Classification de fiches programmes pour l'aide à la décision

Autheurs : Albeiro Espinal

1 octobre 2018

Abstract

Assistant au choix de formation. Rendre les étudiants acteurs de leur formation.

1 Introduction

Après la fusion entre ex Télécom Bretagne et Mines Nantes l'offre académique de l'école IMT Atlantique est devenue beaucoup plus diverse. Les programmes offerts sont construits sur plusieurs domaines interdisciplinaires : le domaine du numérique, de l'énergie, de l'environnement, de l'électronique, des réseaux, entre autres. Les diverses thématiques d'approfondissements de ces domaines représentent un défi au niveau de la planification des programmes et spécialement un défi au niveau du choix des élèves. En effet, les élèves admis à partir de l'année 2019 suivront une formation très diverse et complexe en termes de modules, qui se déroulera de la manière suivante :

Un tronc commun en première année qui sera identique pour tous les élèves qui font leurs études sur les campus de Brest et Nantes. \item Après ce tronc commun, les élèves devront choisir des Thématiques d'Approfondissements (TAF) en deuxième et en troisième année afin de renforcer leurs connaissances sur les différents domaines de l'école qui composent l'offre académique.

Cette offre académique est constituée par plus de 24 Thématiques d'Approfondissements (TAF). Chaque TAF est composée de plusieurs UE (Il y a plus de 100 UE, unités d'enseignements) et à l'intérieur de chaque TAF une UE peut être classifié comme une UE élective ou une UE coeur. Les UE coeur développent le domaine principal concerné par la TAF tandis que les UE électives permettent d'approfondir sur un domaine spécifique de la TAF.

Alors, dans la construction de son parcours, l'élève se trouve en face de plusieurs paramètres qui rendent le processus décisionnel complexe. Entre ces paramètres se trouvent la quantité des TAF disponibles (plus de 24) et la quantité des UE parmi lesquelles ils pourront choisir (plus de 100).

L'objectif de cette plateforme c'est d'aider les élèves dans leur choix de formation, tout à partir des fiches programmes rédigés par les concepteurs de TAF et UE de l'école. À partir de la classification automatique des fiches programmes et d'un document qui décrit le parcours d'un élève (par exemple son CV ou son profil Linkedin) ou d'un document qui décrit son projet d'études (une lettre de motivation par exemple), le système guidera l'élève dans la choix de TAF et UE offertes par l'école.

2 Classification des Documents

La classification des documents est devenue un problème central aujourd'hui face à l'accroissement exponentiel de la quantité d'information disponible. La plupart des documents traités aujourd'hui par l'industrie sont des documents web disponibles sur le World Wide Web : des mails, des documents liés à la structure d'un site web, des documents médicaux, des documents qui sont extraits des réseaux sociaux, entre autres. Aujourd'hui, pour la classification des documents les algorithmes d'apprentissage sont devenus de plus en plus importants pour l'industrie. De façon

générale, il y a deux types d'algorithmes d'apprentissage : les algorithmes supervisés et les algorithmes non supervisés. [1]

Les algorithmes de clustering supervisés permettent de classifier un ensemble d'objets à partir d'un exemple de classements initiaux. Cet exemple initial permet à la machine d'apprendre à classifier un document. Par contre, pour pouvoir mettre en place ces types d'algorithmes, il est nécessaire d'avoir un ensemble d'exemples initiaux très representatifs de la population à analyser et qui représentent la plupart de la variance de cette population.

Il y a d'autres approches statistiques qui permettent de classifier des documents grâce à l'identification des sujets latents d'un ensemble des documents. Un des plus importants, c'est le modèle probabiliste LDA (Latent Dirichlet Allocation). L'objectif de ce modèle, c'est de trouver une description courte d'un ensemble de documents à travers l'identification des sujets latents qui préservent les relations statistiques entre les documents. Cela permet de classifier des documents de la manière suivante : en premier lieu, les sujets latents sont identifiés et représentés sous la forme de distributions des mots présents dans le corpus. Les mots appartenant à un même sujet ont une tendance à apparaître ensembles dans les documents. Après l'identification de chaque sujet latent, la contribution de chaque sujet à chaque document est calculée. [2]

Le modèle LDA a plusieurs variantes qui peuvent être classifiées en deux catégories principales : des modèles basés sur l'échantillonnage et des modèles basés sur l'optimisation. Parmi les modèles basés sur l'échantillonnage, LDA Gibbs Sampling est l'un des plus utilisés. Ce modèle utilise les méthodes de Monte Carlo par chaînes de Markov dans le processus d'identification des sujets et à long terme produit des sujets plus cohérents par rapport aux modèles basés sur l'optimisation. Un des problèmes des modèles basés sur l'échantillonnage, se trouve au niveau de la performance qui est très inférieure à celle des modèles centrés sur l'optimisation. Pour cela, les modèles basés sur l'échantillonnage sont plus utilisés quand la population de documents est réduite. En général, les modèles LDA produiront une classification non supervisée définie par la co-occurrence de mots entre les documents. Cela produit souvent des problèmes de cohérence parmi les sujets générés qui ne sont pas facilement compréhensibles par l'utilisateur. Ce problème a motivé la naissance des modèles LDA guidés qui cherchent aussi à identifier des sujets latents mais qui permettent à l'utilisateur de spécifier les mots de base de chaque sujet afin de guider la classification des documents vers une direction souhaitée.

La plupart des modèles décrits dans cette section demandent une représentation numérique du texte d'un document. En général, cette représentation se fait sous forme de vecteurs ou de dictionnaires qui associent un nombre, obtenu d'une fonction prédéfinie, à chaque mot. Les deux représentations vectorielles et matricielles les plus importants des documents sont les représentations TF-IDF (term frequency–inverse document frequency) et Hashing. [3] Aujourd'hui, la représentation TF-IDF est la plus utilisée. Environ 83\% des systèmes de recommandation l'utilisent [4]. Dans cette représentation, les fréquences de chaque mot dans un texte sont calculées et, à partir de ces fréquences, un vecteur est créé où chaque dimension correspond à la fréquence (term frequency) d'un mot du texte. Ces fréquences sont pénalisées à travers la fréquence inverse du mot (inverse frequency). Cette fréquence inverse permet de donner un poids très faible aux mots qui apparaissent beaucoup dans le texte (en supposant qu'ils ne sont pas importants pour décrire l'essentiel du texte) et un poids plus important aux mots qui apparaissent moins et qui expriment probablement le coeur du document.

Finalement, l'efficacité des résultats des modèles de classification ici décrits dépend très fortement de la qualité du processus de nettoyage des documents qui cherche à supprimer le bruit (des symboles de ponctuation, des mots creux...). Il y a plusieurs librairies sur le marché qui permettent de nettoyer les textes de manière automatique comme NLTK ou Spacy qui offrent des listes des mots creux, des listes des symboles de ponctuation, des fonctions pour la reconnaissance des entités dans un texte et, en général, des fonctionnalités diverses pour réduire le nombre de mots redondants d'un texte comme la racinisation et la lemmatisation. [1]

3 Documents Analysés

Plus de 130 documents qui contiennent les descriptions des TAF et UE de l'école ont été analysés. Parmi ces documents, 24 documents correspondent à la description des TAF et 106 documents décrivent les contenus des UE électives et UE coeurs. Ci-dessous quelques statistiques sur le contenu des documents qui décrivent les TAF. Après avoir enlevé les symboles de ponctuation, contractions et accents de la langue française, la densité lexicale l_i du document d_i a été définie et calculé à travers l'équation suivante :

$$l_i = \frac{w_i}{x_i} \text{ [5]}$$

où w_i est la quantité total des mots du document d_i et x_i correspond à la quantité totale des mots uniques de ce même document. Ci-dessous, la densité lexicale de chacun de 10 documents de TAF analysés :

Document	**Nombre de Mots**	**Nombre de Mots Uniques**	**Densité Lexicale**
11_B_et_isd.pdf	8408	1778	21.15%
08_BN_HEALTH.pdf	3223	1064	33.01%
19_B_OPE.pdf	2729	907	33.24%
03_N-COPSI V5.pdf	2417	810	33.51%
22_B_et_SEH_v3.pdf	2391	772	32.29%
12_N_et_LOGIN.pdf	2332	767	32.89%
05_B_DaSci.pdf	926	465	50.22%
04_R_Cyber.pdf	877	440	50.17%
21_N_Robin.pdf	871	415	47.65%
06_BN_DCL.pdf	867	426	49.13%

Table 1: Densité Lexicale des Documents de TAF

Et le diagramme de dispersion Nombre de Mots vs Densité Lexicale sur l'ensemble des documents :

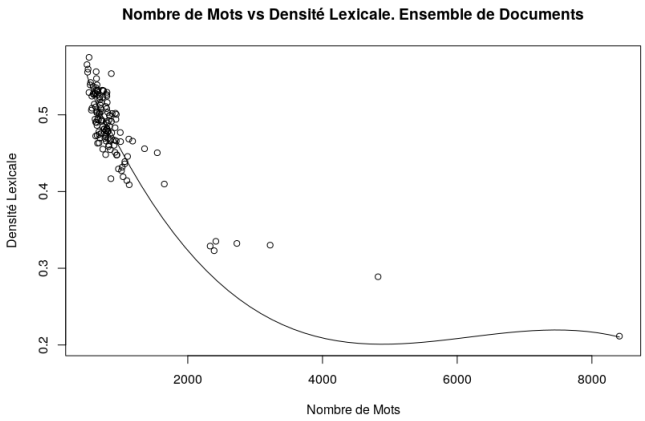


Figure 1: Diagramme de dispersion sur l'ensemble de documents. Nombre des Mots vs Densité Lexicale

La densité lexicale moyenne de l'ensemble de documents de TAF est de 46%. La plupart des documents ont moins de 500 mots uniques et une densité lexicale autour de 50%. Cette homogénéité est une conséquence du modèle de document à remplir fourni aux responsables qui devaient fournir le même type d'informations sous une même structure du modèle de document. Il faut remarquer qu'il y a des documents qui ont un nombre de mots uniques très supérieur aux autres et des autres documents qui ont une quantité faible des mots et qui risquent de

ne pas représenter complètement le domaine de la TAF (après filtrage des mots creux, la quantité des mots uniques sera beaucoup plus faible). Ce risque est réduit par une des fonctionnalités du système de recommandation qui permet de joindre des documents existants. Cela permettra, par exemple, de joindre des documents d'une TAF avec les documents des UE Coeur (qui contiennent le vocabulaire du coeur du métier) respectives afin de construire des documents plus représentatifs du domaine spécifique. Finalement, par rapport aux documents plus gros, il faut remarquer que ces documents risquent de contenir beaucoup de bruit (des mots creux) qui peuvent rendre plus difficile le processus de classification et d'identification des sujets. Alors, un facteur clé d'une classification efficace des documents, c'est un processus de nettoyage qui doit être capable d'enlever la plus grande partie de bruit présent dans les documents.

4 Méthodes

Avant de mettre en place la classification automatique de fiches de programmes pour la recommandation de contenu académique un processus rigoureux de nettoyage des documents est mis en place afin d'enlever la plupart de bruit.

4.1 Réduction de Bruit

Dans un document il y a plusieurs types de bruit à considérer :

- Les symboles de ponctuation.
- Les symboles spéciaux produits par le format PDF des documents.
- Les mots creux (des articles, des adverbes, des mots qui ont un sens vide dans le contexte d'application) qui peuvent compliquer l'identification des sujets plus importants dans les textes.

Pour le premier type de bruit, il y a plusieurs types d'outils sur le marché comme Spacy et NLTK qui permettent d'enlever les symboles de ponctuation dans un texte. Cela est fait souvent grâce aux fonctionnalités POS (Part of Speech) qui permettent de détecter le rôle d'un terme selon le contexte du document. Il y a aussi des listes prédéfinies de symboles de ponctuation fournies par ces librairies qui aident dans le nettoyage de ce type de bruit.

La fonctionnalité POS de Spacy a été testée afin d'enlever les symboles de ponctuation mais son efficacité n'était pas très satisfaisante. En effet, quand dans le texte il y a des mots ensemble avec des symboles de ponctuation, cette fonctionnalité de Spacy n'est pas capable de les détecter d'une manière efficace. La librairie NLTK présent des problèmes similaires quand il y a des symboles de ponctuation ou des symboles spéciaux joint avec les mots du texte. La fonction NLTK *word_tokenize*, qui permet de décomposer un texte en tokens (unités de sens minimales d'un texte) souvent n'est pas capable de séparer de manière propre les caractères non alphanumériques. Par exemple, dans le texte de l'UE Réacteurs Nucléaires, la librairie Spacy présent des problèmes au moment de décomposer le texte en tokens et dans la détection du rôle du terme dans le texte, par exemple :

Token	Rôle
Semestre(s	NOM
concerné(s	NOMBRE
Langue(s	NOM PROPRE
%	NOM
d'	NOM PROPRE
*	AUXILIAIRE
d'	ADPOSITION

Table 2: Exemple d'inconsistances trouvées de la fonction de lemmatisation de Spacy pour la langue française

Parmi les inconsistances l'expression «Semestre(s» a été identifiée comme une unité de sens minimale par la fonction de décomposition en tokens de Spacy pour la langue française. De plus, il y a des inconsistances dans la fonctionnalité POS (Part of Speech) car le token «d'» est identifié comme un nom propre. De plus, il y a des symboles, «%», détectés comme des noms. Ces sont des exemples sur les inconsistances souvent trouvées dans le nettoyage avec les fonctions fournies par la meilleure librairie industrielle pour le traitement de langage naturelle disponible sur le marché pour la langue française.

Ces inconsistances trouvées dans les libraries Spacy et NLTK ont motivé la création d'une fonction générale afin de nettoyer des textes de la langue française. Cette fonction utilise des listes de symboles identifiés, des expressions régulières et la fonctionnalité POS de la librairie NLP Stanford afin de réduire chaque type de bruit (cette librairie est utilisée spécialement pour détecter les verbes qui dans la plupart de cas ne font pas référence aux sujets principaux du document). Cette fonction créée utilise aussi des listes de plus de 1500 mots creux identifiés de la langue française et des mots vides par rapport au contexte de l'école et son offre académique. Ci-dessous, un exemple d'une partie d'un texte avant et après le processus de nettoyage mis en place :

Avant	Après
● métiers à la sortie : ○ Ingénieur R&D dans le domaine de la robotique de production ou de service. ○ Ingénieur conception et modélisation de systèmes robotisés complexes et spécifiques (machines spéciales) ; ○ Chef de projet, Ingénieur concepteur de systèmes robotisés modulaires et versatiles ; ○ Ingénieur contrôle commande de système robotisés et de système cobotique ; ○ Ingénieur conception de logiciel d'interface homme / machine.	metiers sortie ingenieur domaine robotique production service ingenieur conception modelisation systemes robotises complexes specifiques machines speciales chef projet ingenieur concepteur systemes robotises modulaires versatiles ingenieur controle commande systeme robotises systeme cobotique ingenieur conception logiciel interface homme machine

Table 3: Partie d'un texte avant et après nettoyage

4.2 Racinisation (Stemming)

Après le nettoyage des données, il y a encore des mots répétés et, plus important, des mots qui se trouvent dans le texte au singulier et au pluriel. La représentation du texte cible, TF-IDF, n'est pas conçue pour différencier ce type de cas. Les mots robotises, robotisé et robotisés sont des mots différents sous cette représentation. Le premier mot est différent parce qu'il n'y a pas d'accents, un problème qui est adressé, grâce à une fonction du système qui enlève les accents de la langue française. Dans les deux derniers mots, le premier mot est au singulier et le deuxième mot au pluriel. La méthode de stemming utilisée est la méthode SnowBall déjà mise en place dans la librairie NLTK et qui est adaptée pour la langue française. Cette méthode cherche à obtenir une forme tronquée du mot qui sera commune aux variantes morphologiques. Pour le faire, cette méthode enlève les suffixes et les flexions présents dans les mots. Par exemple : ingénieurs et ingénieur appartiennent à la même forme tronquée ingénieur.

Le processus de racinisation a des inconvénients. Parfois, il l'y a des mots qui ont un sens différent mais qui sont tronquées dans une variante morphologique commune. Par exemple, les mots automatique et automatiquement ont la racine «automat» mais sont des mots souvent utilisés dans des contextes différents et ayant des significations différentes dans le cas des TAF. L'alternative au processus de racinisation, c'est la lemmatisation. Mais la meilleure librairie de traitement de langage naturelle pour la langue française, Spacy, qui dépend de sa fonctionnalité POS pour trouver le lemme d'un mot donne souvent des résultats qui sont complètement éloignés du sens réel des mots dans un contexte déterminé. C'est le cas du document de la TAF Sciences de Données où le lemme associé au mot «données» est souvent le verbe donner. Pour l'adjectif «complexe» parfois le lemme associé est le verbe «complexer». Dans la phrase «Je suis» dans un contexte où "suis" fait référence au verbe suivre, les lemmes respectifs sont «Je suivre». Alors, ce processus de lemmatisation souvent produit des résultats qui changent le sens des mots plus importants des textes. Par cela, le processus de racinisation a été choisi pour unifier les mots répétés et trouver la forme canonique des mots après le nettoyage.

4.3 Représentation Vectoriel des Textes : TF-IDF

Pour la mise en place des algorithmes de classification et de détection des sujets, en plus d'un très bon processus de nettoyage des textes, il faut avoir une représentation numérique des textes. Il y a deux types de représentation très populaires. La représentation TF-IDF et la représentation Hashing. Ces deux représentations ont des avantages et des inconvénients selon le contexte. Dans le cas des fichiers des TAF, la densité lexicale montre que pour la plupart des documents, au moins la moitié du texte contient des mots répétés et du bruit . La représentation plus appropriée dans ce cas, c'est la représentation TF-IDF. L'objectif de cette représentation c'est d'extraire les mots qui décrivent l'essentiel de chaque texte, et à la différence de la méthode Hashing Vectoriser, elle pénalise les mots qui apparaissent beaucoup dans l'ensemble du corpus.

Pour la représentation vectoriel de chaque document, la méthode TF-IDF, utilise deux paramètres. Le premier paramètre, c'est la fréquence du mot. La fréquence du mot estime dans quelle proportion un mot apparaît dans un texte. Cette fréquence est normalisée en la divisant par la quantité des mots dans le document. Plus spécifiquement, pour le terme i, la fréquence du terme est égal à :

$$TF_i = \frac{t_i}{m} \, (1)$$

t_i: nombre de fois que le mot i apparaît dans le texte, m est nombre total de mots du texte.

Le deuxième paramètre, c'est la fréquence inverse du terme. Ce paramètre permet de mesurer l'importance d'un terme dans un ensemble des documents [5]. Alors, pour le terme i, le paramètre est calculé grâce à l'équation suivant :

$$IDF_i = log\frac{d}{m_i} \, (2)$$

où d est la quantité des documents et m_i est la quantité des documents qui contiennent le mot i.
Alors, la fréquence finale du mot i d'un document sera donc :

$$F_i = TF_i * IDF_i \, (3)$$

Ci-dessous, un exemple : les mots les plus importants du document de la TAF Ingénierie Nucléaire après réduction de bruit :

4.4 Système d'inférence : Latent Dirichlet Allocation

La détection des sujets sur l'ensemble des documents est faite à l'aide de plusieurs variations du modèle de machine learning LDA (Latent Dirichlet Allocation), très populaire pour la classification des documents, et qui permet d'entraîner un modèle dans la prédiction des sujets inhérents au corpus. En effet, après l'entraînement, le modèle produira un ensemble de sujets intrinsèques au corpus (d'un point de vu de la distribution et de la co-occurrence

Mot	IDF_i Score
nucléaire	0.73
énergétique	0.25
13n	0.19
industrielle	0.19
numérique	0.19
combustible	0.13
cycle	0.13
transitions	0.13
gestion	0.09
instrumentation	0.09
médicales	0.09
physiques	0.09
radioprotection	0.09

Table 4: Mots Plus Importants du Document d'ingénierie Nucléaire sous TF-IDF

de mots présent dans les documents), et permettra de prédire, pour un document donné, quelle est sa composition en terme de ces sujets détectés.

La première variante du modèle LDA Standard utilise Variational Bayes Online Learning, méthode très performante [2] quand il y a des grandes quantités des documents à classifier en peu de temps[Blei]. La deuxième variation du modèle LDA utilisée pour la génération des modèles, c'est le modèle LDA par échantillonage de Gibbs (Connu comme LDA Mallet sous la librairie Mallet de Java). Ce modèle produit des sujets plus cohérents en un ensemble petit de documents et utilise des méthodes de Monte Carlo par Chaînes de Markov, ce qui implique une perte de performance par rapport le modèle qui utilise Variational Bayes Online Learning mais qui permet d'obtenir des sujets plus cohérents et plus robustes dans un ensemble de documents réduit. La troisième variante du modèle LDA utilisé pour l'entraînement de modèles de prédiction, c'est le modèle LDA Guidé [6], qui est une variante du modèle LDA par échantillonage de Gibbs (LDA Mallet). Les modèles LDA et LDA Mallet, avant de commencer l'entraînement et l'identification des sujets, demandent de spécifier le nombre de sujets à détecter. Pendant cet entraînement, la détection des sujets est faite à partir de la distribution et la co-occurrence des mots présents dans le corpus. Cela produit souvent des sujets qui sont difficiles à comprendre pour un utilisateur quelconque. Cet inconvénient a motivé la naissance d'autres variantes du modèle LDA. Une de ces variantes, c'est le modèle LDA Guidé, qui permet à l'utilisateur de spécifier les mots de base de chaque sujet au début de l'entraînement afin d'obtenir des sujets plus souhaitables.

Pour entraîner tous ces algorithmes il faut spécifier la quantité de sujets à détecter sur le corpus. Une métrique très utile pour la détection du nombre optimal des sujets dans l'ensemble des documents choisis, c'est la mesure de cohérence du modèle [7]. Cette mesure aide à différencier entre des mauvais et des bons modèles générés à travers l'algorithme LDA grâce à des analyses intra et extra sujet. Par exemple, pour l'ensemble des documents de l'école analysés cette métrique a été exécuté sur l'ensemble des TAF (sous l'hypothèse que les documents TAF sont un sous ensemble très représentatif de la variance de la population des documents) et sur tout l'ensemble des documents :

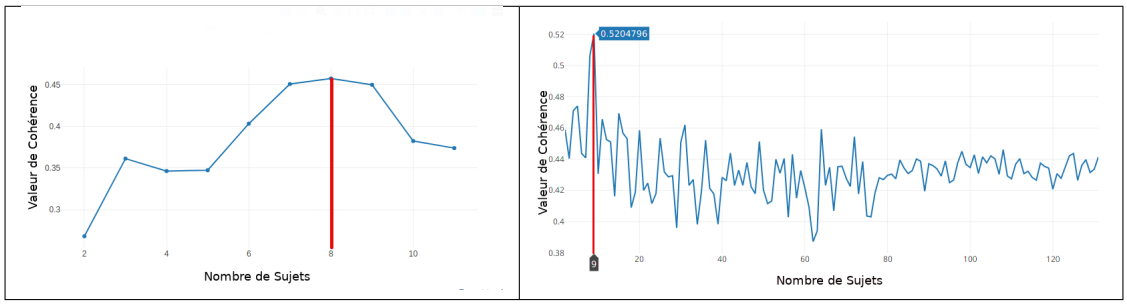


Figure 2: Métrique de Cohérence de Sujets sur l'ensemble de 24 TAF (à gauche) et sur l'ensemble de tous les documents (à droite)

Les résultats trouvés montrent dans les deux cas que la quantité optimale des sujets pour l'ensemble des documents analysés se trouve très probablement entre 7 et 9. Ce facteur est observé spécialement dans le deuxième graphique où la cohérence du modèle avec 10 sujets baisse très fortement. Ce résultat est supporté par les résultats données par le Module de clustering de documents du système de recommandation qui permet d'exécuter divers algorithmes de clustering, principalement l'algorithme K-Means et l'algorithme de clustering Hiérarchique afin de trouver des relations de similitude entre les documents et entre les clusters des documents. Ce module a identifié le nombre optimal de clusters entre 8 et 9.

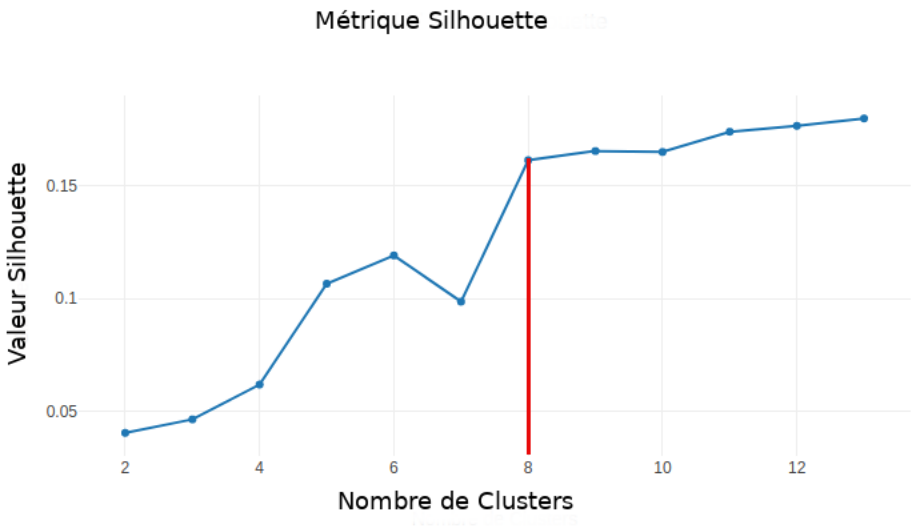


Figure 3: Nombre optimal de clusters. Métrique de Silhouette [8]

À partir de l'identification du nombre optimal de sujets selon les diverses métriques de clustering et de cohérence des sujets, les modèles LDA Standard et LDA Mallet ont été entraînés sur l'ensemble de documents plus représentatifs. Il faut remarquer que les paramètres de chaque modèle ont été adaptés afin d'obtenir des meilleurs résultats sur un ensemble de documents réduit. Ci-dessous quelques sujets qui ont été produits par le modèle LDA Standard sur l'ensemble de TAF :

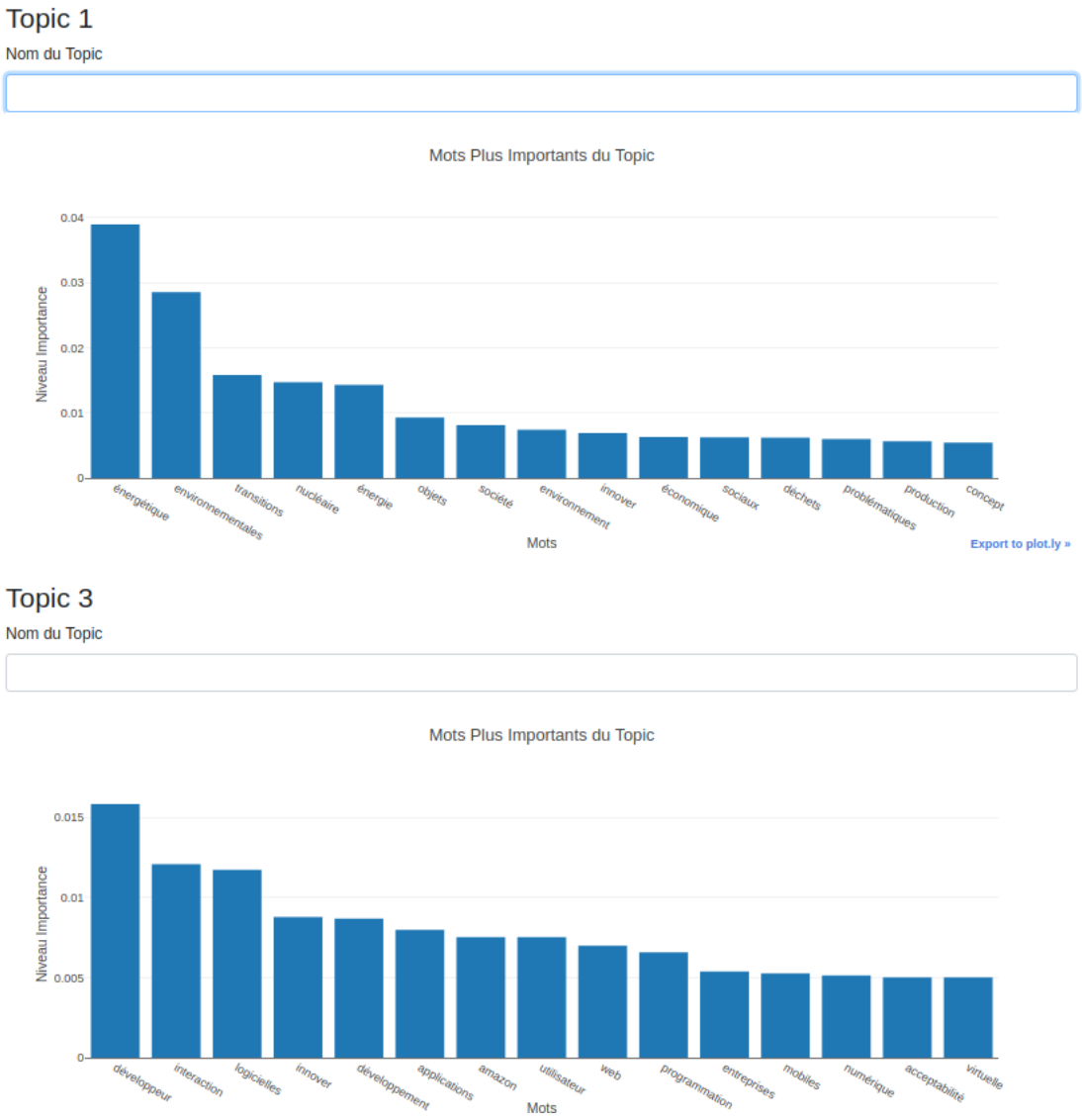


Figure 4: 4 sujets détectés par le modèle LDA Standard

5 Fonctionnalités de la Plateforme

À partir de ces fonctionnalités de nettoyage des documents et d'un moteur d'inférence construit autour des modèles d'analyse de distribution de mots, la plate-forme offre un ensemble de fonctionnalités utiles pour l'aide à la décision des élèves dans leur choix de TAF et d'UE de la nouvelle offre académique. Ci-dessous, les fonctionnalités plus importantes.

5.1 Génération de Modèles de Classification Automatique

Trois modèles déjà décrits sont utilisés pour la génération automatique de modèles d'inférence de sujets des documents. Le premier modèle, c'est le modèle LDA qui utilise Online Variational Bayes. Le deuxième modèle, LDA Mallet, offre une phase d'entraînement plus robuste : en échange d'une perte de performance, ce modèle a produit

des sujets plus cohérents dans un ensemble réduit de documents. Ci-dessous, quelques sujets produits par le modèle LDA Mallet :

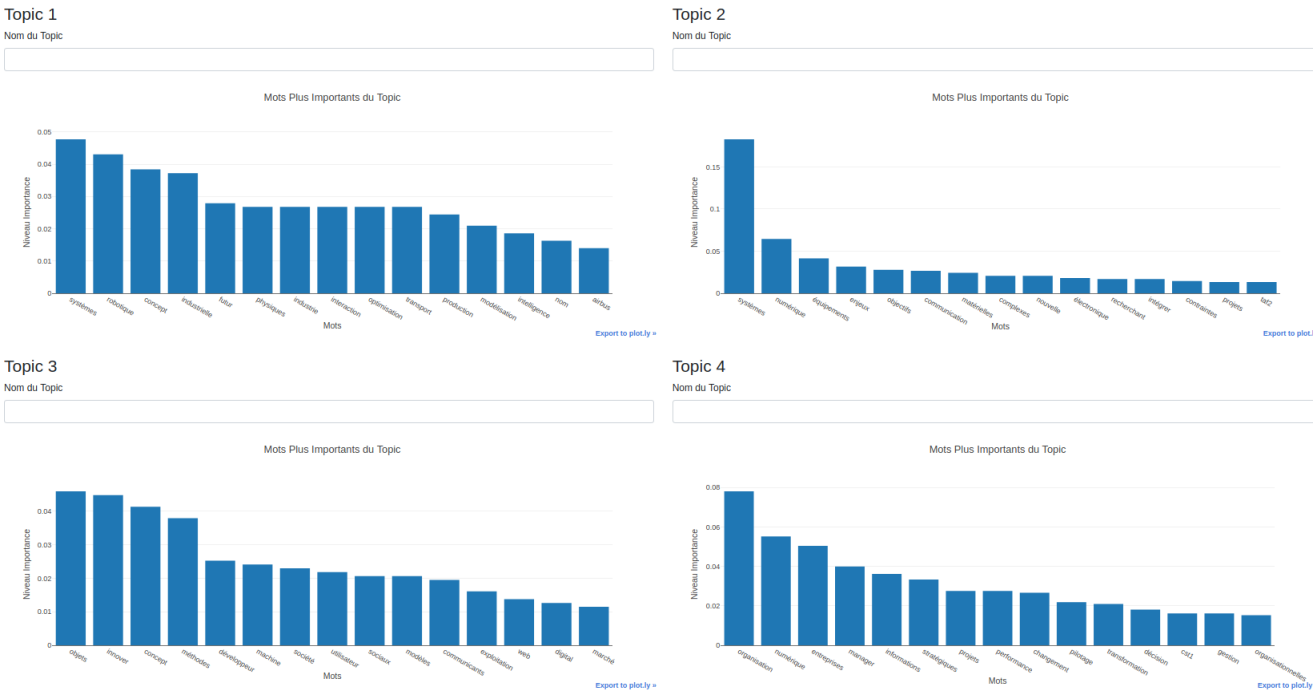


Figure 5: Sujets générés par le modèle LDA Mallet

Le modèle LDA Mallet produit souvent des sujets plus intéressants et plus cohérents. Parmi les sujets détectés se trouve par exemple un sujet constitué par les documents des TAF Internet d'Objets, Objets Communicants, Ingénierie de Systèmes Communicants et Systèmes Embarqués qui fait référence aux TAF liés au domaine d'objets communicants. Il y a aussi un sujet lié aux énergies renouvelables et l'environnement qui comprend les TAF liées au secteur énergétique, la TAF Ingénierie Nucléaire et la TAF Transitions Énergétiques et de l'Environnement. Il y a un autre sujet lié à l'informatique où se trouvent les documents des TAF Développement Collaboratif de Logiciels, Ingénierie de Logiciel et Ingénierie de Systèmes Distribués. Par contre, la TAF Interface Homme Machine très liée au domaine de l'informatique se trouve parmi des sujets liés à la conception et design avec des autres TAF comme Objets Communicants. Cela peut devenir un problème si un des sujets détecté dans la phase d'entraînement n'est pas ce que l'utilisateur espère trouver.

À cause de cette possible situation qui peut introduire des sujets non souhaités parmi les modèles entraînés, un troisième type de modèle est offert par le système de recommandation : le modèle LDA Guidé. Ce modèle permet de fournir des mots de base pour chaque sujet à générer afin de guider le processus itératif du modèle LDA Mallet. Ces mots clés deviennent les mots plus importants de leurs sujets respectifs et s'il y a vraiment un sujet derrière lié à ses mots, très probablement il sera identifié. Très souvent, fournir de manière manuelle les mots clés d'un sujet n'est pas suffisant pour vraiment guider le modèle dans la direction souhaité. Par cela, l'application permet à l'utilisateur de spécifier, pour chaque sujet, la liste des documents qui apporteront les mots clés. Par exemple, si l'utilisateur souhaite produire un sujet lié aux énergies et l'environnement, il peut choisir le document de la TAF Transitions Énergétiques et de l'Environnement comme document source de mots clés pour le sujet respectif. Avant de commencer l'entraînement, le système de recommandation recueille les documents sources de mots clés fournis par l'utilisateur et trouve les mots uniques de chaque document par rapport aux autres. Cela permet d'éviter des collisions des mots parmi les mots clés de plusieurs documents fournis. Par exemple, si l'utilisateur, pour de générer les mots clés d'un sujet, fournit un document lié au domaine de réseaux et pour un autre sujet il fournit un document

lié au domaine d'informatique, très probablement il y aura des mots en commun dans les deux documents qui sont très proches entre l'un de l'autre. Alors, le système filtre les mots clés qui sont communs à plusieurs sujets parmi les documents sources de mots clés fournis.

Ci-dessous, un modèle géneré à partir du modèle LDA Guidé avec 6 sujets plus générales prédéfinis par l'utilisateur (à travers des mots sources) :

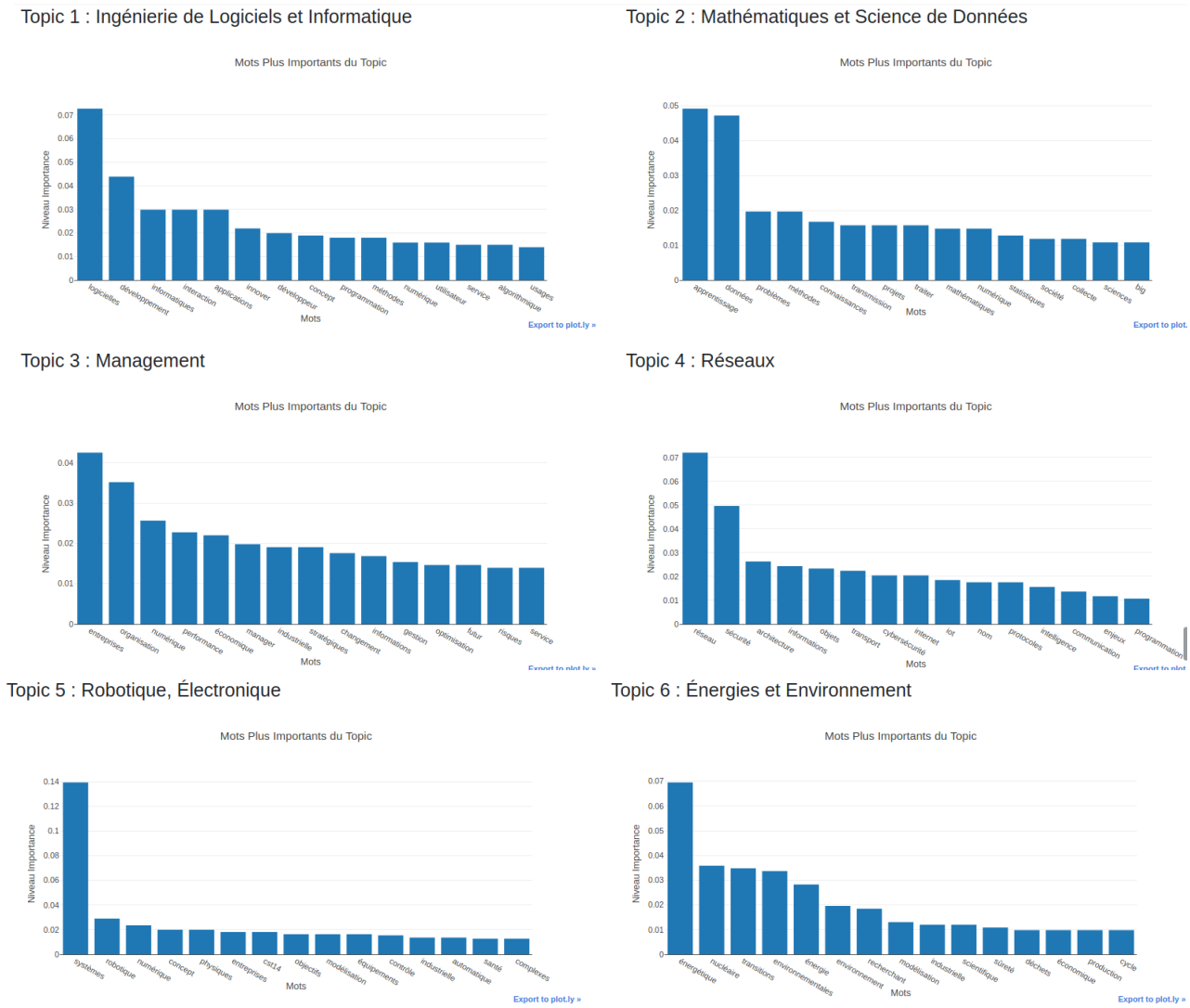


Figure 6: LDA Guidé. Détection de 6 Sujets

5.2 Classification Automatique de Documents

À partir des modèles générés, le système permet de de visualiser l'ensemble des documents existants, les documents plus recommandés de chaque sujet. Ci-dessous les documents de TAF et d'UE recommandés pour le sujet de réseaux et de robotique/physique :

Contribution du Topic	TAF/UE
0.601	TAF ingénierie logicielle et innovation V. 1
0.533	TAF développement collaboratif et multisites de logiciels V. 1
0.517	TAF interaction hommemachine et systèmes collaboratifs V. 1
0.493	TAF ingénierie logicielle des systèmes distribués V. 1
0.280	UE Développement d'applications sur dispositifs mobiles V. 1
0.254	TAF conception d'objets communicants V. 1
0.251	UE Prototypage Rapide : au-delà du design thinking V. 1
0.235	UE Evaluation de l'acceptation des NTIC V. 1
0.224	UE Compression et codage V. 1
0.222	UE Intelligence du Web pour l'IoT et l'IHM V. 1
Contribution du Topic	TAF/UE

Contribution du Topic	TAF/UE
0.632	TAF automatique et systèmes cyberphysiques V. 1
0.594	TAF systèmes embarqués et hétérogènes V. 1
0.559	TAF robotique et interactions V. 1
0.364	UE Commande et observation des systèmes dotés d'actionneurs et capteurs multiples V. 1
0.346	UE Robotique bio-inspirée V. 1
0.337	UE Contrôle des robots V. 1
0.331	UE Transports intelligents V. 1
0.328	UE Modélisation et identification des systèmes dynamiques V. 1
0.321	UE Architecture informatique et implémentation numérique V. 1
0.321	UE Commande robuste des systèmes dynamiques V. 1
Contribution du Topic	TAF/UE

Figure 7: Recommandations pour les sujets d'informatique (à gauche) et de Robotique/Physique (à droite)

5.3 Regroupement de Documents

Afin d'enrichir le vocabulaire d'un sujet spécifique, l'utilisateur peut joindre des documents sur la plate-forme. Cela permet de créer des documents plus représentatifs d'un domaine spécifique. Cela a été très utile quand, par exemple, certains documents ont un vocabulaire très général qui n'est pas très représentatif du domaine. Dans le cas d'application, certains documents de TAF présentaient ce problème qui a été résolu grâce à la fonctionnalité de jonction des documents. Ces documents problématiques de TAF ont été joints avec les documents de leur UE Coeur respectifs (qui contiennent le vocabulaire du coeur du métier) ce qui a permis de produire des documents plus représentatifs du domaine et d'un point de vue sémantique plus significatifs et plus efficaces dans les processus d'entraînement de modèles et de détection de sujets.

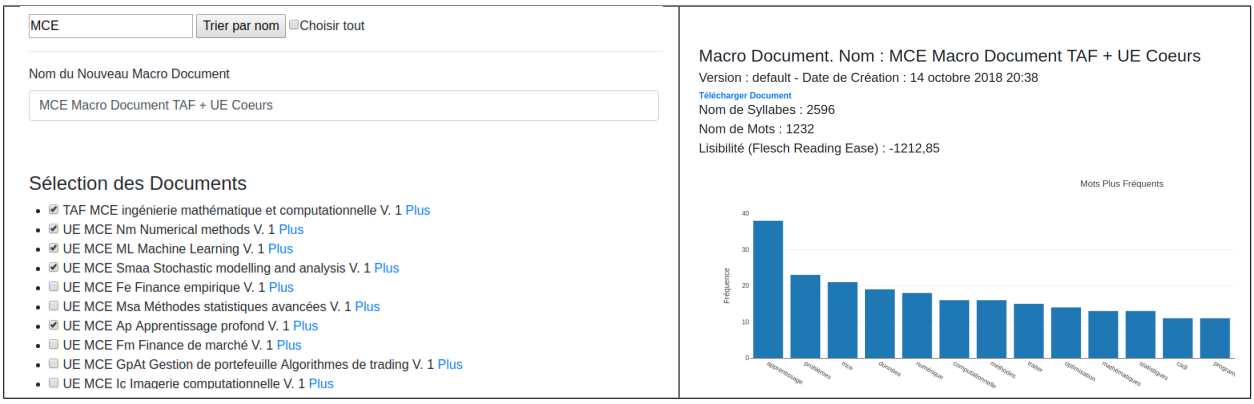


Figure 8: Regroupement des documents. TAF MCE + UE associées

5.4 Recommandation de Contenu Académique Aux Élèves

Les fonctionnalités de production de modèles du système de recommandation permettent de construire un écosystème de raisonnement qui permet de guider les élèves dans leurs choix de formation. Les élèves peuvent télécharger leurs CV qui sont traités et soumis aux modèles déjà disponibles dans le système afin d'identifier la contribution de chaque sujet du modèle au document fourni par l'élève. De cette manière, à partir d'un CV, le système est capable

de recommander du contenu académique selon le parcours de l'élève. Aussi, à partir de la lettre de projet d'études d'un élève le système est capable de recommander le contenu académique plus similaire à son projet d'études ou à un document appartenant à une offre professionnelle souhaitée par l'élève .

De plus, le système de recommandation montre du contenu qui peut être complémentaire pour l'élève. Par exemple, pour un élève dont le CV appartient principalement au sujet de Science de Données; à partir de l'analyse de distribution des sujets sur l'ensemble de documents, le système est capable de recommander du contenu académique qui appartient principalement aux autres domaines (Physique, Électronique, Réseaux...) mais qui ont une contribution importante au sujet de Science de Données. Cela permet aux élèves de diversifier leur parcours en explorant du contenu académique qui est complémentaire à leur domaine d'étude.

6 Résultats

La plate-forme a permis de générer plusieurs modèles représentatifs de l'offre académique de l'école. Ci-dessous, un de modèle généré qui contient 6 sujets correspondant aux sujets plus générales de la nouvelle offre de l'école :

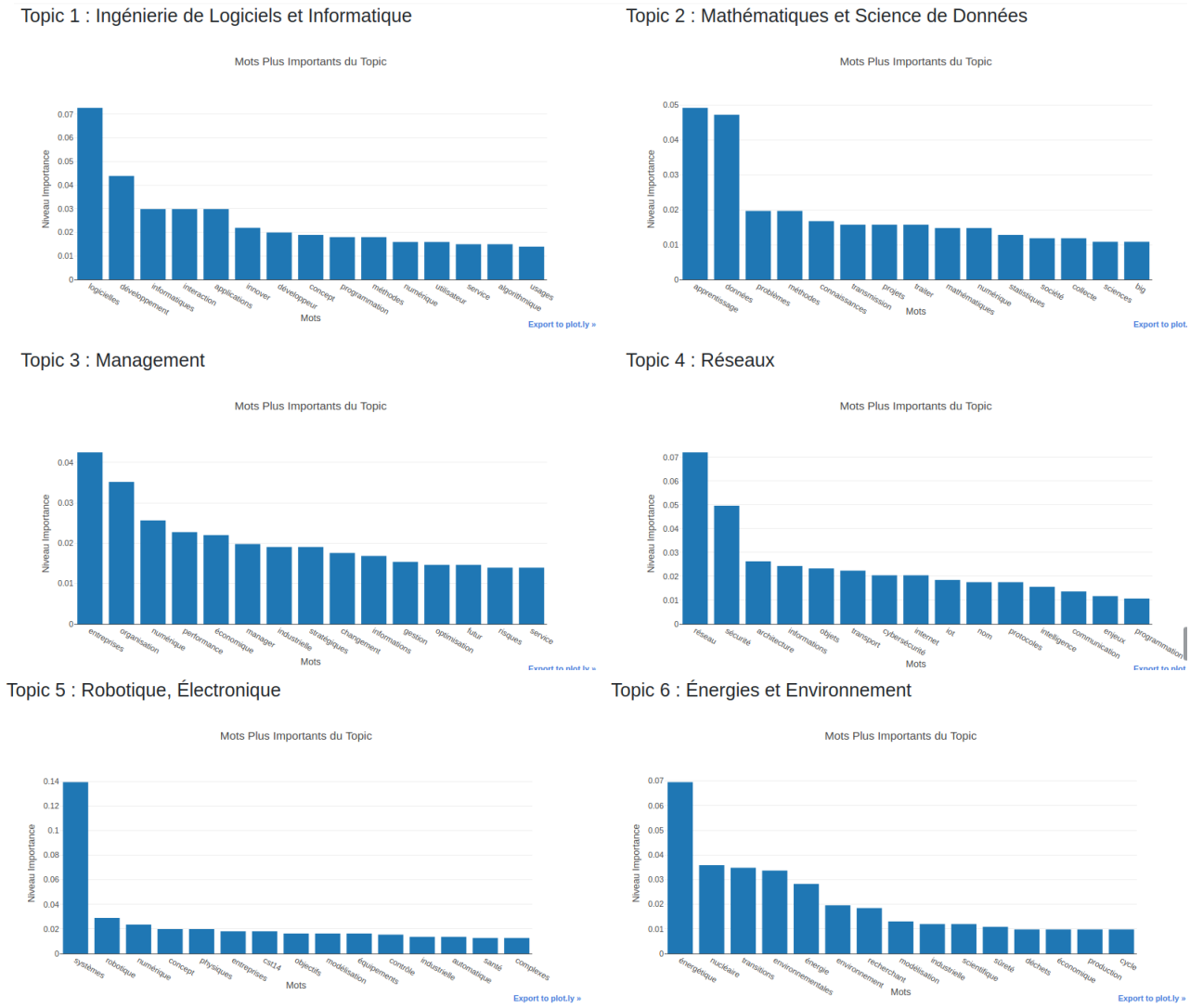


Figure 9: LDA Guidé. Génération de 6 Sujets

Ces modèles ont permis de générer des recommandations académiques pour les élèves à partir des documents diverses. Ci-dessous, un exemple qui illustre la détection de sujets automatique sur un CV d'un élève de l'école :

Stage : Ingénieur-Chercheur Data Science

Formation

2017 - 2019 **Ingénieur Généraliste - Deuxième Année** IMT Atlantique, Brest
Data Science, mathématiques, probabilité et processus stochastiques, réseaux informatiques, circuits numériques, télécommunications, traitement de signal, sciences humaines et finances .

2012 - 2019 **Ingénieur en Systèmes et Informatique - Double Diplôme IMT** Université Nationale de Colombie - Medellin
Conception, développement et gestion des systèmes informatiques.

Expérience

04/18 - 05/18 **Data Scientist - Projet Big Data UV** IMT Atlantique
Complétion de données d'un réseau d'utilisateurs de Linkedin grâce à l'implémentation d'algorithmes de traitement et détection des communautés. Méthodologie CRISP-DM. **Technologies :** NetworkX, Python

01/18 - 05/18 **Développeur - Projet d'Ingénieur (350 h)** IMT Atlantique
Développement d'une Application GPS Android et d'un site web pour faire des balades guidées à Brest. Méthodologie Scrum. **Technologies :** Android API, Mapquest GPS API, FTP Protocol, Bootstrap, CSS, Javascript, GIT

Figure 10: Exemple du CV d'un élève de l'école

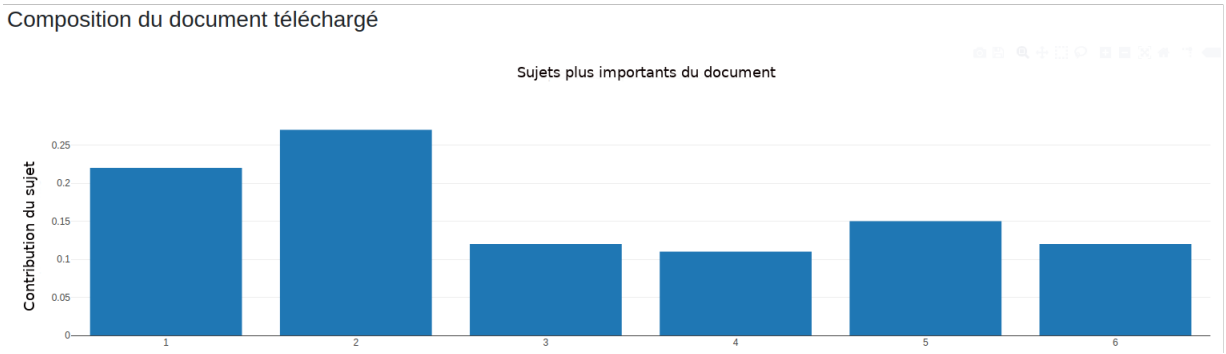


Figure 11: Détection de sujets sur le CV téléchargé par l'élève

Les deux sujets plus importants détectés correspondent au sujet 1 Ingénierie de Logiciels et Informatique et au sujet 2 Mathématiques et Science de Données . À partir de ces résultats, le système a produit les recommandations suivantes en faisant une analyse de similarité selon la distribution de mots du document téléchargé par l'élève à l'intérieur de chaque sujet :

Topic 1. Mots-clés : logicielles, développement, interaction, informatiques, applications, innover, développeur, concept,...

Show 10 entries Search:

Niveau de Similarité	TAF/UE	Document
0.106	UE Développement d'applications sur dispositifs mobiles V. 1	docs/Développement_dapplications_sur_dispositifs_mobiles.pdf
0.093	TAF développement collaboratif et multisites de logiciels V. 1	docs/06_BN_DCL.txt V. 1
0.089	TAF ingénierie logicielle et innovation V. 1	docs/12_N_et_LOGIN.txt V. 1
0.084	TAF ingénierie logicielle des systèmes distribués V. 1	docs/11_B_et_isd.txt V. 1
0.064	TAF interaction hommemachine et systèmes collaboratifs V. 1	docs/09_B_IHM.txt V. 1
0.061	UE Prototypage Rapide : au-delà du design thinking V. 1	docs/Prototypage_Rapide__au-delà_du_design_thinking.pdf
0.037	UE Evaluation de l'acceptation des NTIC V. 1	docs/Evaluation_de_lacceptation_des_NTIC.pdf

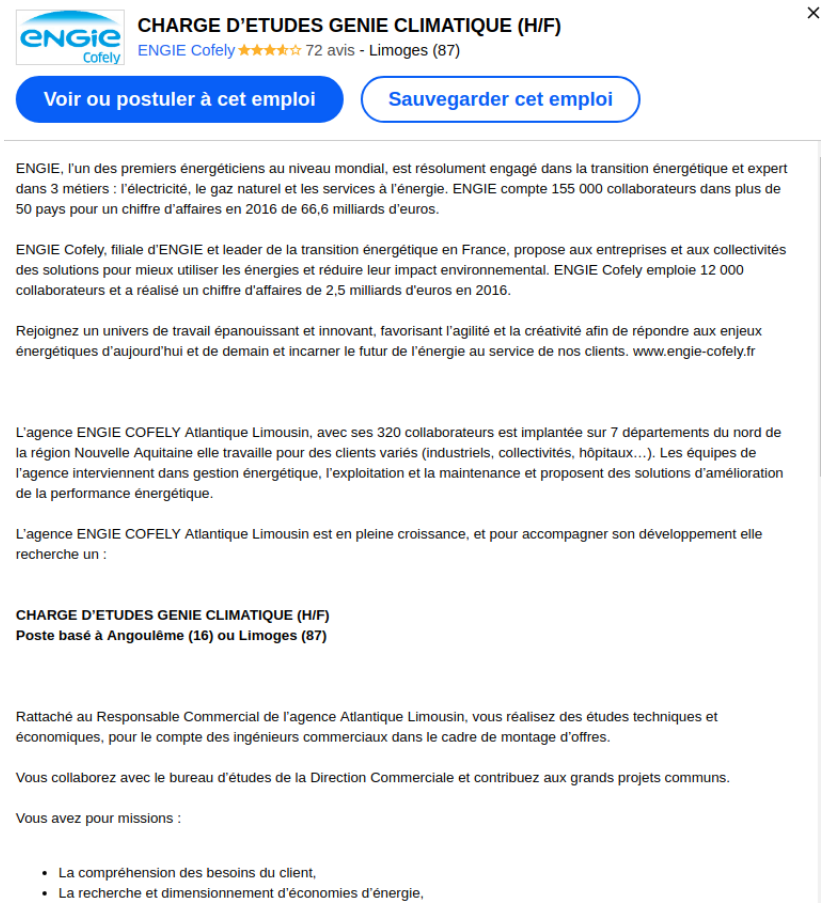
Topic 2. Mots-clés : apprentissage, données, problèmes, méthodes, connaissances, transmission, projets, traiter, mathémat...

Show 10 entries Search:

Niveau de Similarité	TAF/UE	Document
0.105	UE Big Data Analytics V. 1	docs/Big_Data_Analytics.pdf
0.077	UE Architecture Big Data et outils Hadoop V. 1	docs/Architecture_Big_Data_et_outils_Hadoop.pdf
0.075	TAF data science : des données au décideur V. 1	docs/05_B_DaSci.txt V. 1
0.063	UE Storytelling with data: Data Visualization and Data Wrangling V. 1	docs/Storytelling_with_data_Data_Visualization_and_Data_Wrangling.pdf
0.062	UE Fouille de données avancée V. 1	docs/Fouille_de_données_avancée.pdf
0.051	TAF ingénierie mathématique et computationnelle V. 1	docs/17_B_MCE.txt V. 1
0.051	UE Architecture informatique et implémentation numérique V. 1	docs/Architecture_informatique_et_implémentation_numérique.pdf
0.050	UE Machine Learning V. 1	docs/Machine_Learning.pdf
0.046	UE Advanced C++ Programming V. 1	docs/Advanced_C_Programming.pdf

Figure 12: Recommandations générées par la plate-forme

L'élève peut télécharger aussi des documents d'offres professionnelles. Cela lui permet d'obtenir de recommandations de TAF et d'UE qui sont liées à l'offre de stage d'un point de vu de la détection de sujets faite par les modèles déjà entraînés. Ci-dessous, un exemple des résultats d'une offre de stage liée aux énergies et l'environnement :



CHARGE D'ETUDES GENIE CLIMATIQUE (H/F)

ENGIE Cofely ★★★★☆ 72 avis - Limoges (87)

Voir ou postuler à cet emploi | Sauvegarder cet emploi

ENGIE, l'un des premiers énergéticiens au niveau mondial, est résolument engagé dans la transition énergétique et expert dans 3 métiers : l'électricité, le gaz naturel et les services à l'énergie. ENGIE compte 155 000 collaborateurs dans plus de 50 pays pour un chiffre d'affaires en 2016 de 66,6 milliards d'euros.

ENGIE Cofely, filiale d'ENGIE et leader de la transition énergétique en France, propose aux entreprises et aux collectivités des solutions pour mieux utiliser les énergies et réduire leur impact environnemental. ENGIE Cofely emploie 12 000 collaborateurs et a réalisé un chiffre d'affaires de 2,5 milliards d'euros en 2016.

Rejoignez un univers de travail épanouissant et innovant, favorisant l'agilité et la créativité afin de répondre aux enjeux énergétiques d'aujourd'hui et de demain et incarner le futur de l'énergie au service de nos clients. www.engie-cofely.fr

L'agence ENGIE COFELY Atlantique Limousin, avec ses 320 collaborateurs est implantée sur 7 départements du nord de la région Nouvelle Aquitaine elle travaille pour des clients variés (industriels, collectivités, hôpitaux…). Les équipes de l'agence interviennent dans gestion énergétique, l'exploitation et la maintenance et proposent des solutions d'amélioration de la performance énergétique.

L'agence ENGIE COFELY Atlantique Limousin est en pleine croissance, et pour accompagner son développement elle recherche un :

CHARGE D'ETUDES GENIE CLIMATIQUE (H/F)
Poste basé à Angoulême (16) ou Limoges (87)

Rattaché au Responsable Commercial de l'agence Atlantique Limousin, vous réalisez des études techniques et économiques, pour le compte des ingénieurs commerciaux dans le cadre de montage d'offres.

Vous collaborez avec le bureau d'études de la Direction Commerciale et contribuez aux grands projets communs.

Vous avez pour missions :

- La compréhension des besoins du client,
- La recherche et dimensionnement d'économies d'énergie,

Figure 13: Offre de Stage Chargé d'études Génie Climatique

Figure 14: Détection de sujets sur l'offre d'emploi. Les sujets d'Énergies/Environnement et Management sont les plus importants

7 Discussion

Cette plateforme a permis de rendre des documents informatifs à l'intérieur de l'école en sources pour la génération de modèles de classification automatique des documents dans le cadre de la nouvelle offre académique. Sur

l'ensemble de 24 TAF et de plus de 100 unités d'enseignement, le système recommande du contenu académique à partir du téléchargement des documents descriptifs du parcours de l'élève et de son avenir : un CV, une lettre de projet d'études ou un document correspondant à une offre professionnelle. Cela permet aux élèves de mieux se placer parmi la nouvelle offre académique et de les rendre plus conscients et plus acteurs de l'horizon de leur formation.

Le futur de développement de la plateforme est très ouvert. Dans le futur, le système de recommandation permettra de distribuer les sujets des modèles produits de manière indépendante. Une couche sera mise en ouvre au dessus des modèles entraînés. Cette couche permettra de distribuer les sujets de chaque modèle de manière indépendante afin que plusieurs modèles entraînés apportent des sujets divers concernant l'offre académique existante. Cette fonctionnalité d'exploration de sujets permettra aux élèves l'approfondissement dans l'exploration de leurs intérêts et leur avenir. Des diverses sujets : Médicale, Réseaux, Informatique, Traitement de l'Information et du Signal, Management etc. venant de différents modèles entraînés constitueront un moteur intégral d'inférence qui permettra aux élèves l'approfondissement dans la recherche de leur parcours professionnel et l'offre académique de l'école.

De plus, le système de recommandation sera entièrement dynamique. Le base de données lexique du système augmentera de manière dynamique à mesure que de nouveaux documents arrivent. La base de connaissances, intrinsèque à l'écosystème de modèles entraînés, sera élargie avec des nouveaux modèles et sujets qui permettront d'atteindre diverses fonctionnalités du système de recommandation. Cette intégration de modèles permettra d'offrir des fonctionnalités diverses selon les besoins d'un école. Par exemple, dans le cas de l'école IMT Atlantique, le système pourrait aider à la recommandation des combinaisons des TAF pour la deuxième et la troisième année des élèves grâce à l'estimation des paires de TAF (TAF deuxième année et TAF troisième année) plus similaires aux documents téléchargés par un élève.

References

[1] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. Technical report, 2000.

[2] David M Blei, Andrew Y Ng, and Jordan@cs Berkeley Edu. Latent Dirichlet Allocation Michael I. Jordan. Technical report, 2003.

[3] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On Smoothing and Inference for Topic Models. Technical report.

[4] Joeran Beel, Bela Gipp, Stefan Langer, Corinna Breitinger, and Corinna Breitinger breitinger. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17:305–338, 2016.

[5] Tf-idf: Information retrieval and text mining, 2008.

[6] Shuai Wang, Zhiyuan Chen, Geli Fei, Bing Liu, and Sherry Emery. Targeted Topic Modeling for Focused Analysis.

[7] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures.

[8] Tippaya Thinsungnoen, Nuntawut Kaoungku, Pongsakorn Durongdumronchai, Kittisak Kerdprasop, and Nittaya Kerdprasop. The Clustering Validity with Silhouette and Sum of Squared Errors. 2015.