Séance : Text mining

Syllabus

Visseho Adjiwanou, PhD.

21 June 2024

Objectifs

Une grande partie des informations sur notre monde est stockée non pas dans des ensembles de données bien rangés, mais dans des textes. Les manifestes des partis, les publications sur les réseaux sociaux, les procès-verbaux des conseils municipaux, la correspondance par e-mail, les traités, les ordonnances, les décisions judiciaires, les débats présidentiels, les discours au Congrès... la pratique de la politique sous ses nombreuses formes finit par être enregistrée en mots sur une page. Comment prenons-nous tout ce texte non structuré, bien trop vaste pour qu'un seul chercheur puisse le lire, et le convertissons-nous en informations utiles pour la recherche scientifique ?

Dans ce cours, nous explorerons l'avant-garde du « texte en tant que donnée », en nous concentrant sur la manière de récupérer, représenter et analyser les textes dans le cadre d'un projet de recherche. Nous viserons à couvrir à la fois le conceptuel et le pratique, en passant du temps en classe à écrire du code pour reproduire et étendre des résultats significatifs de la recherche sur l'utilisation des données textuelles.

Contenu

Séance 1 : Introduction au Text Mining

- Objectifs :
  - Travailler avec les chaînes de caractères
  - Comprendre les concepts de base du text mining.
  - Explorer les applications du text mining dans différents domaines.
- Contenu :
  - Introduction et définition du text mining.
  - Applications du text mining (analyse des sentiments, extraction d'informations, résumé automatique, etc.).

  - Aperçu des outils et logiciels couramment utilisés (Python, R, NLTK, spaCy).
- Activité pratique :
  - Installation des outils de base pour le text mining dans R.

Séance 2 : Prétraitement des Données Textuelles

- Objectifs :
  - Apprendre les techniques de prétraitement des données textuelles.
  - Nettoyer et préparer les textes pour l'analyse.
- Contenu :
  - Tokenization.
  - Suppression des stop words.
  - Stemming et lemmatization.
  - Normalisation des textes.
- Activité pratique :
  - Mise en pratique des techniques de prétraitement avec R.

Séance 3 : Extraction de Caractéristiques Textuelles

- Objectifs :
  - Extraire des caractéristiques textuelles pour l'analyse.
  - Comprendre les représentations vectorielles des textes.
- Contenu :
  - Bag of Words (BoW).
  - TF-IDF.
  - Word embeddings (Word2Vec, GloVe).
  - Analyse de la fréquence des mots et des n-grams.
- Activité pratique :
  - Extraction de caractéristiques textuelles à partir de textes bruts.

Séance 4 : Analyse des Sentiments

- Objectifs :
  - Comprendre les techniques d'analyse des sentiments.
  - Mettre en œuvre un modèle d'analyse des sentiments.
- Contenu :

  - Concepts et applications de l'analyse des sentiments.
  - Méthodes d'analyse des sentiments supervisées et non supervisées.
  - Utilisation des lexiques et des algorithmes de machine learning.

- Activité pratique :

  - Construction et évaluation d'un modèle d'analyse des sentiments avec un jeu de données réel.

Séance 5 : Modèles de Classification Textuelle

- Objectifs :

  - Apprendre à classifier des textes en utilisant des techniques de machine learning.
  - Construire et évaluer des modèles de classification textuelle.

- Contenu :

  - Concepts de base de la classification textuelle.
  - Algorithmes couramment utilisés (Naive Bayes, SVM, Réseaux de neurones).
  - Évaluation des performances des modèles (précision, rappel, F1-score).

- Activité pratique :

  - Implémentation et évaluation d'un modèle de classification textuelle.

Séance 6 : Text Mining Avancé et Études de Cas

- Objectifs :

  - Explorer des techniques avancées de text mining.
  - Appliquer les connaissances acquises à des études de cas réelles.

- Contenu :

  - Structural topic modelling
  - Études de cas réelles : analyse de réseaux sociaux, analyse de contenu, etc.
  - Défis et perspectives du text mining.

- Activité pratique :

  - Analyse d'un jeu de données textuelles réel en utilisant des techniques avancées.

Ressources Recommandées

- Livres et sites:

– Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691207544/text-as-data

– Silge, J., & Robinson, D. (2017). Welcome to Text Mining with R | Text Mining with R. O'Reilly. https://www.tidytextmining.com/

– Daniel Jurafsky et James H. Martin.2024. Speech and Language Processing. (https://web.stanford.edu/~jurafsky/slp3/)