MLSMM2153 : Web Mining
WebLens: Navigating Modern Challenges

Professeurs : Corentin Vande Kerckhove & Sylvain Courtain

Année académique 2025-2026

1 Contexte et objectif du projet

Ce projet constitue une application concrète des concepts et techniques abordés dans le cadre du cours, à travers l'analyse de données collectées sur des sujets d'actualité. En tant que consultant pour une organisation, vous aurez pour mission de collecter et d'analyser des informations disponibles en ligne sur une problématique spécifique. L'objectif est de fournir des recommandations stratégiques basées sur ces analyses.

Dans le cadre de ce travail, vous utiliserez des méthodes de *Web Mining*, telles que le *Text Mining* et le *Link Analysis*, pour identifier des tendances, extraire des insights et analyser les relations entre les données. Ce projet vise à développer vos compétences dans l'application de ces techniques tout en répondant à des questions stratégiques qui seront définies par votre groupe et l'organisation.

2 Détails du projet

2.1 Constitution du groupe et choix du sujet

Chaque groupe doit être constitué de 3 étudiants et choisir un sujet parmi ceux proposés ci-dessous. Il est essentiel de faire votre sélection rapidement sur la plateforme Moodle, car chaque thématique peut être attribuée à un seul groupe. Le principe de sélection est **premier arrivé, premier servi** : une fois un sujet sélectionné, il devient indisponible pour les autres groupes.

Les thématiques suivantes sont proposées pour ce projet. Pour chaque sujet, une description détaillée des objectifs, des méthodes de collecte des données, des techniques d'analyse à appliquer et des possibilités de personnalisation sont fournies afin de guider votre analyse.

Sujet 1 : Benchmark RSE – Analyse des engagements sur les sites web d'entreprises

- **Objectif** : Comparer les discours RSE de grandes entreprises issues de secteurs variés.
- **Scraping** : Collecte des pages "Développement durable" ou "Responsabilité sociale" de sites corporate (ex : airfrance.com, totalenergies.com, etc.).
- **Text Mining** : Analyse des thèmes, des promesses et des mots-clés dominants dans ces discours.

- **Web Mining** : Création d'un graphe des concepts partagés ou d'un réseau des sous-thématiques abordées.

Sujet 2 : Analyse des blogs de niche dans un domaine d'expertise (mode, sport, tech, food...)

- **Objectif** : Explorer un micro-univers d'experts ou passionnés dans un secteur spécifique.
- **Scraping** : Collecte des blogs personnels (WordPress, Wix, etc.) ou de listes de blogs thématiques.
- **Text Mining** : Identifier les sujets dominants et analyse de l'évolution des thèmes traités.
- **Web Mining** : Création d'un graphe des liens entre blogs (blogrolls, articles référencés, etc.).

Sujet 3 : Discours public et transparence – Analyse des pages "À propos" d'organisations

- **Objectif** : Comparer la manière dont différentes organisations (ONG, entreprises, institutions) se présentent au public.
- **Scraping** : Collecte des pages "À propos" / "Qui sommes-nous" / "Notre mission" des sites web d'organisations.
- **Text Mining** : Analyse de la tonalité, du vocabulaire identitaire et de l'engagement affiché.
- **Web Mining** : Création d'un réseau des concepts partagés ou des mots en co-occurrence dans les discours.

Sujet 4 : Analyse des critiques culturelles sur des blogs (cinéma, théâtre, jeux vidéo)

- **Objectif** : Explorer la manière dont des œuvres culturelles sont critiquées ou valorisées.
- **Scraping** : Collecte de blogs critiques (WordPress, Overblog, etc.).
- **Text Mining** : Analyse du vocabulaire évaluatif, de la fréquence des notes, et de la polarité des critiques.
- **Web Mining** : Création d'un réseau des œuvres citées ensemble ou des thèmes associés.

Sujet 5 : Analyse de la présentation des villes sur les sites touristiques officiels

- **Objectif** : Comparer la manière dont différentes villes mettent en avant leur attractivité.
- **Scraping** : Collecte des pages d'accueil de sites de tourisme locaux (ex : visitbrussels.be, lyon-france.com).

- **Text Mining** : Identification des champs lexicaux dominants et des mots associés au bien-être, à la culture, et à l'économie.
- **Web Mining** : Création d'un graphe des catégories de contenu ou des co-thèmes abordés (gastronomie, événements, etc.).

Sujet 6 : Étude de la mise en avant des produits durables dans les e-boutiques engagées

- **Objectif** : Analyser la manière dont des marques éthiques valorisent leurs produits durables.
- **Scraping** : Collecte des pages produits de boutiques en ligne (HTML simple, pas besoin d'authentification).
- **Text Mining** : Analyse du vocabulaire utilisé pour mettre en avant les engagements écologiques et durables.
- **Web Mining** : Création d'un graphe des catégories ou des produits similaires pour étudier les relations.

Sujet 7 : Analyse comparative de pages FAQ sur des sites d'entreprises tech

- **Objectif** : Étudier les préoccupations des utilisateurs et la manière dont elles sont adressées par les entreprises tech.
- **Scraping** : Collecte des pages FAQ de sites comme Notion, Slack, Miro, Framasoft, etc.
- **Text Mining** : Analyse des thématiques abordées, des types de questions fréquemment posées.
- **Web Mining** : Création d'un réseau des concepts en analysant les co-occurrences dans les questions/réponses.

Sujet 8 : Comparaison de la communication institutionnelle des écoles / universités

- **Objectif** : Étudier la manière dont les établissements se présentent, leurs valeurs et leur positionnement.
- **Scraping** : Collecte des pages "À propos" ou "Pourquoi nous rejoindre" des sites universitaires.
- **Text Mining** : Analyse de la tonalité (excellence, inclusion, innovation...) et des fréquences des termes.
- **Web Mining** : Création d'un graphe des valeurs ou concepts partagés dans la communication des établissements.

Sujet 9 : Exploration du discours écologique dans les sites de festivals ou événements

- **Objectif** : Analyser comment les organisateurs de festivals ou événements intègrent la durabilité dans leur communication.

- **Scraping** : Collecte des pages “engagements”, “durabilité”, “éco-responsabilité” de sites de festivals.
- **Text Mining** : Analyse du vocabulaire, identification des actions concrètes vs promesses vagues.
- **Web Mining** : Création d’un graphe des thématiques abordées (mobilité, déchets, alimentation...).

3 Concepts minimum à explorer et conseils pratiques

Dans cette section sont présentés les concepts minimaux à explorer pour votre projet. Ces lignes directrices, applicables à toutes les thématiques proposées, vous guideront dans la collecte de données, l’analyse textuelle et l’analyse des liens. Vous devrez les adapter en fonction du sujet spécifique que vous aurez choisi.

3.1 Collecte de données (Scraping)

La collecte de données est une étape essentielle du projet. Vous devrez récupérer automatiquement des informations disponibles en ligne, relatives à votre sujet, en utilisant des techniques de *scraping*. Voici les lignes directrices pour cette partie :

- **Sélection des sources** : Choisissez des pages web fiables et pertinentes pour votre sujet. Vous pouvez récupérer des informations provenant de sites institutionnels, de blogs spécialisés, d’articles de presse, de forums ou de toute autre source en ligne liée à votre thématique. L’objectif est de naviguer entre les pages pour obtenir une couverture complète du sujet, en vous concentrant sur les informations les plus pertinentes.
- **Navigation et extraction** : Utilisez des outils de *scraping* pour automatiser la navigation entre les pages et extraire les données pertinentes. Cela peut inclure des textes, des titres, des liens, des dates ou d’autres métadonnées qui enrichiront votre analyse. L’important est de structurer les données récoltées sous un format utilisable pour l’analyse, tel que le format CSV ou JSON.
- **Stratégie et volume de collecte**

  Il est essentiel de définir une **stratégie de collecte** garantissant à la fois une quantité de données suffisante et une pertinence élevée. Votre objectif est de rassembler un volume raisonnable de pages web — suffisamment important pour justifier l’usage d’outils de Web Mining (c’est-à-dire un corpus trop large pour être traité manuellement), mais sans tomber dans une collecte massive et non contrôlée. Il est donc recommandé de planifier une stratégie d’exploration : définir des points d’entrée pertinents, extraire uniquement les pages en lien direct avec votre thématique, et mettre en place des filtres (mot-clés, parties de l’URL, profondeur de navigation, structures de liens, etc.) afin d’éviter de collecter des contenus trop éloignés du sujet. L’objectif est d’obtenir un corpus représentatif, cohérent et exploitable, tout en limitant le bruit et les dérivations thématiques.
- **Analyse du code source HTML** : Une fois les pages identifiées, vous devrez analyser le code source HTML pour extraire les données pertinentes. Utilisez les outils vu au cours pour parser automatiquement les pages et récupérer les informations sous un format structuré.

- **Note** : Travailler en anglais permet, dans certains cas, d'avoir accès à une plus grande quantité d'informations, notamment pour des sujets globaux. De plus, cela simplifie parfois la gestion des erreurs d'encodage liées aux accents, ce qui peut être utile lors de la collecte de données depuis des sites multilingues.

3.2 Text Mining

Dans cette partie, vous allez analyser un ensemble de documents textuels en appliquant les concepts et techniques de *Text Mining* vus au cours. L'objectif est de transformer un corpus brut en données exploitables, d'en extraire des informations pertinentes, puis de produire des analyses quantitatives et qualitatives en lien avec votre thématique.

Votre travail doit intégrer les deux dimensions vues en cours : **concepts** et **applications**. Les étapes ci-dessous représentent le **minimum obligatoire**. Vous êtes encouragés à aller au-delà en mobilisant des méthodes plus avancées ou créatives. Si un des points listés ne s'applique pas à votre corpus, vous devez le justifier clairement dans votre rapport et proposer, le cas échéant, une alternative pertinente.

Concepts à mettre en œuvre :

- **Informations sur le corpus** : Présentez les données utilisées (nombre total de documents, provenance, critères de sélection ou de retrait de certains documents). Évaluez brièvement la qualité du corpus (présence de bruit, lacunes, biais éventuels).
- **Prétraitement et tokenisation** : Nettoyez le texte (normalisation, retrait de caractères non pertinents, etc.) et découpez-le en tokens à l'aide d'un *tokenizer* adapté. Justifiez le choix du tokenizer.
- **Représentation vectorielle** : Transformez le texte en vecteurs (TF-IDF, Doc2Vec, etc.) et expliquez le choix de la méthode retenue par rapport à votre problématique.
- **Réduction et filtrage du vocabulaire** : Appliquez, si nécessaire, des techniques telles que la lemmatisation ou la racinisation, l'utilisation d'un dictionnaire de synonymes, le retrait des stopwords, ou l'élimination des termes trop rares ou trop fréquents. Justifiez les valeurs de seuil utilisées.
- **Mesure de similarité** : Choisissez et appliquez la ou les mesures les plus pertinentes pour comparer vos documents (au choix).

Applications minimales :

1. **Analyse descriptive** : Choisissez **au moins deux** des approches suivantes :
   - Fréquence des mots ou n-grams
   - Nuages de mots
   - Concordance
   - Cooccurrence
   - Évolution temporelle des termes ou thèmes clés
2. **Analyse sémantique** : Utilisez soit :
   - Des méthodes d'apprentissage statistique
   - Un thésaurus de sentiment pour détecter des tonalités ou opinions
3. **Groupement ou classification des documents** : Appliquez soit :
   - Une méthode de classification (supervisée)
   - Une méthode de clustering (non supervisée)

   Interprétez les groupes ou classes obtenus en lien avec votre problématique.

3.3 Link Analysis

Une fois que vous avez créé votre graphe en suivant la description spécifique de la thématique choisie, vous devrez analyser les relations entre les nœuds du réseau à l'aide des concepts de *Link Analysis* vus au cours.

Cette analyse vous permettra d'explorer la structure du graphe, d'identifier les nœuds clés et de repérer les connexions essentielles qui définissent l'organisation du réseau. Les concepts suivants représentent le minimum obligatoire que vous devez explorer dans votre projet. Vous êtes cependant encouragés à aller au-delà de ce minimum et à appliquer d'autres techniques pertinentes pour approfondir votre analyse.

Si un des concepts obligatoires ne s'applique pas à votre réseau, vous devez le justifier clairement dans votre rapport. Expliquez pourquoi ce concept est inapplicable dans le contexte de votre thématique et, le cas échéant, proposez des alternatives adaptées à votre réseau.

Voici les principales mesures et outils à explorer pour guider votre analyse :

- **Centralité de degré (Degree centrality)** : Calculez le nombre de connexions pour chaque nœud afin d'identifier les nœuds les plus connectés. Ces nœuds peuvent représenter des entités centrales ou très actives dans le réseau.
- **Plus court chemin (Shortest path)** : Mesurez la distance minimale entre deux nœuds. Cette analyse permet de comprendre la proximité entre différentes parties du réseau et d'identifier les chemins optimaux de circulation de l'information.
- **Centralité d'intermédiarité (Betweenness centrality)** : Identifiez les nœuds jouant un rôle de "pont" dans le réseau, en calculant la proportion de plus courts chemins qui passent par eux. Ces nœuds peuvent avoir une importance stratégique dans la diffusion ou le contrôle de l'information.
- **PageRank** : Évaluez l'importance d'un nœud en fonction du nombre et de la qualité de ses connexions entrantes. Cette mesure permet de repérer les nœuds influents du réseau, souvent cités ou reliés par d'autres nœuds importants.

4 Informations pratiques et évaluation

4.1 Remise du projet

Votre travail de projet doit être remis selon les modalités suivantes :

- **Rapport écrit** : Vous devez remettre un rapport de 12 pages maximum hors annexes éventuelles [1]. Le rapport doit être structuré, clair, concis (ne détaillant que les éléments essentiels) et détailler les étapes de votre projet ainsi que vos résultats. Il doit contenir votre objectif, vos méthodes, vos résultats (graphes/tableaux) et une discussion des résultats. Le titre du rapport doit correspondre au titre du sujet choisi.
- **Code source** : Vous devez fournir un fichier `.pdf` contenant votre rapport intitulé "rapport_MLSMM2153_group_x". Le rapport doit contenir un lien vers un repo Github contenant votre code. Le repo doit au moins contenir un document intitulé "README.md" (ou README.txt) contenant les instructions pour utiliser votre code. Tout le code que vous avez produit pour votre rapport doit être disponible, de manière à permettre la réplication de vos analyses. Ce code doit être bien commenté et organisé pour faciliter la compréhension de vos démarches.

[1] Les annexes ne comptent pas dans les 12 pages de rapport. Attention, utilisez-les à bon escient ! Le document doit pouvoir se lire sans consulter les annexes. Les graphes essensiels doivent être dans le coeur du document

- **Date limite de remise** : Le rapport `.pdf`(incluant le lien vers le repo public Github) doit être soumis au plus tard le 6 janvier 2026 à 23h59.

4.2 Présentation et défense du projet

Vous devrez également présenter vos travaux devant un jury composé de vos deux professeurs. Voici les détails de la présentation :

- **Date de présentation** : La présentation aura lieu le 13 janvier 2026, dans le Local 19.
- **Durée de la présentation** : Vous disposerez de maximum 8 minutes pour exposer votre projet, suivi d'une demi-heure de questions liées à votre projets, vos choix et à la théorie du cours.
- **Support de présentation** : Vous pouvez utiliser un support de type PowerPoint, PDF, etc. Le fichier de présentation doit être envoyé au plus tard la veille avant 23h59. Votre présentation peut inclure une démo live, une vidéo,... tant qu'elle ne dépasse pas la limite stricte de 8 minutes.

4.3 Critères d'évaluation

Votre note finale sera attribuée par le jury en fonction des critères suivants :

- **Qualité du rapport écrit** : La clarté, la cohérence, la structure et la qualité des analyses et conclusions. Votre rapport doit expliquer de manière détaillée les étapes de votre projet et les résultats obtenus.
- **Pertinence des questions et utilisation des outils** : La pertinence des questions posées dans le cadre de votre analyse, et la manière dont vous avez utilisé la collecte de données, le *Text Mining* et le *Link Analysis*. Vous devez démontrer une bonne compréhension de la complémentarité de ces outils et de leur application dans votre sujet.
- **Définition des objectifs et construction de l'analyse** : La qualité avec laquelle vous posez clairement votre objectif de recherche, formulez vos questions, et utilisez les outils de web mining pour y répondre. L'accent est mis sur la capacité à construire une analyse logique, progressive et menant à des conclusions fondées.
- **Réplicabilité du travail** : Votre travail doit être entièrement reproductible. À la simple lecture du rapport, le lecteur doit comprendre toutes les étapes effectuées et être capable de les reproduire aisément.
- **Justification des choix méthodologiques** : Tous les choix réalisés au cours du projet (prétraitement, sélection d'outils, paramètres, visualisations, filtres, etc.) doivent être motivés et clairement justifiés.
- **Qualité du code** : Le code fourni doit être propre, bien commenté, structuré, et réutilisable. Il doit permettre de reproduire l'ensemble du pipeline sans ambiguïté.
- **Quantité et profondeur du travail réalisé** : L'ampleur de l'analyse effectuée et la qualité des explorations supplémentaires. Les démarches allant au-delà du minimum requis, ainsi que la créativité méthodologique ou analytique, sont particulièrement valorisées.
- **Qualité de la présentation** : La clarté de l'exposé, la structuration du discours, la capacité à présenter des résultats de manière concise, ainsi que la capacité à répondre aux questions du jury.

- **Réponses aux questions du projet** : Votre capacité à expliquer vos choix, les résultats obtenus, et à discuter des limites ou alternatives possibles.
- **Réponses aux questions sur la matière du cours** : Votre compréhension des concepts vus en cours et leur application dans le cadre du projet.

Note importante : Au vu des avancées en intelligence artificielle générative, une part significative de l'évaluation se base sur la discussion orale avec le groupe. Le jury accordera une attention particulière à votre maîtrise des étapes réellement effectuées, de vos choix méthodologiques, et des concepts utilisés (ou qui auraient pu être utilisés). Une compréhension approfondie et personnelle du travail présenté est essentie

Bon travail !
