L'analyse lexicale au service de la cliodynamique : traitement par intelligence artificielle de la base Google Ngram

Jérôme Baray, Albert da Silva, Jean-Marc Leblanc

L'analyse lexicale au service de la cliodynamique : traitement par intelligence artificielle de la base Google Ngram

Jérôme Baray – Albert Da Silva – Jean-Marc Leblanc

UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

UPEC Connaissance - Action — UNIVERSITÉ PARIS-EST CRÉTEIL VAL DE MARNE

Projet PEPS CNRS/UPE Eclavit

Objectifs & Méthode

Qu'est ce que la cliodynamique ?

La cliodynamique est un domaine de recherche assez récent qui considère l'histoire comme un objet d'étude scientifique. L'étymologie du terme est grecque composé de Clio (Κλειώ / *Kleió*, de κλέω / *kleô*, « célébrer, chanter »), la muse de l'histoire qui chante le passé des hommes et des cités, et de dynamique, *dynamikos* (« puissant », « efficace »), dérivé de δύναμις, dynamis (« puissance ») qui sous-entend de façon générale l'analyse des changements. De nature transdisciplinaire, la cliodynamique tente ainsi d'expliquer les processus dynamiques historiques comme la montée ou l'effondrement des empires ou civilisations, les cycles économiques, les booms de population, les modes grâce à la modélisation mathématique, le datamining, l'économétrie ou encore la sociologie culturelle. Les big data agrégeant des données historiques, archéologiques ou économiques sont la matière permettant d'alimenter ces modèles quantitatifs. La cliodynamique comporte également un volet empirique examinant l'adéquation des hypothèses et des prévisions de modèles dynamiques avec les données historiques. Elle s'inscrit dans la démarche de la cliométrie ou « *new economic history* » qui étudie l'histoire grâce à des méthodes tirées de l'économétrie.

Enjeux de recherche

Il s'agissait d'une part de concevoir une méthode robuste d'analyse lexicale capable de traiter une série de très gros corpus datés dont le contenu évolue à travers le temps (big data) avec l'enjeu de cerner les évolutions sociétales et les grandes périodes historiques clés dans le cadre de la cliodynamique.

L'analyse lexicale s'est d'autre part penchée sur les enseignements à tirer de la base de données Google books Ngram (https://books.google.com/ngrams) qui détaille le nombre d'occurrences des mots utilisés année après année dans les publications scannées et intégrées au moteur de recherche Google Books. On considère que cette base a compilé environ 20% des livres publiés dans les langues majeures. Nous nous sommes focalisés sur les ouvrages en langue anglaise publiés aux Etats-Unis et en Grande Bretagne. L'objectif a été de cerner l'utilisation plus ou moins forte de certains mots selon les époques, la période d'étude ayant été fixée de 1860 à 2008. Les ouvrages avant 1860 paraissent être en bien moins grand nombre et la base de données Google books Ngram s'arrête à l'année 2008.

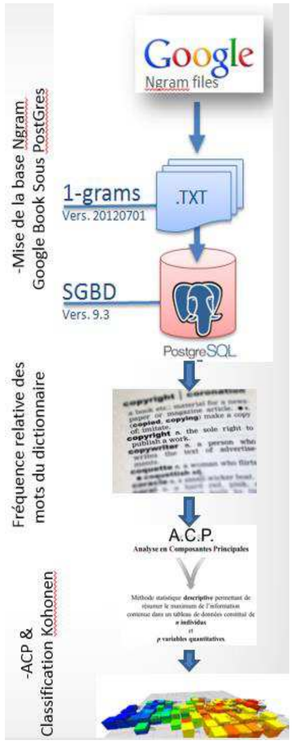
Principe de la méthode

La méthode a été de constituer dans un premier temps un dictionnaire des mots anglais les plus usités en faisant abstraction des termes à double sens, préposition, articles, pronoms. Ce dictionnaire a dans cette version initiale rassemblé 1592 mots couvrant de nombreux aspects de la vie sociale et culturelle avec des termes liés à la politique, à la religion, aux arts et aux sciences, à l'industrie, aux objets, à la famille et aux sentiments.

Dans un second temps, il a été déterminé la représentation en % de chacun de ses mots au sein du dictionnaire année après année après avoir mis l'imposante base Ngram Google Books (1-gram) sur Postgresql. Certains mots comme "king" ou "queen" sont très bien représentés dans le dictionnaire au 19ème siècle avec le règne et la puissance des pouvoirs royaux en Europe, mais l'usage de ces locutions décline au 20ème siècle. On constate donc une évolution constante de la fréquence des mots dans les ouvrages au fur et à mesure des époques.

La troisième étape a été d'effectuer une analyse factorielle en composantes principales centrée et normée sur le tableau décrivant la représentation des mots en % selon les années de 1860 à 2008 (1592 colonnes où les mots représentent les variables et 141 lignes représentant les individus statistiques c'est-à-dire les ouvrages publiés année après année et recensé dans Google Books). Une classification des années en groupes est réalisée grâce à un réseau de neurones (carte auto-adaptative de Kohonen en IA).

Les étapes de la recherche (big data & textmining)

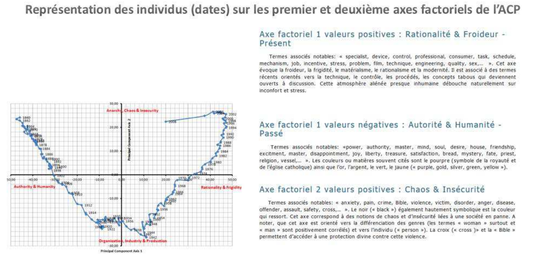


- ***Mise sous PostgreSQL de la base de données Google Book – N-gram***
- ***Création d'un dictionnaire rassemblant les mots les plus significatifs de la langue anglo-américaine (1592 mots) : ces mots font partie du vocabulaire de la culture, science, religion, politique, société, entreprise,…***
- ***Extraction de la fréquence des mots et de leur représentativité au sein du dictionnaire année après année de 1860 à 2008***
- ***Analyse factorielle (ACP) sur le tableau Années (1860-2008 individus) x mots (1301 variables)***
- ***Classification par cartes auto-organisatrices de Kohonen (IA)***
- ***Réalisation d'un site internet et d'un outil d'analyse capable de classifier et « dater » de nouveaux textes***

Résultats

L'analyse des résultats indique une certaine stabilité dans l'utilisation des mots du vocabulaire avec néanmoins des périodes de rupture. L'appréciation qualitative des nuées de mots permet d'interpréter la signification des axes et l'on s'aperçoit que l'axe 1 oppose des notions d'humanisme et de grands sentiments à celles de la rationalité froide. L'axe 2 sépare les années fastes à celles de l'insécurité et du chaos alors que l'axe 3 oppose les périodes plus libérales & individualistes à celles plus révolutionnaires et collectivistes.

Par ailleurs, certaines zones géographiques stratégiques sont bien plus évoquées à la veille de la majorité des conflits majeurs : Afghanistan, Irak, Biélorussie,…

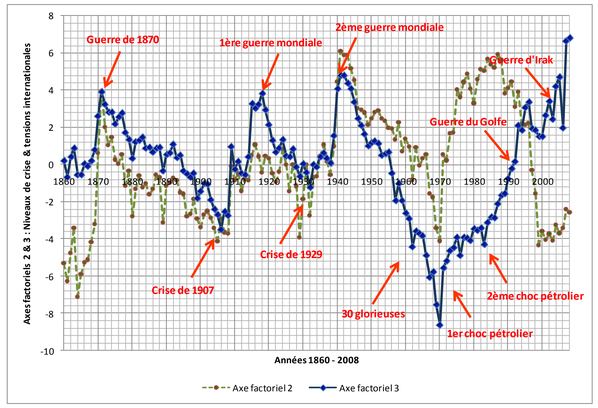


Pays jouant par une évocation plus fréquente de leurs noms un rôle important à la veille de crises ou de conflits (Corrélation / axe factoriel 3)

Les 8 périodes historiques issues de la classification

Groupe 1-1 : 1860-1890 – Religion, Justice & Grandeur
Groupe 1-2 : 1891-1910 – Prospérité, Joie de Vivre & Arts
Groupe 1-3 : 1911-1934 – Industrie, Commerce & Transports
Groupe 2-3 : 1935-1954 –Dictatures, Conflit Mondial & Sources d'Approvisionnement
Groupe 3-3 : 1953-1972 – Décolonisation, Révoltes & Sciences
Groupe 3-2 : 1973-1986 – Menaces, Santé et Innovations Techniques
Groupe 3-1 : 1987-2006 – Communications, Informatique, Services
Groupe 2-1 : 2008 – Amusement, Relations & Risques

Les variations des variables factorielles 2 et 3 montrent une relation surprenante mais logique avec les périodes de prospérité et de conflit :



Cette étude a également débouché sur la réalisation d'un site internet et d'un outil d'analyse capable de classifier et « dater » de nouveaux corpus : http://eclavit.univ-mlv.fr/DaText/index.php

Conclusion

Hormis ces enseignements, la méthode permet de :

- ✓ Dater et catégoriser les discours: commerciaux et marketing, communication institutionnelle, discours politiques,…
- ✓ Analyser des tendances sociales et de consommation, enquêtes marketing en créant un nouveau dictionnaire adapté à certaines disciples : ressources humaines, finance, marketing
- ✓ Réaliser des prévisions de tendances (périodes de crises, de conflits ou d'embellie économique à venir),
- ✓ Analyser des tendances socioéconomiques.

Bibliographie

- Turchin, P. (2003) *Historical dynamics: why states rise and fall*, Princeton University Press, Princeton, NJ.
- Diebolt Claude, Haupert Michael (2015) *Handbook of Cliometrics*, Springer-Verlag Berlin and Heidelberg GmbH & Co.