ISOCO : Produire des statistiques à partir de données textuelles

Patrice Baubeau

Journée mutualisée des PUD de Nanterre et des Grands Moulins

PUDN – MSHM – 09/12/2021

Maison des Sciences de l'Homme Mondes

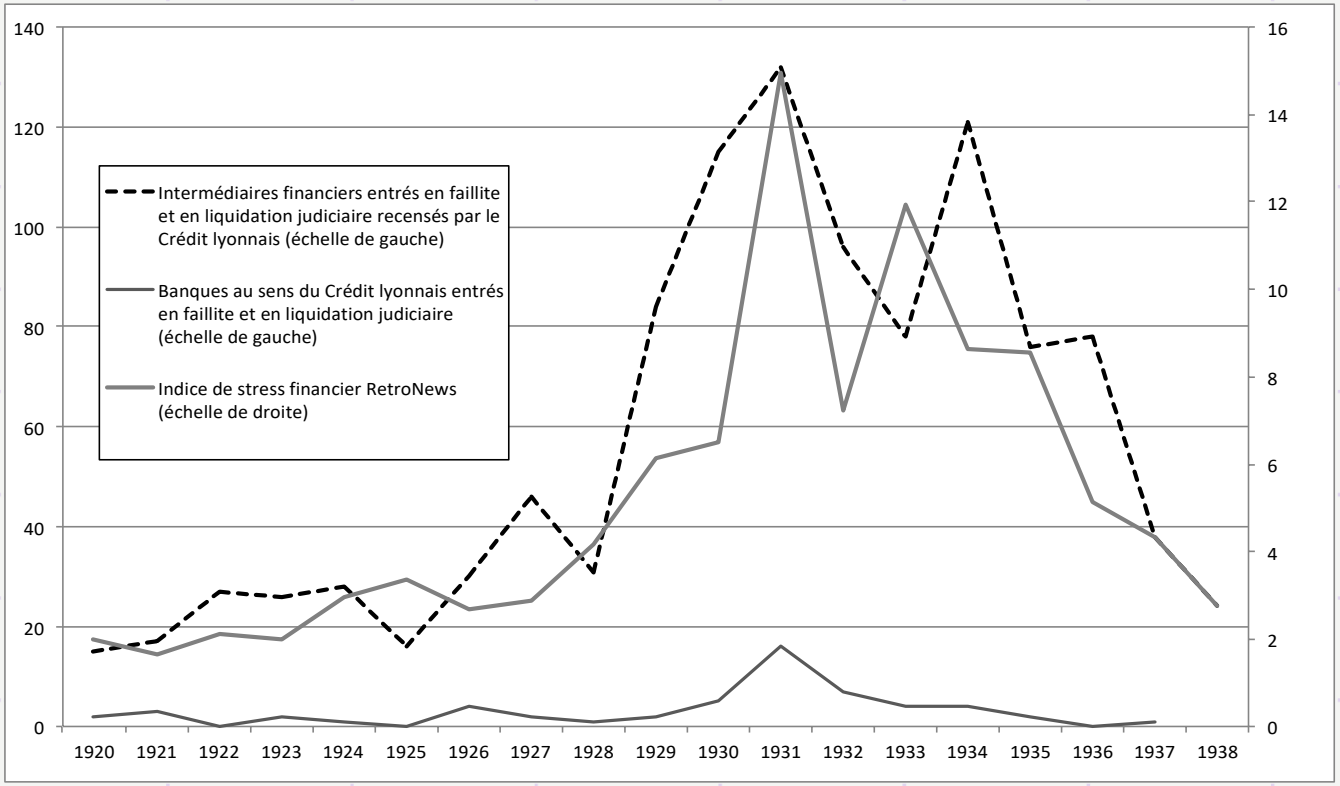
IDHES UMR 8533

Université Paris Nanterre

Produire des statistiques à partir de données textuelles
Le projet ISOCO

- A l’origine du projet :
  - Comprendre les motivations de l’action des pouvoirs publics face à une crise de grande ampleur : 1930
    - Cela suppose de rapprocher la lecture de la situation par les contemporains des données dont ils disposaient et de la manière dont ils les ont intérprétées
  - Source privilégiée : projets et propositions de lois au Sénat, débats, amendements
    - Travail d’identification et de calage chronologique réalisé en janvier-mars 2020
    - Travail d’archives programmé pour mars 2020
- Au bonheur de la sérendipité et du Covid réunis
  - Les archives du Sénat ferment en mars 2020
    - Il faut trouver une métrique alternative à l’action du Sénat : remonter de l’action aux opinions, c’est-à-dire à la lecture de la crise
    - Solution : comptabiliser les éléments qui affectent en « + » ou en « - » l’opinion sur la situation du secteur financier
  - 35 termes et expressions testés entre 1920 et 1938, 34 retenus :
    - Aéropostale, Banque, Banque Adam, Banque Charpenay, Banque d‘Alsace et de Lorraine, Banque de Grenoble, Banque de l‘Union parisienne, Banque nationale de Crédit, Banqueroute, BNC, Comptoir Lyon Alemand, Contrôle bancaire, Contrôle des banques, Corruption, Crédit lyonnais, Crise, Déflation, Dépression, Faillite, Garantie des dépôts, Hanau, Krach, Liquidation judiciaire, Oustric, Panique, Panique bancaire, Protection de l‘épargne, Récession, Ruée bancaire, Run bancaire, Scandale, Société centrale des banques de province, Société générale, Stavisky, Suspension de paiement
  - Une pondération terme par terme
    - Par une mesure de fréquence relative correspondant au ratio du nombre d’occurrences pour l’année considérée par le total des occurrences sur l’ensemble de la période
    - En effet, il y a 759 457 occurrences du terme «Banque» recensées par Retronews entre 1920 et 1938, mais seulement 21 occurrences de l’expression « Run bancaire »

Produire des statistiques à partir de données textuelles
Le projet ISOCO



Le problème clé : peut-on généraliser ce résultat ?

Indice de stress financier – Retronews. Cet indice repose sur la fréquence relative des termes suivants (voir encadré) : Garantie des dépôts, Protection de l'épargne, Contrôle des banques, Contrôle bancaire, Banque, Run bancaire, Ruée bancaire, Panique bancaire, Panique boursière et Krach.

Difficultés bancaires – Crédit lyonnais. Il s'agit : 1) du recensement des faillites et liquidations judiciaires d'intermédiaires financiers (banques et banquiers, quelle que soit leur forme juridique et leur activité principale) réalisé par les services du Crédit lyonnais et conservé dans les archives historiques de Crédit agricole S.A. (129AH110) et 2) des faillites et liquidations judiciaires des banques *stricto sensu* au sens du Crédit lyonnais (voir *infra*) recensées dans l'Album.

Lecture. Lorsque l'indice de stress financier se trouve en-dessous de 5 % environ, on peut considérer que l'image de la situation financière que renvoient les journaux ne correspond pas à une situation de crise généralisée ; lorsque cet indice s'élève, cela signifie une dégradation de la perception de la situation financière.

Produire des statistiques à partir de données textuelles
Le corpus : RETRONEWS

- Naissance de la « civilisation de la périodicité » et de l'opinion publique
- D. Kalifa et ses co-auteurs emploient l'expression dans*La Civilisation du journal. Histoire culturelle et littéraire de la presse française au XIXe siècle*, Paris, Ed. du Nouveau Monde, 2011
- Leur travail légitime largement l'idée d'une étude du pays à partir de la presse
- Mais peut-on relier presse et opinion publique ?
- Oui ; en dynamique, la presse reflète autant qu'elle façonne l'opinion, sans cela elle ne se vendrait pas
- Non : « l'opinion publique n'existe pas », ou plus exactement elle est une construction sociale
- Pas toute la presse… mais presque
- D'où deux choix :
- Un échantillon aussi large que possible de la presse française→ D'où Retronews – 1500 titres, dont 921 et plus de 11 millions de pages entre 1880 et 1938
- Une période resserrée sur l'apogée de la grande presse et de la presse populaire : 1880-1938
- Et deux conséquences :
- La nécessité de tenir compte des biais et des caractéristiques de la base Retronews
- Donc la nécessité d'établir un partenariat avec Retronews

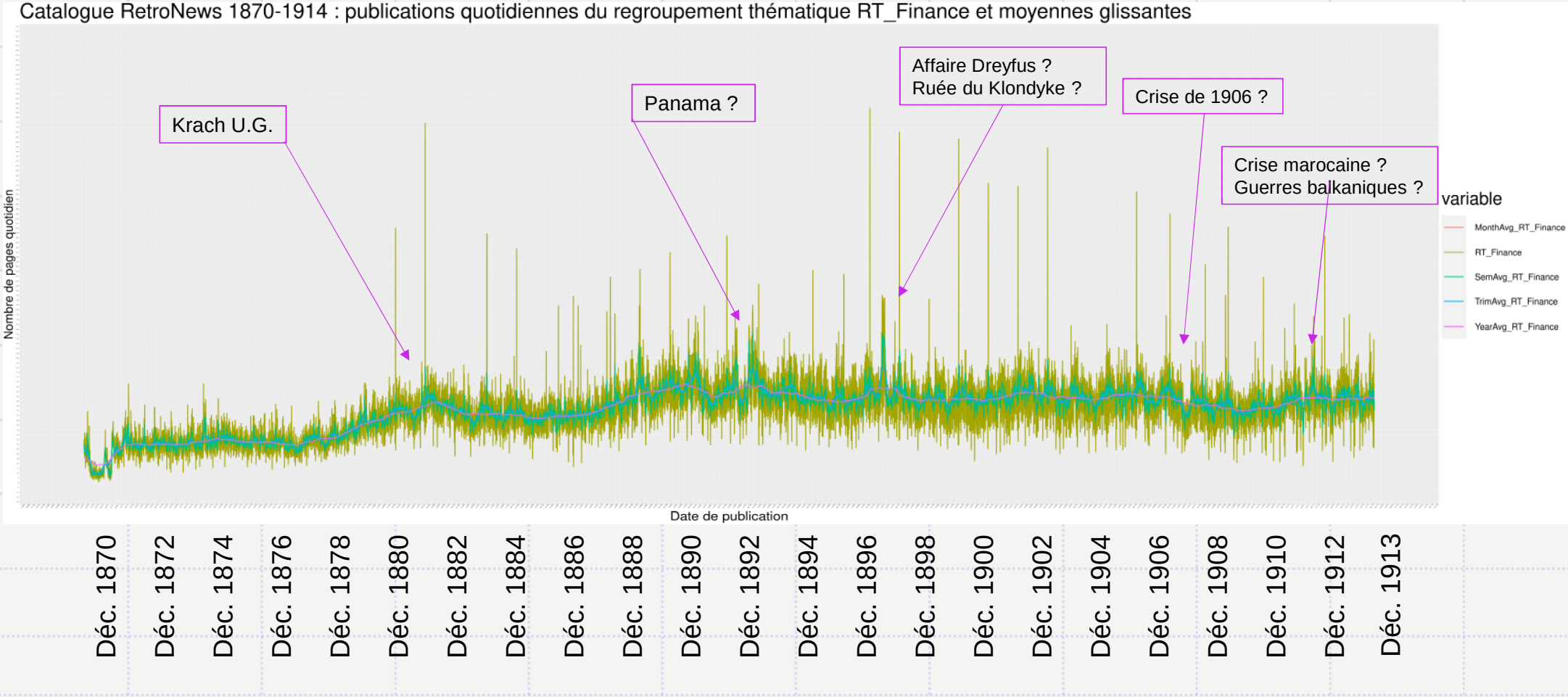
Produire des statistiques à partir de données textuelles
De l'aiguille à la meule

- L'exemple de Robert Shiller – *Narrative economics*
- Publié en 2019, l'ouvrage de Shiller (prix Nobel d'économie, 2013) prolonge une réflexion plus ancienne sur les « motivations » des agents économiques.
- Pour Shiller – vidéo – ces motivations s'actualisent dans des «*narratives* », des récits, qui sont susceptibles de se diffuser de manière « virale » et d'influence ainsi les perceptions, les anticipations et donc les actions des agents économiques
- Mais l'ouvrage est construit sur le suivi de termes ou d'expressions isolés à partir de Ngrams et de ProQuest, même si l'auteur est conscient des limites d'une telle approche
- Lien évolutif entre termes et concepts
- Nécessité de tenir compte des modalités d'usage des termes et des expressions
- Un domaine en plein essor
- Logiquement, du fait du développement des humanités numériques et du succès des techniques de butinage sur Internet…
- Par exemple l'utilisation de Twitter ou des requêtes Google pour prédire le mouvement des épidémies
- … des travaux de plus en plus nombreux, et très récents, exploitent cette veine :
- Baron, Verner and Xiong, « Banking crises without panics »,*QJE*, 2021
- Kabiri, James, Landon-Lane, Tuckett, Nyman, « The Role of sentiment in the economy: 1920 to 1934 », Cesifo WP, Feb. 2021
- Lennard, « Uncertainty and the Great Slump »,*EHR*, 2020
- Dans tous les cas, l'enjeu est de sortir du repérage d'éléments isolés ou uniques pour restituer des modalités qui traduisent une opinion, un sentiment. Donc de ne pas chercher l'exemple idoine mais de repérer des structures

Produire des statistiques à partir de données textuelles
Des outils très techniques

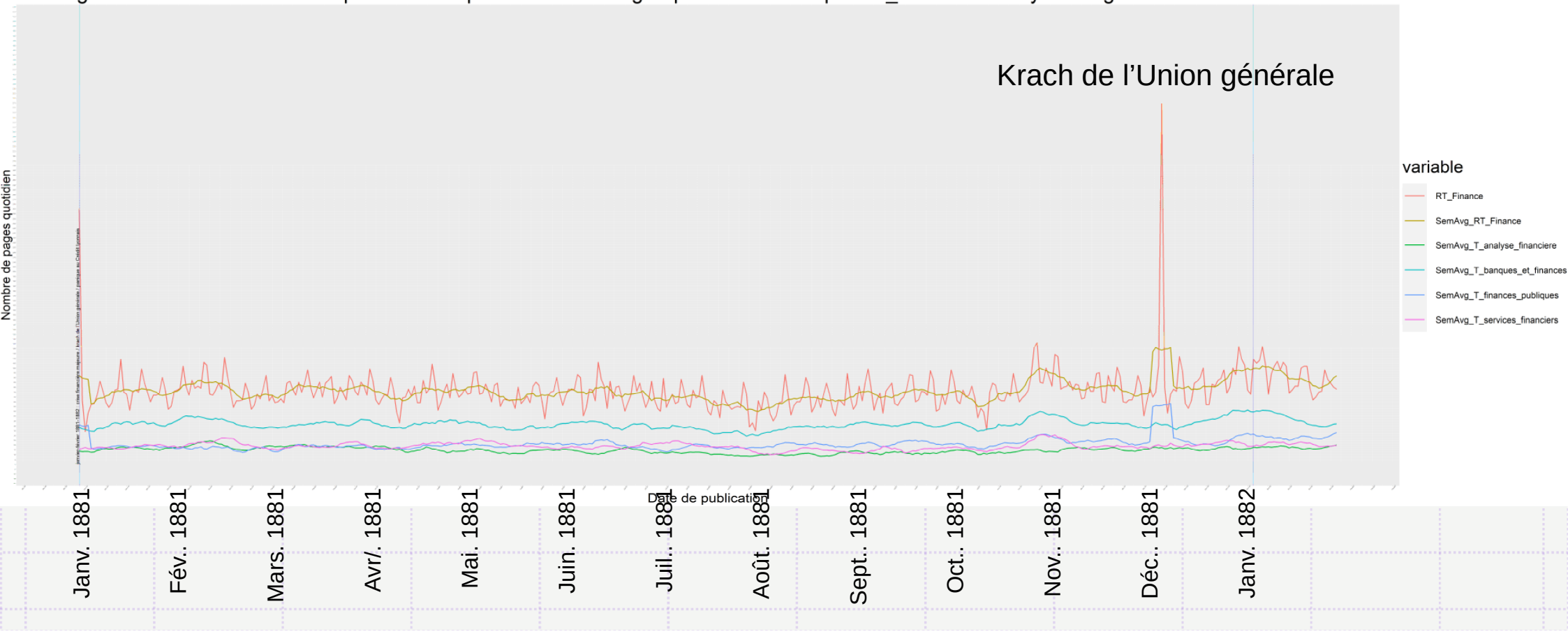
- Moissonner des masses de données
- Cela suppose un traitement massif des données numérisées
- D'où une extraction année par année
- En leur appliquant
- Des filtres
- Des analyses syntaxiques allant au-delà de la proximité pour toucher aux modalités
- Déceler les erreurs
- C'est tout le sens du projet : tester pour apprendre…
- Croiser les données
- En effet, un indice robuste ne peut pas reposer sur quelques termes isolés
- Il doit combiner
- Un aussi grand nombre de termes significatifs que possible
- Leurs modalités
- Tout en tenant compte du rapport évolutif entre un concept et son expression
- Et si possible un principe d'agrégation permettant de mesurer l'évolution « en + ou en – » de l'indice qui en résulte

Produire des statistiques à partir de données textuelles
Des outils très techniques



Produire des statistiques à partir de données textuelles
Des outils très techniques

Catalogue RetroNews 1881-1882 : publications quotidiennes du regroupement thématique RT_Finance et moyennes glissantes



Produire des statistiques à partir de données textuelles
Les objectifs

- Un projet collaboratif
  - IDHE.S et MoDyCo Nanterre ; PSE
  - PUDN – MSH.M
  - Retronews
  - BU Nanterre
- Surmonter la césure introduite par « l'ère statistique »
  - Intervient en France vers 1950
  - Pose le problème de la période 1938-1950 : discontinuités dans la presse / censure / pénuries
- Obtenir une résolution inférieure à l'année et à la nation
  - Mais suppose de tenir compte des périodicités spécifiques aux publications et aux différents thèmes
  - Soulève le problème de la représentativité à l'échelle des départements ruraux
- Explorer des champs nouveaux de l'opinion
  - Élargir l'enquête des indices économiques à des indices sociaux ou politiques

ISOCO

- Merci Mathilde !
- Merci Suzanne !

Merci Brian !

- Merci Philipp !