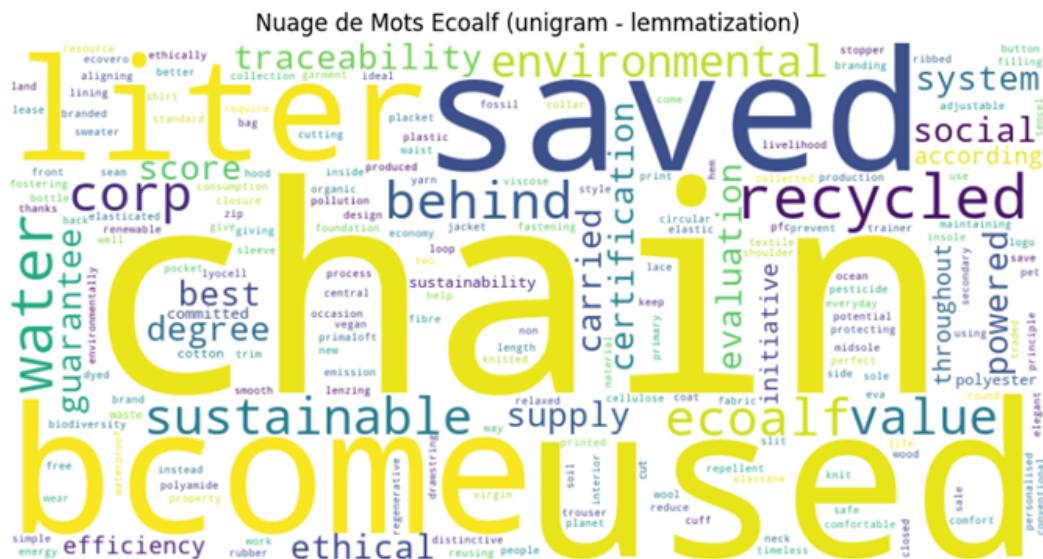


FIGURE A10 – Nuage de mots 3 (Ecoalf - Unigram - Lemmantisation)

Top 20 Most Frequent Words:

- chain: 3085
 - used: 3031
 - bcome: 3028
 - saved: 2978
 - liter: 2976
 - recycled: 2496
 - water: 2360
 - corp: 1726
 - ecoalf: 1724
 - behind: 1708
 - sustainable: 1621
 - value: 1569
 - environmental: 1566
 - traceability: 1523
 - social: 1523
 - carried: 1523
 - best: 1520
 - supply: 1516
 - system: 1516
 - ethical: 1516

FIGURE A11 – Top 20 des mots les plus fréquents (Ecoagf - Unigram - Lemmantisation)

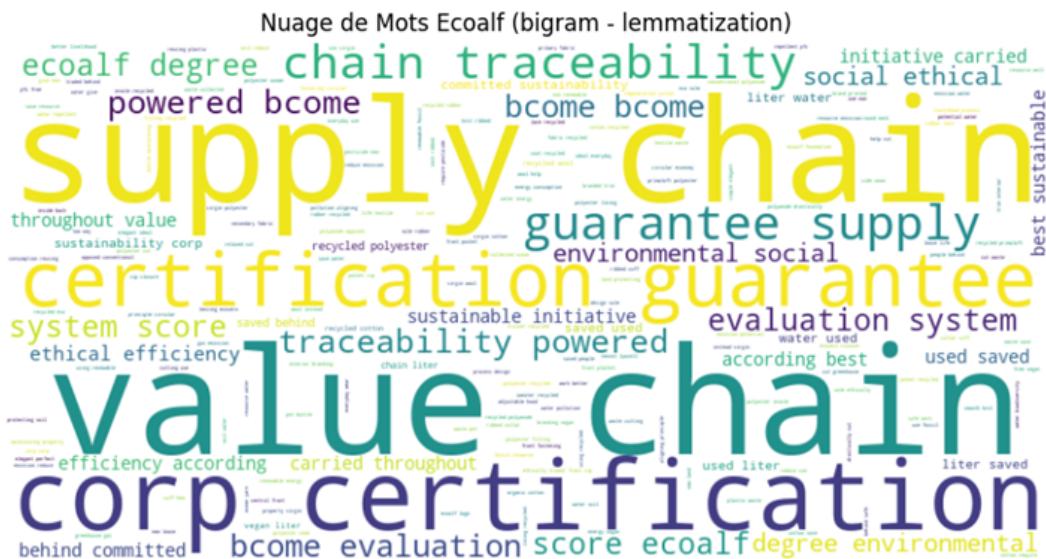


FIGURE A12 – Nuage de mots 4 (Ecoalf - Bigram - Lemmantisation)

Top 20 Most Frequent Words:

- value chain: 1569
- supply chain: 1516
- corp certification: 1514
- certification guarantee: 1514
- guarantee supply: 1514
- chain traceability: 1514
- traceability powered: 1514
- powered bcome: 1514
- bcome bcome: 1514
- bcome evaluation: 1514
- evaluation system: 1514
- system score: 1514
- score ecoalf: 1514
- ecoalf degree: 1514
- degree environmental: 1514
- environmental social: 1514
- social ethical: 1514
- ethical efficiency: 1514
- efficiency according: 1514
- according best: 1514

FIGURE A13 – Top 20 des mots les plus fréquents (Ecoalf - Bigram - Lemmantisation)

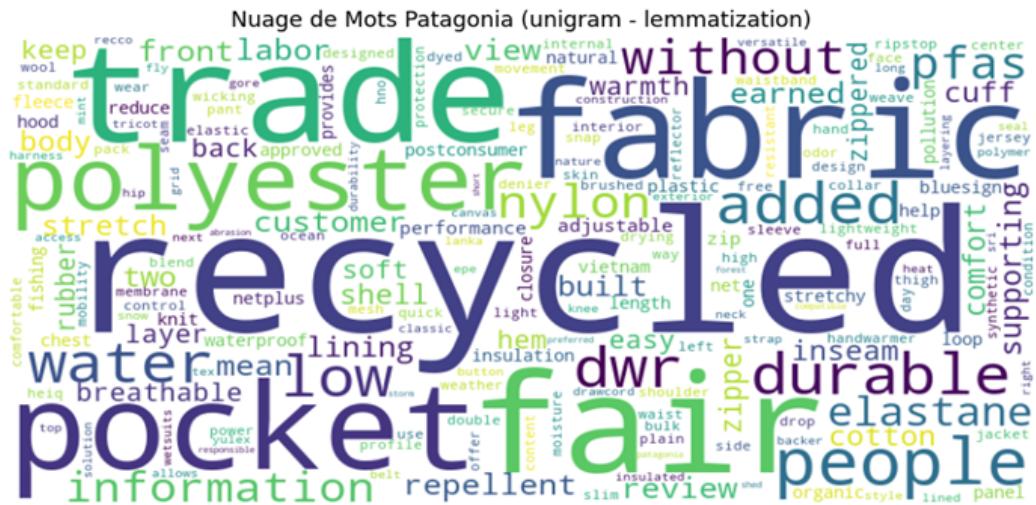


FIGURE A14 – Nuage de mots 5 (Patagonia - Unigram - Lemmantisation)

Top 20 Most Frequent Words:	
- recycled:	3067
- trade:	1843
- fair:	1842
- fabric:	1819
- pocket:	1714
- polyester:	1465
- people:	1199
- durable:	932
- water:	921
- added:	903
- low:	902
- dwr:	880
- without:	847
- nylon:	825
- pfas:	760
- information:	676
- elastane:	633
- repellent:	632
- view:	630
- customer:	626

FIGURE A15 – Top 20 des mots les plus fréquents (Patagonia - Unigram - Lemmantisation)

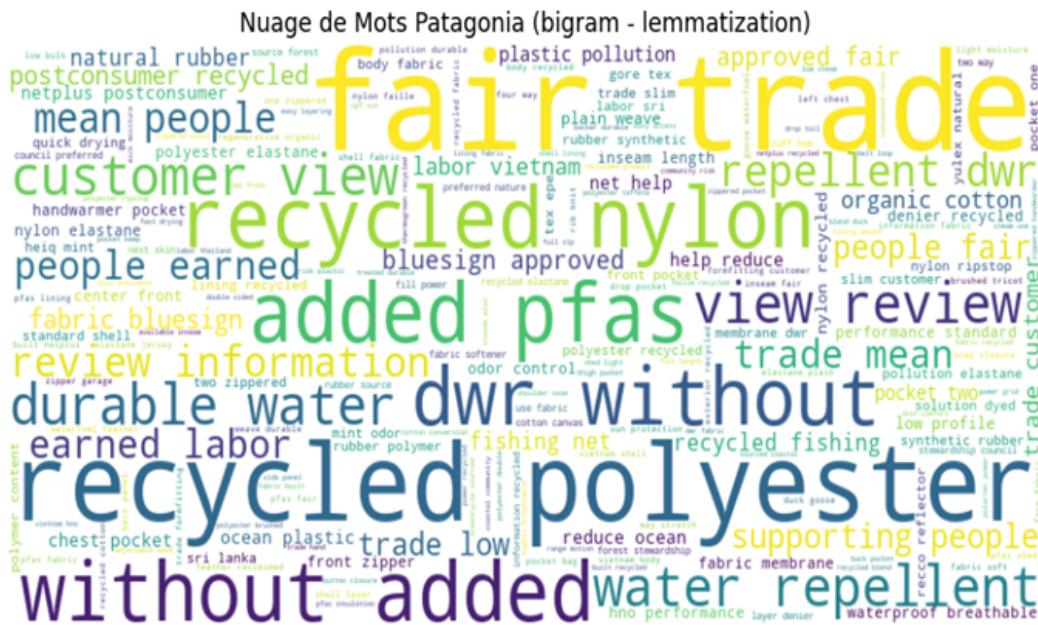


FIGURE A16 – Nuage de mots 6 (Patagonia - Bigram - Lemmantisation)

Top 20 Most Frequent Words:

- fair trade: 1842
 - recycled polyester: 1365
 - recycled nylon: 778
 - added pfas: 752
 - without added: 749
 - dwr without: 696
 - water repellent: 632
 - durable water: 630
 - customer view: 626
 - view review: 626
 - repellent dwr: 625
 - review information: 622
 - mean people: 597
 - people earned: 597
 - earned labor: 597
 - trade mean: 596
 - supporting people: 592
 - people fair: 532
 - trade low: 500
 - postconsumer recycled: 3

FIGURE A17 – Top 20 des mots les plus fréquents (Patagonia - Bigram - Lemmantisation)

Peta : People for the Ethical Treatment of Animal, association internationale à but non lucratif, qui oeuvre à protéger le droit et la dignité des animaux. La certification PETA, approuve que les produits certifiés ne contiennent aucune matière animale.

BCOME : plateforme technologique permettant de vérifier l'impact du vêtement sur la planète et les personnes impliquées dans sa production. Le système d'évaluation de BCOME note le degré d'efficacité environnementale, sociale et éthique d'Ecoalf en fonction de ses meilleures initiatives durables tout au long de sa chaîne de valeur, sur la base du fait que le dépassement de 0 % est déjà un impact positif.

DWR = Durable Water Repellent (Déperlant Durable) : c'est un traitement chimique appliqué sur les tissus extérieurs de leurs vêtements techniques pour que l'eau perle et glisse, les empêchant d'absorber l'humidité

PFAS « ajouté » : substance qui est volontairement incorporée dans un produit pour lui conférer une propriété spécifique (résistance à l'eau, antiadhésif, etc.)

Annexe 5 – Méthode du coude

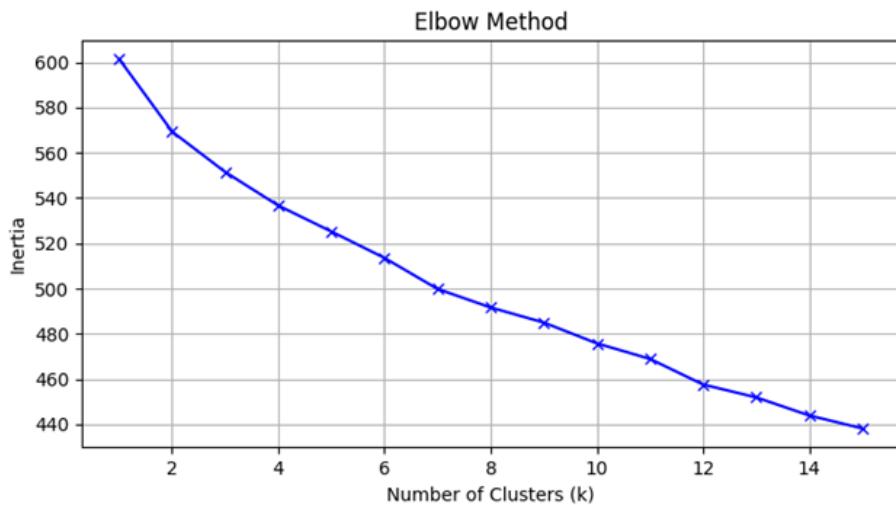


FIGURE A18 – Méthode du coude - Patagonia - Unigram

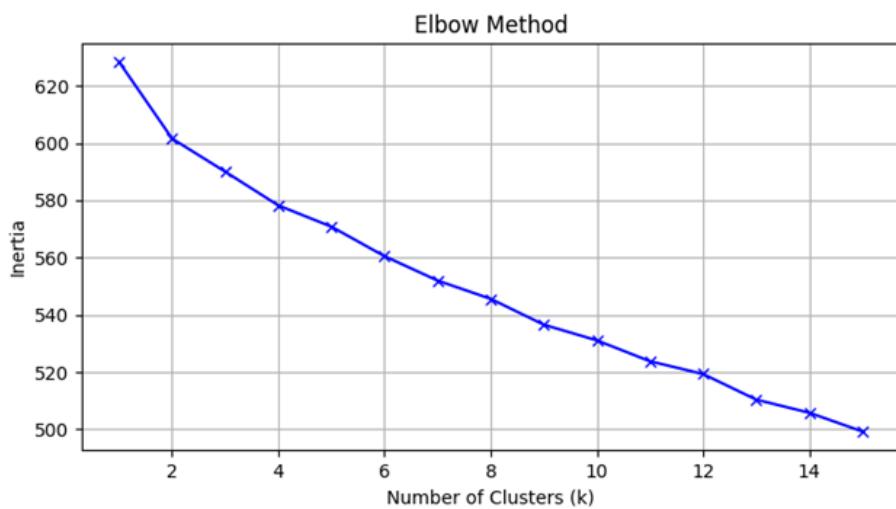


FIGURE A19 – Méthode du coude - Patagonia - Bigram

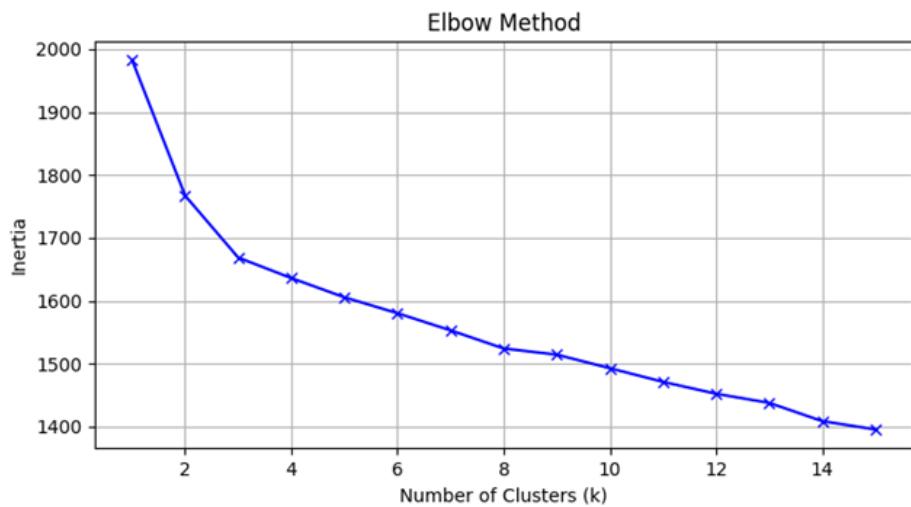


FIGURE A20 – Méthode du coude - Tous les produits - Unigram

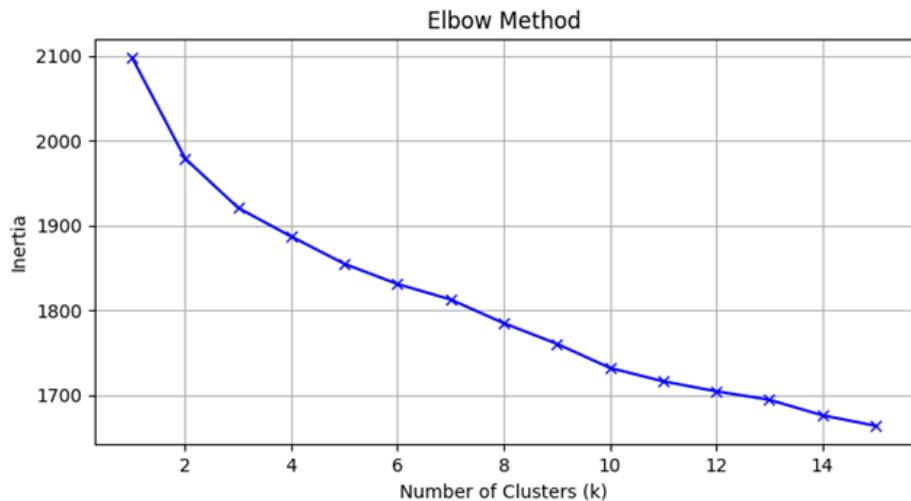


FIGURE A21 – Méthode du coude - Tous les produits - Bigram

Annexe 6 – Clustering

--- ANALYSE : LEMMATIZATION + UNIGRAM		
K	Silhouette Score	Inertia
2	0.0609	569.34
3	0.0630	551.49
4	0.0735	536.68
5	0.0761	525.20
6	0.0858	513.54
7	0.0958	499.84
8	0.1017	491.66
9	0.1055	485.02
10	0.1114	475.81
11	0.1135	468.96
12	0.1260	457.65
13	0.1234	451.97
14	0.1290	443.98

FIGURE A22 – Résultat du Clustering 1 : Patagonia - Unigram

```
Clustering (K=7)
> Silhouette Score: 0.0958

CLUSTER 0 :
Keywords : waterproof, membrane, reflector, hno, layer, tex, gore, epe, shell, performance
Size : 78 products

CLUSTER 1 :
Keywords : organic, cotton, canvas, regenerative, crown, button, brim, conversion, hemp, inseam
Size : 93 products

CLUSTER 2 :
Keywords : odor, heiq, control, grid, mint, polartec, wicking, power, chafing, pure
Size : 63 products

CLUSTER 3 :
Keywords : rubber, natural, yulex, polymer, content, wetsuits, synthetic, council, stewardship, preferred
Size : 42 products

CLUSTER 4 :
Keywords : shirt, cotton, crewneck, tee, mexico, postconsumer, reliance, virgin, raw, utilizing
Size : 36 products

CLUSTER 5 :
Keywords : elastane, fleece, insulation, panel, stretch, stretchy, strap, dwr, pocket, pfas
Size : 226 products

CLUSTER 6 :
Keywords : wool, netplus, ocean, plastic, fishing, net, pollution, postconsumer, reduce, help
Size : 109 products
```

FIGURE A23 – Résultat du Clustering 2 : Patagonia - Unigram

--- ANALYSE : LEMMATIZATION + BIGRAM ---		
K	Silhouette Score	Inertia
2	0.0402	601.59
3	0.0475	590.03
4	0.0541	578.25
5	0.0594	570.81
6	0.0670	560.51
7	0.0722	551.98
8	0.0772	545.59
9	0.0873	536.57
10	0.0841	531.08
11	0.0939	523.79
12	0.0949	519.33
13	0.1040	510.39
14	0.1081	505.80

FIGURE A24 – Résultat du Clustering 3 : Patagonia - Bigram

```
5. Clustering (K=7)
> Silhouette Score: 0.0722

CLUSTER 0 :
Keywords : fabric membrane, tex epe, gore tex, hno performance, performance standard, recco reflector, membrane dwr
, waterproof breathable, standard shell, without added
Size : 89 products

CLUSTER 1 :
Keywords : natural rubber, polymer content, synthetic rubber, rubber polymer, rubber synthetic, yulex natural, pref
ered nature, stewardship council, council preferred, forest stewardship
Size : 42 products

CLUSTER 2 :
Keywords : help reduce, netplus postconsumer, reduce ocean, ocean plastic, net help, recycled fishing, fishing net,
plastic pollution, postconsumer recycled, nylon faille
Size : 101 products

CLUSTER 3 :
Keywords : recycled cotton, cotton postconsumer, recycled fabric, labor mexico, postconsumer recycled, ribbing neck
, reliance virgin, covered almost, almost anything, virgin raw
Size : 33 products

CLUSTER 4 :
Keywords : organic cotton, regenerative organic, cotton canvas, cotton recycled, cotton conversion, netplus recycle
d, button closure, rib knit, industrial hemp, brim netplus
Size : 99 products

CLUSTER 5 :
Keywords : polyester elastane, elastane jersey, nylon elastane, power grid, polartec power, recycled polyester, way
stretch, four way, dwr without, recycled nylon
Size : 230 products

CLUSTER 6 :
Keywords : rws wool, control union, nylon ripstop, responsible standard, wool standard, responsible wool, land mana
gement, insulation eco, animal welfare, shell lining
Size : 53 products
```

FIGURE A25 – Résultat du Clustering 4 : Patagonia - Bigram

Annexe 7 – PCA

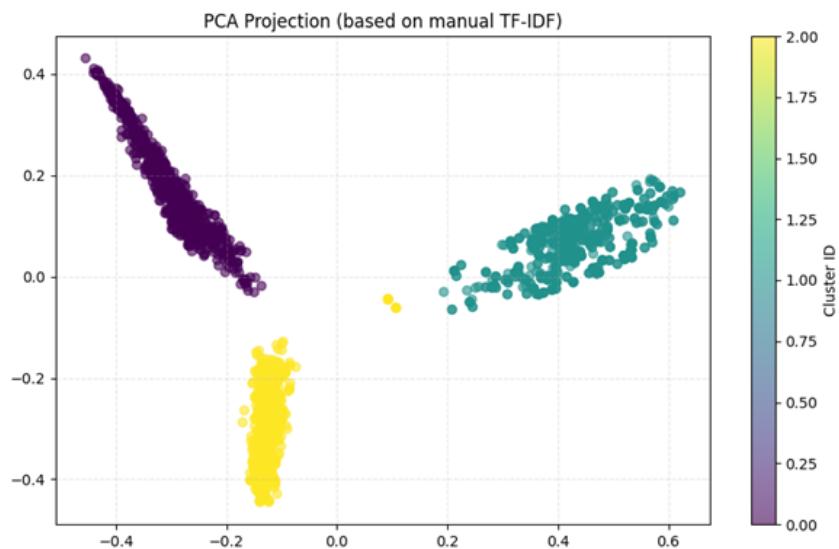


FIGURE A26 – Analyse en Composantes Principales 1 : Tous les produits - Unigram

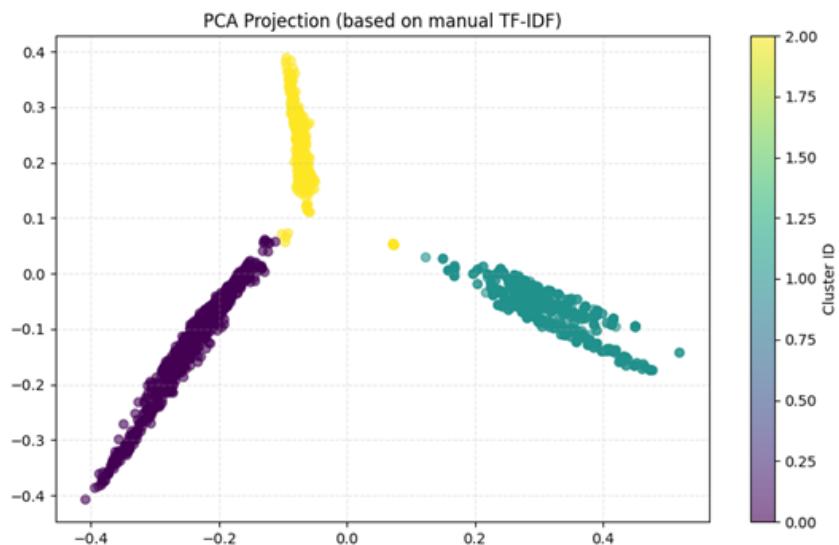


FIGURE A27 – Analyse en Composantes Principales 2 : Tous les produits - Bigram

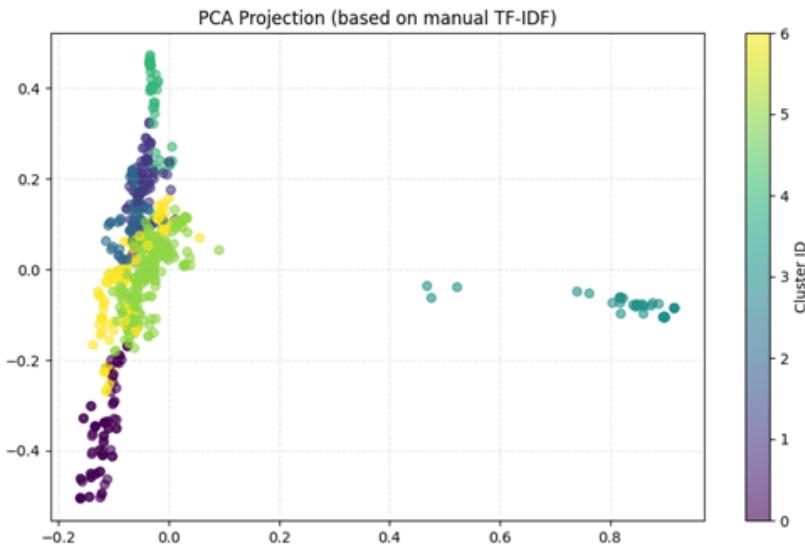


FIGURE A28 – Analyse en Composantes Principales 3 : Patagonia - Unigram

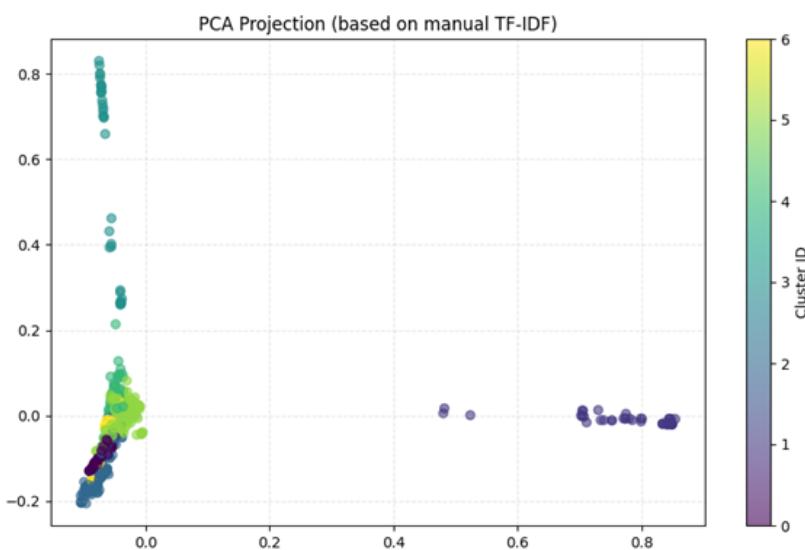


FIGURE A29 – Analyse en Composantes Principales 4 : Patagonia - Bigram

Annexe 8 – Statistiques de similarité

```
Similarity Statistics:  
– Min similarity: 0.2000  
– Max similarity: 1.0000  
– Mean similarity: 0.3947  
– Median similarity: 0.3006
```

FIGURE A30 – Statistiques de similarité 1 : Patagonia - Unigram

```
Similarity Statistics:  
– Min similarity: 0.1500  
– Max similarity: 1.0000  
– Mean similarity: 0.3834  
– Median similarity: 0.2817
```

FIGURE A31 – Statistiques de similarité 2 : Patagonia - Bigram

```
'ESG_DURABILITE': {
    'activism', 'bcome', 'biosoft', 'blended', 'bluesign', 'carbon',
    'certified', 'chain', 'circular', 'climate', 'corp', 'cottonrecycled',
    'downcycled', 'eco', 'economy', 'ecosystem', 'efficiency',
    'emissions', 'environment', 'environmental', 'ethical', 'fair',
    'fair trade', 'fishery', 'footprint', 'fossil', 'fsc', 'gots',
    'greenhouse', 'grs', 'hemprecycled', 'incinerators', 'initiatives',
    'labor', 'landfills', 'liters', 'net', 'netplus', 'netting', 'ocean',
    'oceancycle', 'ocs', 'organic', 'people', 'peta', 'pfas',
    'pfcsfpfas', 'planet', 'plastic', 'pollution', 'postconsumer', 'rcs',
    'recyc', 'recycled', 'regenerative', 'renewable', 'resources',
    'responsib', 'responsibilitee', 'responsible', 'rubber',
    'rwscertified', 'saved', 'sewn', 'social', 'solar', 'standard',
    'stewardship', 'supply', 'sustainab', 'traceab', 'traceable',
    'trade', 'used', 'vegan', 'waste', 'wastewater', 'worker'
},
```

FIGURE A32 – Extrait dictionnaire lexical 1 : Durabilité

Annexe 9 – Dictionnaires lexicaux

```
'TECHNIQUE_PHYSIQUE': {
    'abrasionresistant', 'adjustablesnap', 'ankle', 'articulated',
    'backzip', 'baffle', 'baffled', 'bifit', 'binding', 'branding',
    'breathable', 'button', 'buttonclosure', 'buttonned', 'buttonfront',
    'buttonsnap', 'buttonthrough', 'chin', 'closerfitting', 'closure',
    'collar', 'collared', 'compressible', 'control', 'cropped', 'cuff',
    'cufftohigh', 'dobby', 'doublecuff', 'doublesnap', 'drawcord',
    'drawcordadjustable', 'drawcords', 'drawstring', 'dropin',
    'dropped', 'droptail', 'durable', 'dwr', 'elastic', 'elasticated',
    'elasticized', 'embroidery', 'fastening', 'fill', 'filling',
    'fillpower', 'finish', 'fit', 'fitted', 'fivepocket', 'flap',
    'flatseam', 'foursnap', 'frontzip', 'fullzip', 'gasket', 'glove',
    'glovefriendly', 'guard', 'gusset', 'gusseted', 'h2no',
    'halfelastic', 'halfzip', 'heiq', 'hem', 'hemline', 'high', 'hood',
    'hooded', 'hoodie', 'hoodless', 'hoody', 'hookandloop', 'inseam',
    'insole', 'insulated', 'knee', 'laces', 'layer', 'layering', 'light',
    'lightweight', 'liner', 'linerfree', 'linerless', 'lining', 'loft',
    'longsleeved', 'loop', 'loose', 'metalbutton', 'mid', 'midi',
    'midlayer', 'midsole', 'mobility', 'mock', 'moisturewicking',
    'neck', 'odor', 'onseam', 'outseam', 'oversized', 'packable',
    'parka', 'placket', 'pocket', 'pocketrouted', 'print', 'pure',
    'quarterzip', 'quilt', 'quilting', 'raglan', 'regular',
    'regularfit', 'relaxed', 'relaxedfit', 'repellent', 'reversible',
    'ribbed', 'ribknit', 'rise', 'round', 'scuff', 'seam', 'seaming',
    'seamless', 'seamsealed', 'securezip', 'shankbutton', 'shell',
    'shirttail', 'shortsleeved', 'silhouette', 'singlebutton',
    'singleseam', 'singlesnap', 'sleeve', 'sleeveless', 'sleeves',
    'slim', 'slimfit', 'slimfitting', 'slimzip', 'snap',
    'snapadjustable', 'snapclosure', 'snapfront', 'snaponoff', 'snapt',
    'snatab', 'sole', 'standup', 'stoppers', 'storm', 'straight',
    'stretch', 'stretchy', 'taped', 'terryloop', 'threequartersleeved',
    'tophem', 'trims', 'twosnap', 'ultralight', 'unbuttoned',
    'verticalzip', 'verticalzippered', 'waist', 'waistband',
    'waistbandclosure', 'waistbelt', 'warmth', 'water-repellent',
    'waterproof', 'waterproofbreathable', 'weatherresistant', 'wick',
    'wicking', 'wide', 'windproof', 'yoke', 'zip', 'zipfly', 'zipneck',
    'zipout', 'zipped', 'zipper', 'zippered', 'zipperfly', 'zipsecured',
    'zipthrough'
```

FIGURE A33 – Extrait dictionnaire lexical 2 : Technique

```
'MATERIAUX_TEXTILES': {
    'acetate', 'airmesh', 'baggies', 'bio', 'bottles', 'brushedtricot',
    'canvas', 'capilene', 'cashmere', 'chiffon', 'coating', 'corduroy',
    'corozo', 'cotton', 'cottonelastane', 'cottonrecycled', 'crepe',
    'denim', 'denimstyle', 'doublefabric', 'down', 'downdrift',
    'downinsulated', 'downlike', 'duck', 'ecovero', 'elastane',
    'elasthane', 'elasticized', 'elastomultiester', 'eucotton', 'eva',
    'fabric', 'fabricstrap', 'feather', 'fiber', 'fibre', 'flannel',
    'fleece', 'fleecelike', 'fleecelined', 'gore-tex', 'gridfleece',
    'heather', 'hemp', 'herringbone', 'insulation', 'jacquard',
    'jersey', 'jerseyknit', 'knit', 'knitfleece', 'knitted', 'lace',
    'laminate', 'laminated', 'leather', 'lenzing', 'linen', 'lyocell',
    'material', 'materials', 'membrane', 'merino', 'mesh', 'meshback',
    'meshlined', 'microfiber', 'microfleece', 'microfleecelined',
    'microgridfleece', 'modal', 'mohair', 'nylon', 'nylonbound',
    'nyloncoated', 'nylonelastane', 'pertex', 'pet', 'pile', 'plaid',
    'plumafill', 'polartec', 'polyester', 'polyesterelastane',
    'polyestermesh', 'polyurethane', 'powermesh', 'primaloft', 'puff',
    'pulp', 'rib', 'ribbing', 'ripstop', 'selffabric', 'shearling',
    'sorona', 'spandex', 'stretchmesh', 'suede', 'sweat', 'synchilla',
    'synthetic', 'taffeta', 'tencel', 'textile', 'textile-to-textile',
    'thermogreen', 'tpu', 'tpufilm', 'tricot', 'tricotlined', 'twill',
    'twilllined', 'velour', 'velvet', 'viscose', 'wool', 'woolblend',
    'woven', 'yarn', 'yulex'
}
```

FIGURE A34 – Extrait dictionnaire lexical 3 : Matériaux

Annexe 10 – Distribution lexicale

Brand	ESG (%)	Technical (%)	Material (%)	Other (%)
Armedangels	7,43	13,24	8,62	70,72
Ecoalf	27,07	6,05	5,36	61,52
Patagonia	12,01	15,23	7,84	64,91

FIGURE A35 – Distribution lexicale

Annexe 11 – Similarité rapports ESG

Score de similarité global : 10.68%				
Top 10 des mots qui rendent les documents similaires :				
	token	freq_norm_1	freq_norm_2	part_commune
2243	production	0.024976	0.006823	0.006823
2036	organic	0.015522	0.005117	0.005117
1750	make	0.011913	0.004342	0.004342
604	cotton	0.016420	0.004187	0.004187
1215	gots	0.006303	0.004187	0.004187
2346	recycled	0.008462	0.003877	0.003877
1243	grs	0.003530	0.003566	0.003530
2196	portugal	0.006256	0.002946	0.002946
2431	right	0.011062	0.002791	0.002791
3204	wool	0.003766	0.002326	0.002326

FIGURE A36 – Similarité ESG 1 : Armedangels

Score de similarité global : 20.23%				
Top 10 des mots qui rendent les documents similaires :				
	token	freq_norm_1	freq_norm_2	part_commune
924	ecoalf	0.017174	0.016474	0.016474
2466	recycled	0.024865	0.009957	0.009957
3296	water	0.023510	0.007784	0.007784
1017	environmental	0.015600	0.005431	0.005431
2962	sustainability	0.014943	0.004254	0.004254
479	chain	0.030733	0.004164	0.004164
984	emission	0.003556	0.004888	0.003556
715	cotton	0.008468	0.003440	0.003440
2052	ocean	0.003367	0.004707	0.003367
3221	value	0.015630	0.003349	0.003349

FIGURE A37 – Similarité ESG 2 : Ecoalf

Score de similarité global : 18.94%				
Top 10 des mots qui rendent les documents similaires :				
	token	freq_norm_1	freq_norm_2	part_commune
4718	organic	0.003530	0.003259	0.003259
4927	people	0.011693	0.002905	0.002905
3170	help	0.003335	0.002834	0.002834
5490	recycled	0.029910	0.002267	0.002267
6424	standard	0.002516	0.002197	0.002197
1516	cotton	0.004622	0.002126	0.002126
7420	water	0.008982	0.002090	0.002090
1020	center	0.002106	0.001949	0.001949
2500	fair	0.017963	0.001949	0.001949
4670	one	0.002789	0.001842	0.001842

FIGURE A38 – Similarité ESG 3 : Patagonia

GOTS : Certification (Global Organic Textile Standard)

GRS : Global Recycled Standard

Annexe 12 – Degree Centrality

Degree Centrality :	
Top 10 Degree Centrality :	
Men's Granite Crest Rain Jacket	109
Women's Granite Crest Rain Jacket	109
Men's Endless Run Shorts - 6"	105
Women's Outdoor Everyday Rain Jacket	102
Men's Swiftcurrent® Wading Jacket	102
Men's Endless Run Tights	101
Women's Granite Crest Rain Pants	100
Women's Swiftcurrent® Wading Jacket	100
Men's Granite Crest Rain Pants	100
Men's Baggies™ Lights - 6"	96
dtype: int64	
Moyenne :	35.0015455950541

FIGURE A39 – Top 10 produits Patagonia par degré de centralité

Annexe 13 – Shortest Path

```
Shortest Path :  
  
Chemin de : 'Women's Long-Sleeved Rugby Top'  
Vers       : 'PowSlayer Beanie'  
Distance   : 4.0  
Etapes     : Women's Long-Sleeved Rugby Top -> Cotton Down Jacket
```

FIGURE A40 – Résultat demo Shortest Path

Annexe 14 – Betweenness Centrality

Betweenness Centrality :	
Natural Blend Retro Cardigan	16304.591382
Fieldsmith Hip Pack 5L	14866.034063
Wavefarer™ Bucket Hat	11768.825265
Men's Nomader Joggers	11266.415277
Women's Nano-Air® Light Bottoms	10581.822815
Men's Reversible Down Better Sweater™	10259.099131
Women's All Seasons Bomber Hoody Work Jacket	9092.603790
Women's Cord Fjord Jacket	7427.171893
Women's Outdoor Everyday Rain Jacket	7062.095352
Women's Pack Out Hike Tights	6973.743304

FIGURE A41 – Top 10 produits Patagonia par centralité d'intermédiarité

Annexe 15 – Gephi



FIGURE A42 – Visualisation des clusters Patagonia

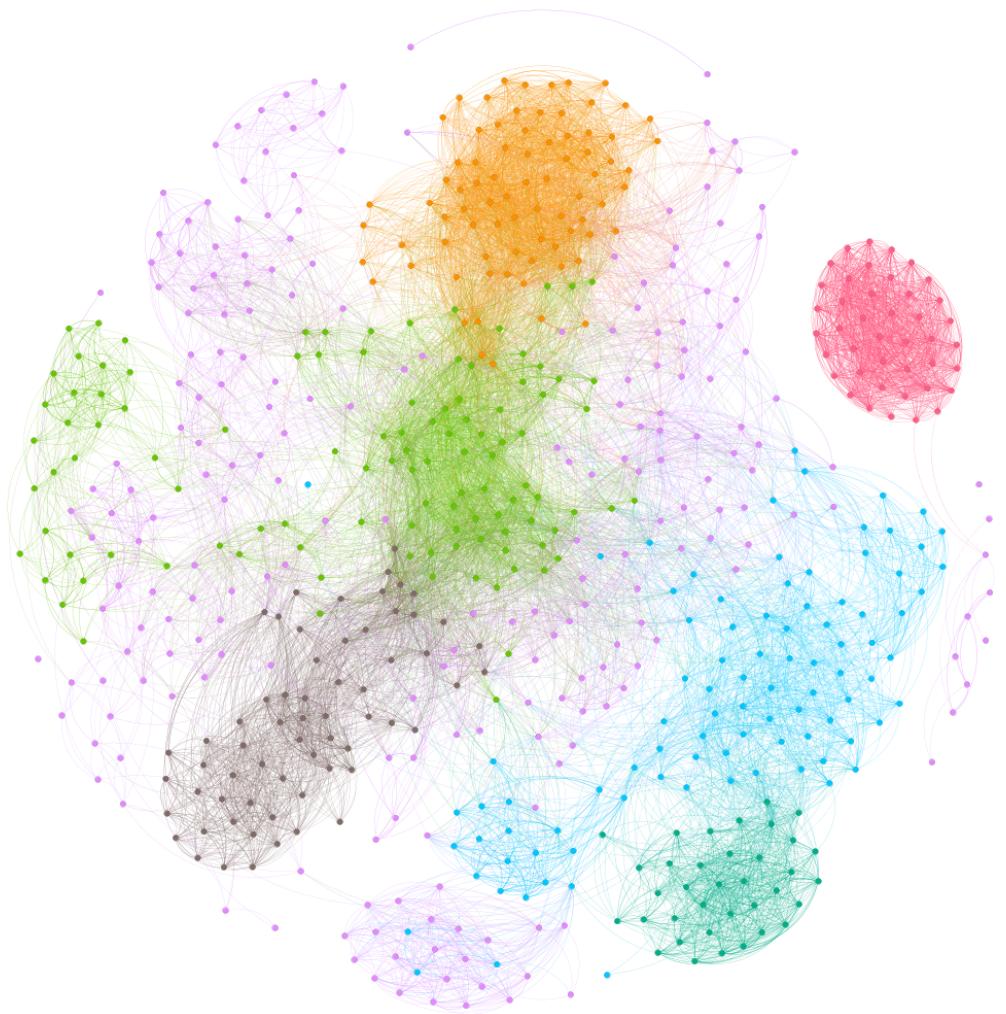


FIGURE A43 – Visualisation sans labels

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons Belgique | www.uclouvain.be/lsm