Introduction à la statistique textuelle

Olivier Godechot

Pourquoi compter les mots ?

- Limite des enquêtes par questions fermées
  - La standardisation des réponses est une déformation
  - Alternative : questions ouvertes, entretiens, etc.
- Existence aujourd'hui de sources écrites numérisées de grande taille
  - Internet
  - Base de données de presse (lexis-nexis, europress)
  - Base de données de revue (etc.)
  - Œuvres complètes, etc.
- Limite de nos capacités de lecture analytique
  - Comment analyser des milliers de pages

Qu'est-ce que compter les mots ?

- On travaille sur des chaînes de caractères
  - Les chaînes de caractères semblables = mots semblables
  - Les chaînes de caractères dissemblables = mots différents
    - Ex: Evénement différent de événement
- Pour repérer les mots, il faut repérer les séparateurs de mots :
  - Espace, signes de ponctuation, caractères spéciaux (sauts de paragraphe, de ligne, etc.)
- Choisir ou non de confondre majuscule et minuscule, voyelle avec accents et voyelle sans accents
- Regroupement de mots ou de chaînes de plusieurs mots
  - Segments de deux mots, de trois mots, etc.

Les unités de compte

- Statistiques ➔ corrélation. Deux variables corrélées : Des individus répondent la même chose à plusieurs variables.
- Deux mots corrélés ? Définir un individu ou « unité de contexte élémentaire »
  - Un individu = une personne.
    - Adapté pour réponse courte à question ouverte
    - Pb si réponse longue. Corrélation entre deux mots si énoncés à 15 minutes d'intervalle.
  - Un individu = une phrase
    - Unité de contexte, relativement cohérente
    - Mais phrases de taille différente
  - Un individu = un paragraphe
    - Idem
  - Unité de taille fixe : 10 mots / 12 mots, etc.
    - Taille fixe
    - Pas forcément le même contexte

Comment compter ?

- L'effectif par unité de compte …
- Ou la présence absence par unité ?

Quels mots compter ?

- Les mots les plus fréquents : Les mots outils…(de, le, et, …)
- Les mots les plus singuliers : Les hapax … en général, quelle que soit la taille du texte, 50% des formes graphiques
- Quels mots apportent le plus d'information ? les mots singuliers ou les mots les plus fréquents ?
- Tous les mots : ne rien rater vs beaucoup de mots n'apportent pas d'information. Quantité d'information trop importante
- Une sélection de mots qui ont du sens : information plus précise vs arbitraire de la sélection.

La lemmatisation

- Regroupement de forme lexicale sous un même lemme.
- Regroupement des pluriels et des singuliers, des conjugaisons et des déclinaisons d'un même mot.
- À la main vs automatique (dictionnaire)
- Risque d'erreur important : ex. bus → boire

Les limites de la statistique textuelle

- Le problème de l'homonymie
  - Ex : La petite brise la glace / Le boucher sale la coupe.
- Le problème de la polysémie
- Le problème de la synonymie
- Le problème de la modalisation
  - Formules négatives : travailler et ne pas travailler
  - Nuances
- Arbitraire des modalités de réponse fermées vs arbitraire des catégories du langage
- Le résultat : différence de contenu ou différente manière (sociale) de s'exprimer

Les traitements statistiques sur les mots

- Pas de spécificité statistique – seulement des affinités. On a juste des variables qui sont des mots.
- L'analyse des données
  - Analyse factorielle / Analyse de correspondance multiple
  - Classification ascendante hiérarchique
- La régression
- Les analyses en termes de distance lexicale

Les usages

- Bons résultats : Questions ouvertes dont les réponses sont standardisées et stéréotypées
- Résultats discutables : Entretiens, si texte très différent.
- Les alternatives
  - La lecture simple (utilisation de toute l'information, difficulté à accéder aux phénomènes de nombre, pas d'aspect systématique)
  - Le codage exhaustif (très bon résultats, coûts)
- Les hybridations.
  - Outil possible en complément d'autres techniques
  - Outil élémentaire pour un codage
  - Outil confirmatoire de la lecture simple

Les logiciels disponibles

- Mes macros pour SAS
- Spad-t et spad
- Alceste
- Hyperbase
- Prospero