Comptes Rendus Réunions

Projet Inter Promo - G5 Classification

Personnes :

**M2** : JJ (Chef), LD (Sous-Chef), AA (Qulité), DI (Sous-Chef 2), FB

**M1** : NA, NK, TV, VR, AG (abs)

**L3** : PM, GR, JC, AS, MV

Lundi 04 Janvier 2021

* **Réunion du matin (9h - 10h) :**

Sujets abordés lors de la réunion :

* Présentation du Brown Paper pour rappeler ce que l’on va faire pendant les 2 semaines du Projet Inter Promo.
* Présentation de Google collab grâce à un notebook permettent d'importer et de lire les données.
* Présentation de la charte qualité à l’aide d’un notebook expliquant les notions principales.
* Planification des premiers jours de travail.

Travail à effectuer :

* Réalisation du travail préparatoire.
* Familiarisation avec les données (compréhension des fichiers tf-idf et bow) et analyse du fichier data (valeurs manquantes, etc).
* Création de features basés sur la syntaxe (nombres de phrases).
* **Réunion début d’après-midi (13h - 13h20) :**

Sujets abordés lors de la réunion :

* Répondre aux questions sur le travail préparatoire.
* Rappel du travail à faire pour l'après-midi.
* **Réunion fin d’après-midi (16h30 - 16h45) :**

Sujets abordés lors de la réunion :

* Faire le point sur le travail effectué dans la journée dans tous les groupes.
* Répondre aux questions s’il y en a.
* Dire le travail du lendemain.
* **Réunion des Chefs (17h - ...)**

.

✓ L’intégralité du travail prévu ce jour a été effectué.

Mardi 05 Janvier 2021

* **Réunion du matin (9h - 9h30) :**

Sujets abordés lors de la réunion :

* Re-Présentation des 3 sous-groupes et présentation des sous-chefs.
* Planification du travail à faire aujourd’hui.

Travail à effectuer :

1. Clustering
   * Labellisation manuelle des articles. On détermine les thèmes des articles à partir des tags.
   * Début de l’implémentation d’algorithmes.
2. Innovation
   * Labélisation manuelle d’articles pour créer un petit échantillon d’entraînement avec certains articles qui parlent avec certitude d’innovation et d’autres qui ne parlent pas du tout d’innovation.
3. Doublon
   * Tester des méthodes de non-supervisé avec Scikit-Learn (Isolation Forest, Local Outlier Factor (LOF), One-Class SVM et Robust covariance) en utilisant les features d’hier ainsi que le TF-IDF et Bag-of-Words (BoW) réalisé par les non-alternants.
   * Quelques représentations visuelles pour analyser les premiers résultats des modèles.

* **Réunion avec Mathieu Serrurier (10h30 - 11h30) :**

Sujets abordés lors de la réunion :

* + Clustering : 2 approches possibles :

1. Labels : Méthode choisie initialement. Association des thèmes à une liste d’articles. Pour le choix des thèmes on utilise les tags déjà présents. Puis utilisation de ces données labellisées pour le calcul d’une erreur. Avec pour algorithmes les algorithmes de clusterings.
2. Clustering : Utilisation d’algorithmes de clusterings également (mais différents des premiers) comme la CAH. Puis une fois les clusters acquis choix des thèmes associées à chaque clusters (encore une fois grâce aux tags).

* Innovation :
* Labellisation manuelle de certains articles. Pour obtenir certains articles dont ont est sûr que l’article évoque une innovation ou qu’il n’en évoque.
* Utilisation d’algorithmes de machine learning supervisés/semi-supervisés.
* Doublon :
* Approche avec le non-supervisé (Pauwels).
* Représentation visuelle ? (Nuage de Points, …).
* Définir ce qu'est une nouveauté ? Dans le temps ?
* Utiliser LOF : qu’avec les documents datés avant lui.
* BoW pourrait être une baseline.
* Définir clairement les définitions pour “Innovation” et “Doublon” !
* Faire de la représentation en 2D (ACP, t-SNE, ...) et nuage de mots (wordcloud) pour un texte et pour un ensemble de texte. (API toute simple). ➨ Pour représenter les résultats !
* Créer un sous groupe représentation qui va créer une représentation vectorielle. Faire nos propres représentations vectorielles en réalisant nos représentations à partir de ceux du groupe 3 en réalisant un auto-encoder (keras, pytorch).
* **Réunion début d’après-midi (13h30 - 13h45) :**

Sujets abordés lors de la réunion :

* Résumé vite fait de la réunion de ce matin avec M. Serrurier.
* Dire le travail à faire pour l'après-midi, continuer le travail du matin !
* Travail pour des L3, représenter les résultats visuellement (2D + nuage de mots).
* **Réunion fin d’après-midi (16h45 - 17h00) :**

Sujets abordés lors de la réunion :

* Faire le point sur le travail effectué dans la journée dans tous les groupes.
* Répondre aux questions s’il y en a.
* Dire le travail du lendemain.
* **Réunion des Chefs (17h - ...)**

.

✓ Nous avons commencé à travailler par sous-groupe.

Mercredi 06 Janvier 2021 Présentiel - Salle U3-108

* **Réunion du matin (9h - 9h10) :**

Sujets abordés lors de la réunion :

* Présentation du travail à faire pour les 3 groupes.

Travail à effectuer :

* Clustering :

- Labellisation (manuelle et k-Mean).

* Innovation :

- Labelliser les documents.

- Mise en place/finalisation de la v0.

- Création d’algos de preprocessing (lemmatisation).

* Doublon :

- Comparer les résultats et tests d’algorithmes.

- Optimisation des hyper-paramètres des algos.

* **Réunion avec Mathieu Serrurier (9h45 - 10h) :**

Sujets abordés lors de la réunion :

* + Clustering :

- Faire du clustering sur les tags (pour aider à classer les textes), (pour donner des catégories). Pour essayer de prédire un texte qui n'a pas de tag.

* Innovation :

- Attention aux biais ! (Fortement présent pour la V0).

- Pour la suite (et surtout la V1) attention de taguer plus finement les données.

* Doublon :

- (Aucune remarque)

* **Réunion Jeremy & Damien (10h30 - 11h)**
* **Réunion des Chefs (17h - ...)**

✓ Nous avançons tranquillement dans le projet.

Jeudi 07 Janvier 2021

* **Réunion du matin (9h - 9h10) :**

Sujets abordés lors de la réunion :

* Présentation du travail fait par les autres groupes.
* Présentation du travail à faire pour les 3 groupes.

Travail à effectuer :

* Clustering :

- Labellisation

* Innovation :

- Tester plusieurs modèles.

* Doublon :

- Tester le nouveau BoW fourni par le groupe 3.

* Finir la v0 pour les 3 sous-groupes.
* Faire un petit rapport pour expliquer le fonctionnement des v0.
* **Réunion Sous-Chefs (9H30 - 10h)**

Mise au point sur la situation actuel de chaque sous groupe :

* Clustering :

- Fin de la labellisation manuelle (Regroupement par tag).

- Fin de la labellisation automatique (K-Mean puis analyse des clusters).

- Comparaison des deux labellisations .

- Calcul d’un score de pertinence de chaque cluster (similarité entre le thème et le titre du document).

- Préparation du code de la V0.

* Innovation :

- Mise en place de labelling manuel (On classifie comme pertinent les documents avec beaucoup de mots clés liés à l’innovation et comme non innovant les documents qui n’ont aucun mot clés liés à l’innovation).

- Suivie d’une étape de pseudo labelling (On entraîne un modèle sur les données avec labélisation manuelle, puis on prédit et si jamais un article est prédit dans une classe à plus de 99% on l’associe dans cette classe.)

- Entraînement de façon supervisée avec le labelling obtenus après le pseudo labelling.

- On remarque la présence d’un biais dans les labels car on obtient une prédiction de 100%. Donc pour la suite essayer de retirer le biais.

* Doublon :

- Mise en place de l'Isolation Forest et d’un LOF.

- Entraînement et prédiction sur 3 jeux de données différents (TF-IDF, BOW et features syntaxique).

- Analyse des résultats de façon qualitative (Avec des graphiques), avec des plots 3D d’une ACP réalisées sur les données BOW. On réalise une analyse qualitative des résultats et non quantitative car on ne possède pas de labels et donc il nous est impossible de calculer des métriques.

* **Réunion début d’après-midi (13h30 - 13h40) :**

Sujets abordés lors de la réunion :

* Rappel du travail à faire.
* **Réunion Sous-Chefs (16h40 - 16h50)**
* **Réunion fin d’après-midi (16h50 - 17h00) :**

Sujets abordés lors de la réunion :

* Faire le point sur le travail effectué dans la journée dans tous les groupes.
* Répondre aux questions s’il y en a.
* Dire le travail du lendemain.
* **Réunion des Chefs (17h - ...)**

✓ Dans chaque sous-groupe nous avons fini la v0.

Vendredi 08 Janvier 2021 Présentiel - Salle U3-03

* **Réunion avec Berger-Levrault & les Chefs (9h - 10h30)**
* Nouveau document (Doublon) : Tout se passe bien et aucune remarque particulière du client
* Document Innovant (Innovation) : Le client aimerait bien que la même chose soit appliquée pour le thème “Gamme de Gestion”. Pour ce thème on dispose d’une liste de mots clés (comme pour le thème “Innovation”).
* Clustering : A ce jour nous disposons d’une méthode ou le nombre de clusters est déterminé de façon manuelle, et le titre de chaque cluster est déterminé de façon manuelle également. Le client souhaiterait lui une version adaptative, concrètement cela se traduit par un algorithme qui va trouver de façon automatique le nombre de clusters et qui va associer un nom (automatiquement) à chaque clusters.
* **Réunion des Chefs (10h30 - 10h45)**
* **Point Chef, Sous-Chefs et Qualité (10h45 ~ 11h15)**

Conclusion :

On casse le sous groupe Nouveau et on réattribue les personnes dans ce groupe dans les groupes Innovation (Valentine et Maylin) et Clustering (Théo et Jérémy), et Alice est à la qualité.

* **Point du Sous-Groupe Clustering (13h30 - 14h) :**

Sujets abordés lors de la réunion :

* Appliquer les remarques de la réunion avec Berger-Levrault.
* **Point avec Chefs & Profs (14h00 - 14h40)**

Nos codes doivent être lancés en même temps que les codes du G6 ()

* **Réunion avec Mathieu Serrurier (14h45 - 15h30) :**

Sujets abordés lors de la réunion :

* + Serveur : Driver Invidia, install GPU, …, man sceen !

@Jeremy Johann Tu pourras compléter le compte rendu de réunion ! Oui Madame !

Étude des résultats du point avec le client.

* + Clustering :

- Finir la V0 même si on a un peu de mal.

- Le problème de créer un algo de clustering avec une faculté d’adaptabilité a été soulevé. Les problèmes de cette méthode sont :

- la détection automatique du nombre de clusters.

- La détection automatique de l’étiquette de chaque cluster.

- Solution :

- Pour le nombre de clusters : Certains algorithmes permettent l’obtention automatique du nombre de clusters.

- Pour trouver l’étiquette des clusters : On utilise les tags les plus proches du centroïdes. Ou alors on prend le mot qui revient le plus dans le titre de chaque article.

- Proposition de l’algorithme de Clustering Spectral.

- Mise en place de métrique de clustering (Calinsky : recommandé par Damien et M.Serrurier)

- Et trouver un nom automatique (LDA).

* Innovation :

- Continuer le semi-supervisé

- Comme on est en SID et que par conséquent on a fait du code propre le code pour la liste Innovation est facilement applicable à la liste “Gamme de Gestion”.

- Par la suite : On peut tenter des méthodes “exotiques” en entraînant des modèles en non supervisé (Nuance entre Anomaly Detection et Novelty Detection)

* Doublon :

- Aucune remarque, continuer à faire tourner notre code lorsqu’on reçoit une nouvelle représentation du groupe 3.

* Présentation :

Tips pour le rendu final :

1. Bien poser le problème (Exemple du problème du nombre de cluster et des étiquettes de clusters)
2. Présenter les métriques (Exemple le Fscore incompréhensible)
3. Présenter les résultats quantitatifs et les expliquer avec du contexte (Exemple la détection des spam 0.9/0.99 contre la prédiction de la bourse 0.50/0.53)
4. Présenter des résultats qualitatifs à l’aide de graphiques (Pour avoir un résultat plus compréhensible pour les non connaisseurs).

Important de mettre la différence entre les graphes de la V0 et la V1 pour dire qu’on a bien progressé.

Conseil sur des nouvelles représentation vectorielles :

* Récupérer le vecteur laten d’un VAE (Variational Auto Encoder)
* La dernière couche cachée d’un BERT

Lundi 11 Janvier 2021

* **Réunion du matin (9h - 9h10) :**

Sujets abordés lors de la réunion :

Gros Point de début de semaine (et rappel de la réunion avec le client)

* Nouveau Document:
  + Rien de nouveau. On attend de meilleures représentations vectorielles pour fournir des résultats plus représentatifs de la réalité (car document embedding meilleur que word embedding).
  + Rééquilibrage des groupes avec les effectifs du groupe nouveau qui sont divisé dans les groupes innovation et clustering.
* Document Innovant :
  + Préparer les Codes et la liste de lexique pour “Gamme de Gestion”
  + Améliorer la phase de semi-supervisé (Gros problème de biais il faut mieux détecter les document innovant initiaux).
  + Analyser les premiers résultats obtenus avec le semi-supervisé (Cloud Word).
* Clustering :
  + Finir la V0 (Static)
  + Commencer la V1 (Adaptable), pour ça on va utiliser des métriques qui vont nous fournir automatiquement le nombre de clusters et des méthodes qui vont nous fournir automatiquement le nom de chaque cluster.