Sorbonne Nouvelle
université des cultures

LZML041 - Statistiques textuelles

Séance 1 - Introduction au Text mining

Lila Kim lila.kim@sorbonne-nouvelle.fr

Paris Sorbonne Nouvelle – 2024-2025

Description du cours

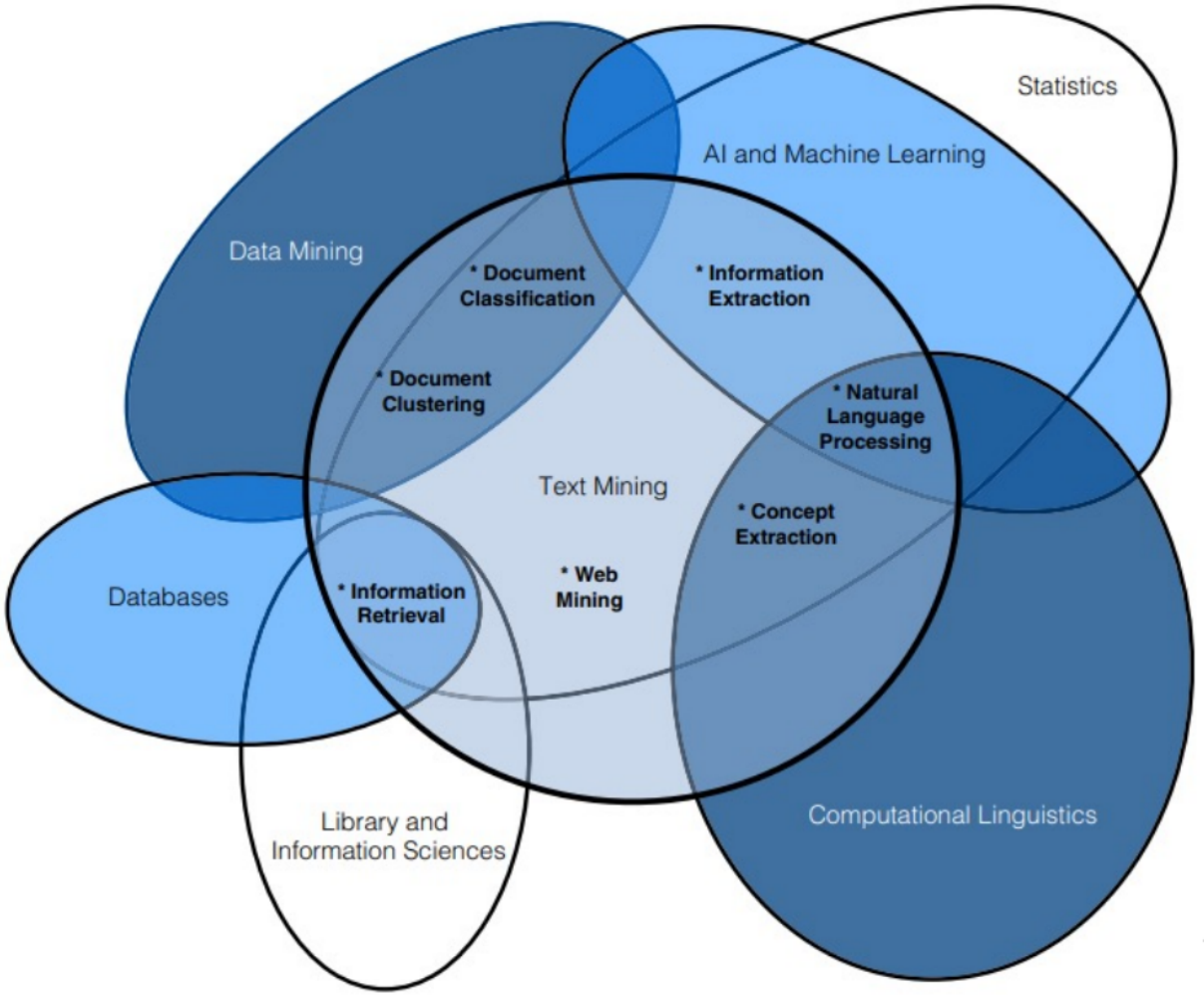
- 13 séances de 2h pour **s'initier au Text mining**
- Un mix de théorie et de pratique avec **l'outil Python/Jupyterlab**
- **Évaluation en contrôle continue** avec des tests courts toutes les 3 séances et un travail final à rendre en mai (50/50)

Plan du cours

- Introduction Générale
- Introduction à Python
- Qu'est-ce qu'un corpus ?
- Technique de pré-traitement (nettoyage, normalisation, segmentation, tokenization, ...)
- Text to numbers (Bag of words, frequency, TF-IDF, n-gramm model)
- Text similarity (mesures d'association, mesures de similarités)
- Text Clustering (K-means and Hierarchical algorithms)

Qu'est-ce que le Text mining ?

- **Le text mining**, également appelé exploration de texte (Knowledge-Discovery in Text, KDT), est **le processus consistant à extraire des informations significatives et des motifs** à partir d'une vaste collection de **textes**.



Une histoire brève

- L'un des premiers exemples (Frost 1976) est le catalogue de bibliothèque attribué à Thomas Hyde (1674) pour la Bodelian Library de l'Université d'Oxford.

- Des années 1800, la classification décimale de Dewey a été introduit : une fiche de bibliothèque contenant des métadonnées.

- En 1957, Hans Peter Luhn a fait la démonstration d'un premier ordinateur IBM 701 indexant des textes en utilisant son algorithme KWIC (Key Word in Context) pour générer des résumés de documents.

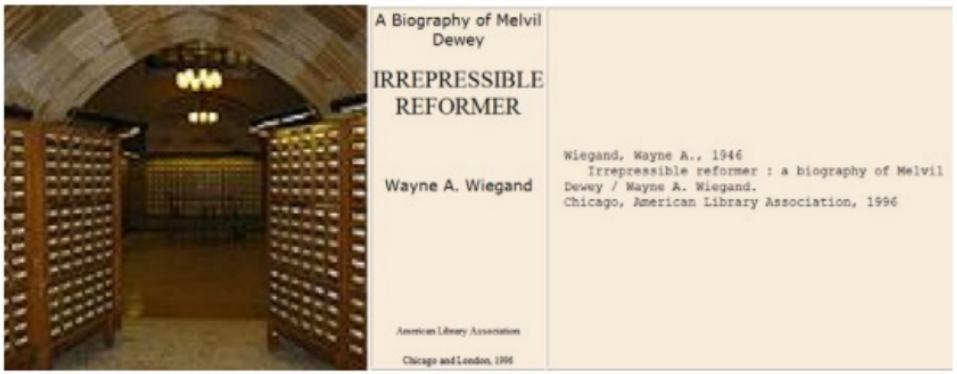


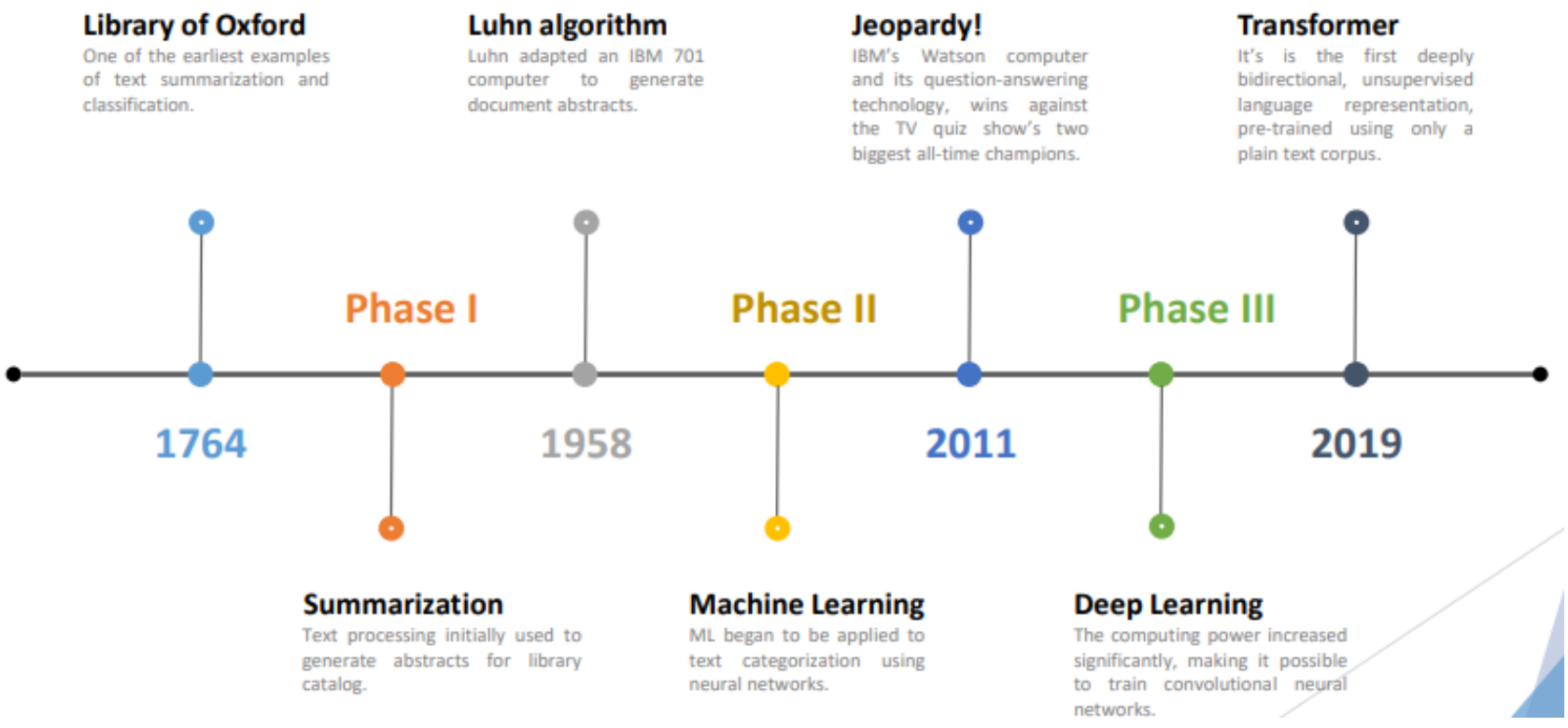
FIGURE 1.1
The library card catalog at Yale University and an index card. *Source: http://commons.wikimedia.org/wiki/File:Yale_card_*

IBM 701 Electronic analytical control unit

Qu'est-ce que le Text Mining Aujourd'hui ?

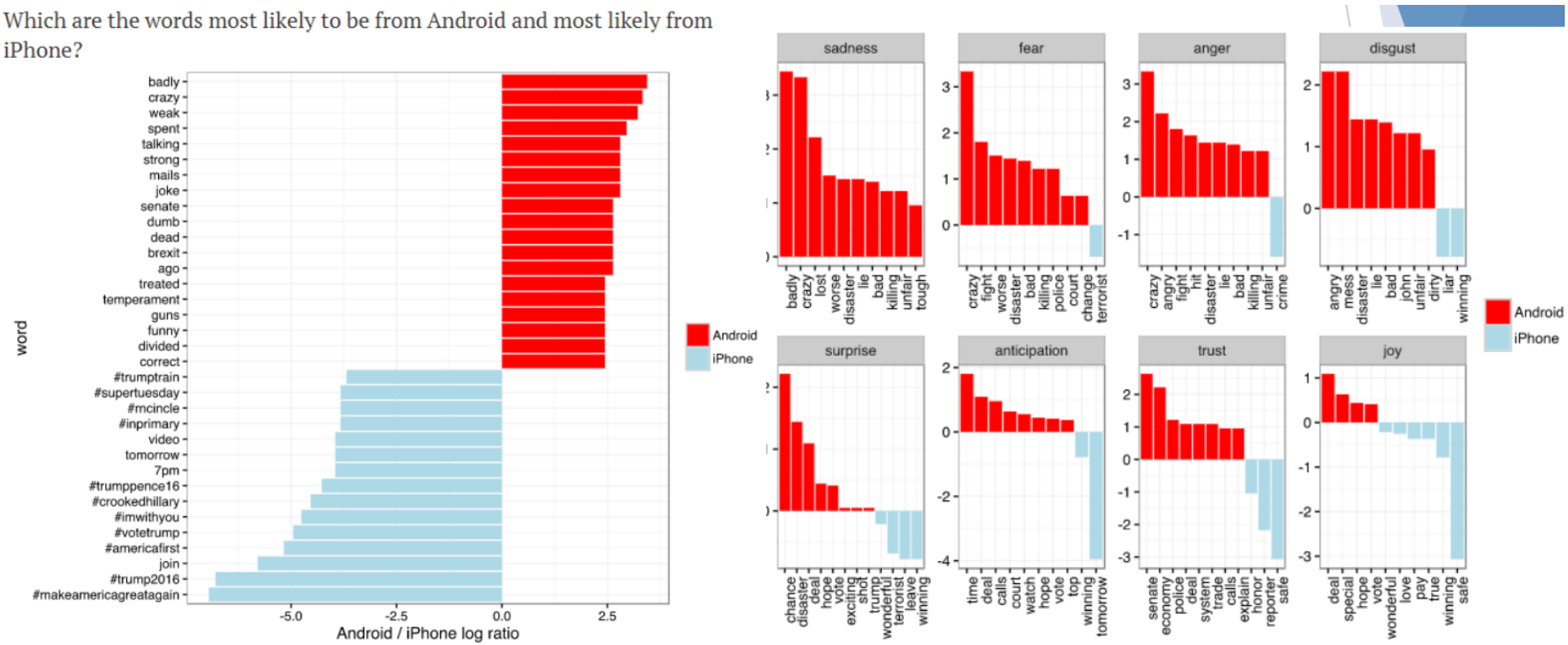
- Aujourd'hui, le Text Mining est un terme générique qui décrit une série de technologies permettant d'analyser et de traiter des données textuelles.

- Nous pourrions définir le text mining moderne comme **le processus d'extraction d'informations significatives** à partir de **textes non structurés** par l'application de **techniques analytiques avancées**.



Étude de cas - Les tweets de Trump

- http://varianceexplained.org/r/trump-tweets/
- beaucoup de mots annotés comme des sentiments négatifs (à quelques exceptions près, comme "crime" et "terroriste") sont plus fréquents dans les tweets Android de Trump que dans les tweets iPhone de la campagne.



ChatGPT

Un exemple d'application : https://chat.openai.com/chat

Exemples d'application

Dans le monde des affaires, les techniques de text mining sont utilisées pour **révéler des informations, des modèles et des tendances** à partir de grands volumes de données non structurées :

- Analyse des données des réseaux sociaux
- Publicité contextuelle
- Enrichissement de contenu
- Filtrage de spam
- Prévention de la cybercriminalité
- Amélioration du service client
- Investigation des réclamations simplifiées
- Gestion des risques
- Gestion des connaissances
- Intelligence d'affaires

Pratiques de Text Mining (1)

- **Recherche d'information** (Information Retrieval - IR)
  - Utilisation d'algorithmes basés sur un ensemble de requêtes prédéfinies pour retrouver des documents correspondant à une recherche par mots-clés.

- **Regroupement de documents** (Document Clustering)
  - Regroupement de paragraphes ou documents similaires (sur le plan lexical ou sémantique).

- **Classification de documents** (Documents classification)
  - Attribuer des étiquettes ou tags à des documents en utilisant des catégories prédéfinies en fonction de leur contenu.

Pratiques de Text Mining (2)

- **Fouille du web** (Web Mining)
  - Adaptation des techniques de text mining au format des informations disponibles sur le Web (blogs, e-mails, tweets, etc.).

- **Extraction d'informations** (Information Extraction - IE)
  - Découverte de données structurées à partir de données non structurées, nécessitant souvent des algorithmes spécialisés et une expertise thématique.

Pratiques de Text Mining (3)

- **Traitement du Langage Naturel** (Natural Language Processing - NLP)
  - Utilisation d'algorithmes d'apprentissage automatique pour exécuter différentes tâches basées sur des données textuelles.

- **Extraction de concepts** (Concept Extraction)
  - Composante complexe, car un concept peut représenter plusieurs termes selon le contexte du texte et les ressources linguistiques disponibles.
  - Cette extraction nécessite souvent une combinaison d'expertise humaine et d'intelligence artificielle pour interpréter le sens du texte dans son contexte.

Supervisé vs. Non supervisé

- **Apprentissage supervisé** (Supervised Learning)
  - Les algorithmes apprennent un classificateur ou déduisent une fonction à partir d'un ensemble de données d'entraînement étiquetées.
  - Cela leur permet de réaliser des prédictions sur des données non vues.

- **Apprentissage non supervisé** (Unsupervised Learning)
  - Aussi appelé apprentissage sans enseignant (learning without a teacher).
  - Contrairement à l'apprentissage supervisé, seul un ensemble de données d'entrée est fourni au modèle.
  - L'objectif est de découvrir des schémas cachés et des informations utiles dans cet ensemble de données inconnu.

Sous-catégorie de l'apprentissage non supervisé

- **Clustering** (Regroupement)
  - Identifie des données textuelles similaires entre elles et détecte des regroupements naturels (ou clusters) existants dans l'ensemble de données.

- **Associations**
  - Utilise une méthode basée sur des règles pour trouver des relations entre les variables d'un ensemble de données donné.

Comment fonctionne la tâche de Text Mining ?

Six phases qui décrivent naturellement le cycle de vie de la science des données d'un projet d'exploration de données :

1. Déterminer l'objectif de l'étude
2. Compréhension des données
3. Préparation des données
4. Modélisation
5. Évaluation
6. Déploiement

Projet de Text Mining en pratique ?

Le processus d'exploration de texte comprend les étapes suivantes :

1. Consntruction
2. Pré-traitement
3. Transformation
4. Application d'une technique
5. Évaluation
6. Applications/visualisations