Projet de traitement de données massives

Pour ce projet, vous aurez accès aux données de CN2i, une compagnie gérant les données de six journaux au Québec. Votre projet consistera à résoudre une problématique réelle et actuelle de la compagnie : comment prédire la popularité d'un article auprès du public avant sa publication afin de faire de l'aide à la rédaction pour les articles plus faibles.

Les données

Le dossier *articles* contient un fichier par article. Le fichier <guid>.json contient le contenu et les médatonnées de l'article en format structuré, avec les champs suivants :

type: le type de contenu
templateName: l'affichage du contenu, dépendant du type de contenu
visual: le visuel d'en-tête de l'article, qui peut être une image, une galerie d'images, un vidéo, etc.
availableInPreview: si le contenu est disponible gratuitement ou non. Tous le contenu est gratuit.
url: l'adresse unique du JSON
creationDate: date de création de l'article
modificationDate: date de dernière modification de l'article
externalIds: URL de l'article s'il provient d'une source externe
authors: auteur de l'article
id: le <guid> de l'article
title: le titre de l'article
channel: un mot clé décrivant l'article
lead: le résumé de l'article qui apparaît au début
chapters: le texte de l'article

Les fichiers <guid>--publication-info.json contiennent les informations de publication. Il contient un document JSON par emplacement où l'article a été publié. On notera surtout la clé organizationKey qui indique le journal auquel cette publication est rattachée, la clé publications.publicationDate qui indique quand est-ce qu'elle y a été publié, la clé publications.type qui indique le type de publication, et la clé publications.slug qui contient un URL pour l'article dont la première partie est la catégorie de l'article en question. Pour simplifier, vous pouvez ignorer les publications de type « dossiers » (qui ne sont pas des nouveaux articles mais des collections d'articles déjà publiés sur un même sujet).

Les fichiers analytics enregistrent les évènements à chacun des six journaux de CN2i. Il y a un fichier contenant tous les évènements de la journée pour cinq des six journaux, et plusieurs fichiers pour Le Soleil. Chaque évènement est un document json de quatre clés :

"createdAt": "2019-01-02T01:42:01.519Z",
"hash": "556f1db1873df9403fbd0261f85cceaa",
"name": "View30",
"product": "latribune/web"

La clé *createdAt* est simplement le moment de l'évènement, la clé *hash* est le <guid> de l'article, et la clé *product* est le journal et la plateforme (site web ou application mobile) de laquelle l'évènement provient. Le nom de l'évènement a six valeurs possibles, soit *view, view5, view10, view30,* ou *view60*,

et représente un click, un temps de lecture de 5 secondes, 10 secondes, 30 secondes, ou 60 secondes, respectivement. Notez qu'un temps de lecture plus élevé implique que les évènements précédents ont été créés : une personne qui passe 15 secondes à lire un article aura trois documents json dans le fichier analytique correspondant aux évènements *view*, *view5* et *view10*.

Vous avez des articles depuis 2013 et les données analytiques de janvier à juillet 2019. Il y a environ 120Go de données. Vous pouvez l'archive compressée (8Go) ici : http://fsg-p-ftp01.fsg.ulaval.ca/RichardKhoury.zip. Notez cependant qu'un des fichiers dans cette archive est corrompu. Le fichier 2019-01-24T13_00_00.000Z doit être téléchargé du site web du cours et remplacer celui dans l'archive.

La problématique

La popularité d'un article est une combinaison de deux aspects, soit combien de personnes le voient, et combien de temps elles passent à le lire. Il est facile de maximiser l'une de ces métriques au détriment de l'autre. Par exemple, un article « click-bait » attire un maximum d'utilisateurs, mais pour un minimum de temps. À l'inverse, un article présentant une analyse technique longue et approfondie d'un sujet maximisera le temps de lecture des utilisateurs le consultant, mais n'intéressera qu'un minimum du public. Le défi est plutôt d'optimiser ces deux métriques simultanément. CN2i réalise ceci simplement avec une somme pondérée des six évènements.

Exemple : Supposons un article avec les évènements suivants :

évènement	compte	pondération
view	500	1
view5	450	1
view10	300	2
view30	120	5
view60	10	10

La valeur pondérée de cet article est : $(10 \times 10) + (120 \times 5) + (300 \times 2) + (450 \times 1) + (500 \times 1) = 2\,250$.

L'objectif spécifique du problème est de créer un outil d'aide à la rédaction pour les articles plus faibles. Lorsqu'un journaliste entre un nouvel article dans cet outil, sa valeur est prédite. Les 10% d'articles plus faibles (arrondissez au besoin pour ne pas avoir de fractions d'articles) sont identifiés et des suggestions de modifications sont données aux auteurs pour améliorer le tout.

Cependant, étant donné que la valeur de l'article est directement liée au volume de lecture, il est important de différentier les sujets plus monotones (ex. : l'économie) des sujets « sexy » attirant l'attention (ex. : les potins de célébrités). Ne tombez pas dans le piège de simplement laisser passer ces articles bonbons au détriment d'articles plus sérieux! Vous devez noter chaque article relatif aux autres articles dans sa catégorie, c'est-à-dire sa section dans le journal (la valeur de la clé slug).

L'évaluation

En plus de la base de données, vous aurez une liste d'identifiants de nouveaux articles tests (pour le mois d'août 2019). Vous devez prédire la valeur de ces articles et soumettre ces prédictions avec votre rapport final. Votre système sera évalué sur votre habileté à prédire quels sont les 10% pires articles dans chaque catégorie. Nous allons donc sélectionner les 10% articles auxquels vous avez donné la valeur la plus faible dans chacune des catégories (avec une sélection aléatoire en cas d'égalité) et les comparer aux vrais 10% articles les plus faibles. La métrique d'évaluation est simplement le ratio d'intersection entre ces deux listes :

$$ratio = \frac{(vos\ articles) \cap (vrais\ articles)}{(vos\ articles) \cup (vrais\ articles)}$$

Votre système n'a pas besoin d'afficher des suggestions d'amélioration des articles faibles. Cependant, les attributs utilisés pour prédire la valeur des articles doivent être présentés dans vos rapports.

L'aide à la rédaction

L'équipe produisant le meilleur projet sera considérée pour le prix Pierre Ardouin (voir la section appropriée). Afin de se qualifier pour ce prix, les étudiants doivent produire un travail innovateur, écrire un rapport complet de haute qualité, et avoir une très bonne performance dans la métrique d'évaluation. Une équipe réalisant un vrai système d'aide à la rédaction (avec une interface graphique et des suggestions utiles) sera naturellement avantagée pour ce prix.

La compagnie CN2i est très intéressée aux résultats de votre étude, et consultera les rapports de projets. Elle se réserve le droit de faire une offre pour acheter un système intéressant ou pour engager les étudiants l'ayant créé.

Premier rapport : Analyse et prétraitement des données (10%)

L'objectif de la première partie du projet est de vous familiariser avec les données avec lesquelles vous allez travailler.

Analysez vos données et de leurs propriétés statistiques. Portez attention autant aux valeurs normales qu'aux cas problématiques, comme la présence de bruit, le fléau de dimensionnalité, les informations manquantes, le déséquilibre des classes, les valeurs aberrantes, etc. Discutez de vos observations. L'objectif ici n'est pas de faire une grande liste de statistiques sur les données, mais d'en tirer des leçons pour guider la réalisation du projet. (3 points)

Prévoyez les attributs que vous allez utiliser pour votre algorithme de traitement de données. Il y a une immense variété d'attributs qui peuvent être obtenus de ces données, incluant des attributs linguistiques (les mots utilisés dans l'article, les adverbes, etc.), des attributs numériques (la longueur de l'article en mots, en paragraphes, le nombre de photos, le niveau de difficulté du texte, etc.), des attributs stylistiques (le contenu du titre, le contenu des photos, l'utilisation de citations, etc.), des attributs sociaux (l'inclusion de tweets, de vidéos YouTube, etc.), et des méta-attributs (le sujet de l'article, l'heure de publication, etc.). Vous devez prévoir les premiers attributs sur lesquels vous allez

vous concentrer (il vous est bien entendu possible d'en rajouter à n'importe quel moment au cours de la session). Justifiez votre choix d'attributs initiaux. (3 points)

Prévoyez comment vous allez traiter les données d'un point de vue pratique. C'est-à-dire premièrement les algorithmes que vous allez implémenter, mais aussi l'optimisation de ceux-ci afin de pouvoir traiter la quantité massive de données disponible pour ce projet de manière efficace. (2 points).

Discutez également de la procédure de tests que vous envisagez. Vous ne pouvez pas tester votre système avec la liste d'identifiants d'évaluation, car vous n'avez pas les résultats réels pour ces articles. Vous devez donc prévoir votre propre procédure de tests afin de savoir si chaque variation que vous implémentez pour votre solution améliore ou non vos prédictions, et ainsi guider votre travail de développement. (2 point)

Le rapport pour cette partie devrait être d'environ 10 pages et est dû le 24 février 2021.

Deuxième rapport : traitement des données (10%)

Pour ce rapport, vous devez présenter les algorithmes de traitement de données que vous avez implémenté, leur fonctionnement et les résultats que vous avez obtenu. Je m'attends que vous ayez un processus de développement itératif : implémentez un système, testez-le, découvrez ses points faibles, et raffinez-le en conséquences (en ajoutant des attributs, en modifiant l'entraînement, en corrigeant l'algorithme, etc.).

Décrivez les algorithmes que vous avez choisi d'implanter. Décrivez, d'un point de vue technique, comment ils fonctionnent et les composantes clefs. Le but ici n'est pas de répéter les notions de base des algorithmes que je vous ai enseigné au cours de la session, mais plutôt d'expliquer comment vous avez adaptés et utilisés ces algorithmes pour ce projet. Justifiez vos choix pour les décisions de design et d'implémentation que vous avez pris. Décrivez leur efficacité étant donné le volume de données à traiter. (2 points)

Décrivez également les tests que vous avez faits. Pour chaque test, présentez les résultats attendus et les résultats obtenus. Présentez des statistiques pertinentes (taux de succès, précision, rappel, temps moyen de calcul, complexité algorithmique, etc.). Discutez des leçons que vous avez prises de chaque test et comment elles ont guidé votre travail. (2 points)

Présentez la version finale de votre système. Quels attributs des données sont utilisés par votre algorithme, quels attributs ont une valeur prédictive plus importante, et pourquoi? (2 points)

Présentez des études de cas (i.e. des articles) comme exemples spécifiques du fonctionnement de votre algorithme. Décrivez autant les cas qui fonctionnent bien que ceux pour lesquels le test échoue, et discutez des raisons pour cette différence. (2 point)

Présentez la capacité de votre système à offrir une aide à la rédaction. Sur quels attributs et valeurs est-ce que les articles plus faibles sont identifiés? Quels conseils peut-on donner aux auteurs de ces articles pour les améliorer? (1 point)

Offrez une rétrospective sur le projet. Comparé à vos réflexions au début de la session, en quoi avez-vous eu raison, et quelles surprises avez-vous eu en chemin? Si le projet était à refaire, que feriez-vous différemment? (1 point)

Le rapport pour cette partie devrait être d'environ 10 pages et est dû le 22 avril 2020.

Évaluation des résultats (5%)

Soumettez votre prédiction sur les données de test. Cette prédiction vaudra 5 points. Il sera calculé comme ceci :

Ratio	Points
[0.90, 1.00]	5
[0.70, 0.90[	4
[0.50, 0.70[	3
[0.30, 0.50[	2
[0.20, 0.30[	1
[0.00, 0.20[	0

Évaluation des rapports

Les rapports seront remis en-ligne à travers le site web du cours. Une seule soumission par équipe. Chaque rapport doit inclure une page titre indiquant les membres de l'équipe, leur niveau (1er, 2e, ou 3e cycle), et la date de soumission. Les rapports doivent être écrits en Word ou LaTeX (pas de rapports écrits à la main) et soumis en format PDF.

La majorité des points du rapport seront donnés sur l'analyse et la discussion de votre système et de vos résultats. Il est donc important (pour vous) d'écrire une analyse approfondie et scientifique. La question centrale n'est donc pas « qu'est-ce qui se produit », mais « pourquoi est-ce que ça se produit » et « qu'est-ce qu'on peut y faire ». Vous ne devez pas simplement écrire un algorithme et générer des résultats. Vous devez être en mesure de justifier vos décisions qui ont mené à votre algorithme, et expliquer pourquoi il a généré ces résultats.

Un exemple peut clarifier les choses. Supposons que vos tests démontrent que les articles à propos du film *Sharknado* sont les mieux côtés de manière démesurée. Vous pouvez rapporter ce résultat de plusieurs manières :

- « Notre algorithme côte les articles à propos du film Sharknado trop haut. » Ceci n'est pas une analyse, mais simplement une observation des faits. Les points donnés seront minimaux.

- « Notre algorithme côte les articles à propos du film Sharknado trop haut parce que ce film obtient un score élevé dans notre algorithme trop souvent. » Ceci est l'inverse d'une analyse utile. Je ne donne pas de points, et je me réserve le droit de rire de vous.
- « Notre algorithme côte les articles à propos du film Sharknado trop haut parce que la description du film contient une immense liste de tous les mots clefs possibles, donc il se trouve à obtenir un score élevé par la somme des mots. » Vous avez identifié et analysé le problème et découvert sa source, bien joué! Vous avez des points.
- « Notre algorithme côte les articles à propos du film Sharknado trop haut parce que la description du film contient une immense liste de tous les mots clefs possibles, donc il se trouve à obtenir un score élevé par la somme des mots. Nous allons résoudre ce problème en assignant un poids aux mots clefs en fonction de la longueur du texte. » Non seulement vous avez découvert la source du problème, mais vous l'avez comprise assez bien pour proposer une solution, c'est fantastique. Vous aurez une bonne note.
- « Notre algorithme côte les articles à propos du film Sharknado trop haut parce que la description du film contient une immense liste de tous les mots clefs possibles, donc il se trouve à obtenir un score élevé par la somme des mots. Une solution possible serait de tronquer le texte, mais nous jugeons que ceci nous ferait perdre trop d'information utile étant donné que la vaste majorité des articles ont des contenus précis et informatifs. Une autre solution possible serait de noter la valeur des textes des articles selon le nombre de lectures. Cependant, nous anticipons plusieurs problèmes avec cette solution, par exemple que faire d'un nouvel article qui n'a pas encore été lu, et comment éviter les articles viraux? Finalement, on pourrait assigner un poids aux mots clefs en fonction de la longueur du texte, ce qui pénaliserait les articles trop longs sans affecter les autres. C'est la solution que nous avons choisie d'appliquer. » Vous avez identifié le problème, vous l'avez analysé pour trouver sa source, puis vous avez exploré plusieurs pistes de solutions et justifié votre choix d'une en particulier. C'est parfait. 100%.

Notez finalement que jusqu'à 10% des points d'un rapport peuvent être enlevés en pénalité pour une mauvaise qualité. Ceci inclut particulièrement les fautes d'orthographe et de grammaire, les figures mal préparées (ou dessinées à la main), les rapports écrits à la main, les irrégularités de polices et tailles de caractère, et les textes incohérents.

Équipes

Le projet doit être réalisé en équipes de 2 ou 3 étudiants. La note sera donnée pour l'équipe, et non par individu. Choisissez bien vos coéquipiers.

Plagiat

Le plagiat est une offense académique sérieuse. Tout étudiant qui tente de soumettre un travail qui n'est pas le sien sera pénalisé. Ceci inclut de copier le travail ou rapport d'un autre étudiant du cours ou un système trouvé ailleurs. Un étudiant coupable de plagiat recevra automatiquement la note de zéro pour le projet entier (c'est-à-dire toutes les parties) et s'exposera à d'autres sanctions telles que décidées par l'Université.

Conseils

- Les deux rapports produits par l'équipe ayant gagné le Prix Pierre-Ardouin l'an dernier est disponible sur le site web. Je vous conseille de le lire avant de commencer pour avoir un survol du travail déjà réalisé et des idées de solutions fonctionnelles. Pour référence, ils ont obtenu 26% de bonnes prédictions.
- Essayez plusieurs de vos idées et parlez-en dans vos rapports. Décrivez quelle est l'idée, pourquoi vous pensez que c'est intéressant à essayer (qu'est-ce que vous voulez découvrir ou que pensez-vous va arriver), et quel est le résultat obtenu (est-ce celui que vous attendiez, et sinon pourquoi). À force de réfléchir et d'expérimenter, vous trouverez une bonne solution. Et ce n'est pas mauvais que plusieurs de vos idées ne fonctionnent pas; c'est la nature même de la recherche! De plus, ça justifie expérimentalement que la version finale de votre système est la meilleure, et non simplement la première que vous avez essayé. Pour l'évaluation, je donne des points pour les explorations intéressantes (à condition qu'elles soient bien présentées, justifiées, et analysées, bien entendu). Je ne donnerai pas de points pour des idées farfelues ou mal présentées. Mais par contre je n'enlèverai jamais de points pour avoir essayé quelque chose. Et en contrepartie, si vous ne décrivez pas vos idées et expériences dans votre rapport, je ne peux pas vous donner de points du tout.
- Considérez les extrêmes logiques de vos idées. Par exemple, si augmenter le poids d'un attribut améliore les résultats, pourquoi ne pas l'augmenter encore plus, ou ne conserver que cette variable? Ce sera rarement le bon choix, mais d'explorer le comportement de votre système dans les cas extrêmes peut souvent aider à mieux comprendre le problème et à développer une nouvelle intuition pour sa solution.
- Justifiez votre analyse avec des démonstrations mathématiques lorsque possible.
- Ne soumettez pas un copier-coller de votre code au complet dans votre rapport. Expliquez comment votre algorithme fonctionne en utilisant des descriptions du processus et des étapes, la logique du système, des formules mathématiques, et du pseudo-code.
- Je suis disponible durant mes heures de bureau pour vous aider en discutant de votre projet, des difficultés que vous rencontrez, et en suggérant des idées et des pistes. Je n'ai pas de solutions pour chaque projet. Et je ne vais pas déboguer votre code pour vous.

Prix Pierre Ardouin

Depuis l'automne 2013, le Département d'informatique et de génie logiciel a mis en place un concours récompensant l'équipe qui aura produit le meilleur TP/projet dans le cadre d'un cours. Ces travaux de session ont l'envergure d'un mini-projet qui est admissible par rapport aux normes fixées par le Département. À la suite des évaluations des travaux, l'enseignant du cours détermine l'équipe gagnante; chaque membre de l'équipe gagnante reçoit alors une bourse de 50$ ainsi qu'une attestation remises par le Département.

De plus, le Département d'informatique et de génie logiciel a mis en place une bourse Élite, appelée bourse « Pierre Ardouin », qui vise à récompenser le meilleur projet de session, tous cours confondus. Deux principaux critères guident le choix des évaluateurs dans l'identification du lauréat :

l'excellence du travail (par rapport à ce qui est demandé dans l'énoncé) et l'aspect créativité/innovation. Il est actuellement prévu une bourse de 200$ pour récompenser chaque membre de l'équipe « élite » gagnante (pour un maximum de 1000$ pour toute l'équipe). Aussi, le Département veille à publier l'information sur un site Web dédié : http://www.ift.ulaval.ca/vie-etudiante/prix-pierre-ardouin.

À la deuxième moitié du mois de mai de chaque année universitaire, le Département organise une cérémonie pour honorer les finalistes et le lauréat du prix «Pierre Ardouin» des sessions d'automne et d'hiver, et leur remettre une attestation.