Cadet Aurélie
M2 HCS

Livrable Humanités numériques

Pour ce livrable j'ai choisi d'utiliser l'outil Voyant Tools afin de réaliser une analyse textuelle de document.

Résumé de mon affaire

Dans le cadre du Master, je travaille sur une affaire criminelle du XIXe siècle. Il s'agit de l'affaire Louis Menesclou, le viol et le meurtre d'une petite fille de quatre ans le 15 avril 1880 par son voisin de 20 ans. Le jour suivant le meurtrier a tenté de se débarrasser du corps en le brûlant dans son poële mais a été pris en flagrant délit par la police. L'affaire et le procès ont été très relayés dans les journaux et la population de Paris notamment a été très touchée par cette histoire et ce peu importe la classe sociale. Après un long procès, Menesclou a été jugé coupable et a été condamné à la peine capitale. Avec ma directrice de recherche j'ai décidé de centrer ma réflexion sur la place de la folie dans les affaires criminelles du XIXe siècle en prenant celle-ci en et plus particulièrement à l'image qu'on en avait à cette époque que ce soit les spécialistes ou la population dans son ensemble.

Présentation de l'outil

Voyant Tools est un outil d'analyse de textes ou de corpus de textes en ligne, on parle d'analyse automatique de texte ou de *text mining*. Il s'agit d'un projet open-source dirigés par deux universitaires canadiens Stéfan Sinclair et Geoffrey Rockwell qui a pour particularité d'être le plus accessible et le plus simple d'utilisation possible. Une première version moins aboutie a vu le jour en 2014 mais a rapidement été améliorée en 2015 afin d'arriver aux résultats obtenus aujourd'hui. Il s'adresse à la fois aux chercheurs, aux étudiants ou à toute personne ayant besoin d'un tel outil et les textes soumis à analyse peuvent être de différents formats comme HTML, XML, PDF etc.

L'interface est simple et facile à comprendre, il s'agit de déposer son texte dans le cadre prévu à cet effet par copié-collé d'un texte, en chargeant un corpus de documents ou par le biais des URL de textes. L'analyse est ensuite répartie en cinq parties distinctes : le nuage de tags appelé

Cirrus qui met en lumière les termes importants en les présentant de façon originale et colorée. Puis le texte analysé où Voyant Tools met en avant les mots importants en les surlignant directement sur le document. Ensuite les courbes de fréquence des mots les plus présents dans le texte qui permettent d'obtenir des résultats sous forme de graphiques et qui sont de ce fait plus facilement identifiables. Mais aussi le nombre de documents analysés, de mots dans le document et les mots-clés les plus présents. La dernière partie présente les mots-clés dans leur contexte. Il faut cependant nettoyer ses données pour que le résultat soit probant en enlevant par exemple les mots vides de sens et en concentrant l'analyse sur ce qui vraiment important par rapport aux attentes de l'utilisateur.

Etat de l'art

Les logiciels d'analyse textuelle ont pour but de réaliser des analyses de textes ou de corpus de textes en partant du principe que les textes sont des données organisées. Ces données doivent être exploitées de manière tant qualitative que quantitative conformément à la linguistique structurelle et à l'analyse de discours qui ont fortement inspirées la conception de ces logiciels. Les travaux sur des corpus littéraires et politiques de Jean-Paul Benzécri en 1960 ont beaucoup servis pour la conception de tous ces logiciels. D'ailleurs on peut dire que tous ces outils informatiques ont permis une évolution importante dans le domaine de la textométrie. Voici quelques exemples d'autres outils similaires à Voyant Tools :

Alceste (Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte) dont la première version a été créé en 1979 par la société IMAGE en collaboration avec le CNRS. C'est un outil qui est principalement utilisé dans le domaine des sciences humaines et sociales, il vise surtout à réaliser une analyse du vocabulaire et de la fréquence des mots employés. Les résultats sont, comme pour Voyant Tools, accompagnés de représentations graphiques afin de permettre à l'utilisateur une meilleure visualisation. Malgré son ancienneté il est aujourd'hui reconnu comme un outil très performent dans son domaine, la dernière version, Alceste2018 présente de nouveaux graphiques et de nombreuses innovations afin d'améliorer encore ses performances.

Tropes, logiciel d'analyse sémantique aussi appelé logiciel de « fouille de textes » crée en 1994 par Rodolphe Ghiglione, Agnès Landré et Pierre Molette. Les fonctions proposées par le

logiciel Tropes sont un éditeur d'ontologies, une classification dite arborescente, une analyse chronologique du récit, le diagnostic du style du texte… La dernière version de Tropes date de 2014 et il existe également une version plus aboutie, Tropes Zoom à partir de 1999. C'est un logiciel téléchargeable et gratuit utilisé principalement par des professionnels. Tropes peut être utilisé dans différents domaines comme la psychologie, la sociologie, marketing, linguistique, sciences politiques ou informatique.

Prenons également l'exemple de Wordstat dont la première version date de 1998 et a été produite par la compagnie Recherche Provalis. La technique de Wordstat se base sur une analyse de contenu à l'aide d'un dictionnaire mais aussi via des algorithmes d'exploration de textes, sa dernière version 8.0 date de 2018. Cependant contrairement à Voyant Tools par exemple il ne s'agit pas d'un logiciel gratuit et il est principalement utilisé par des professionnels en « analyse d'affaires et en analyse compétitive de sites web », il n'est donc pas accessible au grand public.

On peut en citer encore beaucoup, il y a également Hyperbase à partir de 1989, IraMuTeQ en 2009, Langage R créé en 1993, TXM dès 2007, ou encore Statistica dont la première version date de 1993.

Pourquoi ce choix d'outil ?

J'ai choisi cet outil car le document sur lequel j'ai souhaité travailler pour ce devoir est l'intégralité du dossier de l'affaire qui comprend de nombreux documents très différents, le dossier étant très complet et imposant. Initialement je désirais utiliser un logiciel de transcription tel que Transkribus ou From the page afin de retranscrire plus rapidement toutes mes archives mais je n'ai pas réussi à trouver la bonne version de From the page ni à réinstaller Transkribus sur mon ordinateur. J'ai donc cherché un autre outil et mon choix s'est porté sur Voyant Tools.

Venant de terminer la transcription je voulais avoir du dossier une vision globale afin de voir les principales idées qui en ressortent. De plus j'ai trouvé de prime abord que cet outil était facile d'utilisation car il est gratuit et il n'est pas nécessaire de le télécharger pour s'en servir. Il existe effectivement de nombreux logiciels de text mining mais celui-ci est, à mon sens, un des plus simple d'utilisation et donc d'exploitation des données. De plus, de nombreuses

personnes qui travaillent dans d'autres domaines, par exemple en Lettres ou dans les langues, m'ont conseillé Voyant Tools car elles s'en servent régulièrement et que cet outil a toujours plutôt bien répondu à leurs besoins.

Comment je m'en suis servie

J'ai d'abord commencé par découper le dossier en thématiques, j'ai mis d'un côté les témoignages, d'un autre les rapports de police et les rapports médicaux et d'un autre les autres documents inclassables. Pour me familiariser avec le logiciel j'ai commencé par faire un essai sur un extrait de livre de littérature que je connaissais bien afin d'avoir un aperçu des résultats où je savais d'avance ce qui allait principalement ressortir, une fois habituée à Voyant Tools j'ai commencé l'analyse du dossier.

Je me suis ensuite servi de Voyant Tools et j'ai analysé individuellement chacune de ces parties cependant les résultats obtenus n'étaient pas forcément très intéressants pour moi. Par exemple la partie « Témoignages » était très répétitive, les termes qui se répétaient étaient ceux des documents officiels prévus pour les témoignages et non les déclarations des témoins en elles-mêmes. Il en va de même pour les rapports de police et les rapports médicaux où rien ne ressort réellement, en revanche Voyant Tools m'a été plus utile pour les documents inclassables c'est-à-dire des lettres ou des descriptions de dessins et de croquis. J'ai donc décider d'analyser l'intégralité du dossier afin d'observer ce qui en ressortait principalement, grâce aux différents graphiques j'ai ensuite noté tous les termes qui ressortaient et qui me semblaient bien plus parlant qu'en analysant le document en plusieurs parties.

Les résultats obtenus

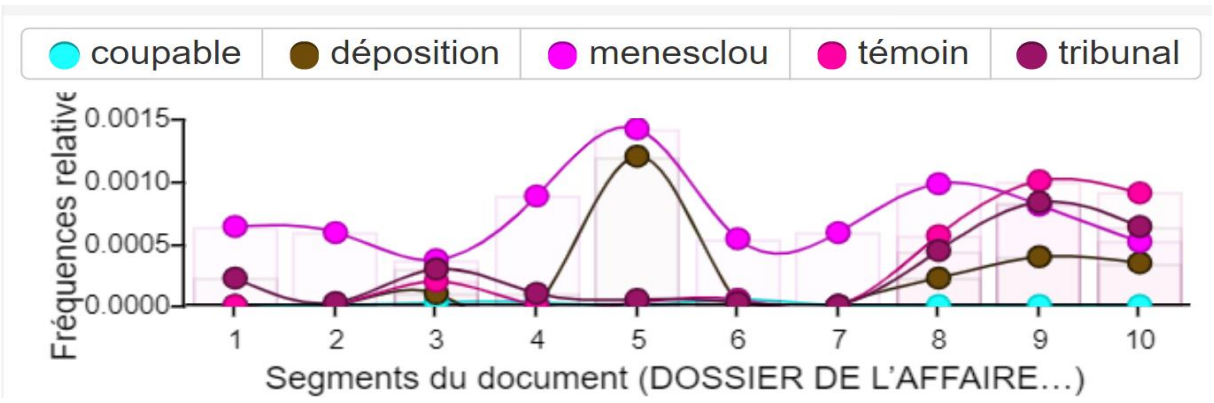
Je vais donc présenter dans cette partie les résultats que j'ai obtenu pour l'analyse de l'intégralité du dossier de l'affaire Menesclou. Pour mieux illustrer mes propos, j'ai réalisé des captures d'écran des représentations graphiques qui m'ont été les plus utiles pour la réalisation de mon mémoire en choisissant de mettre ici les 3 graphes qui m'ont le plus parlé dans ce travail.

Cadet Aurélie
M2 HCS

Nuages de mots

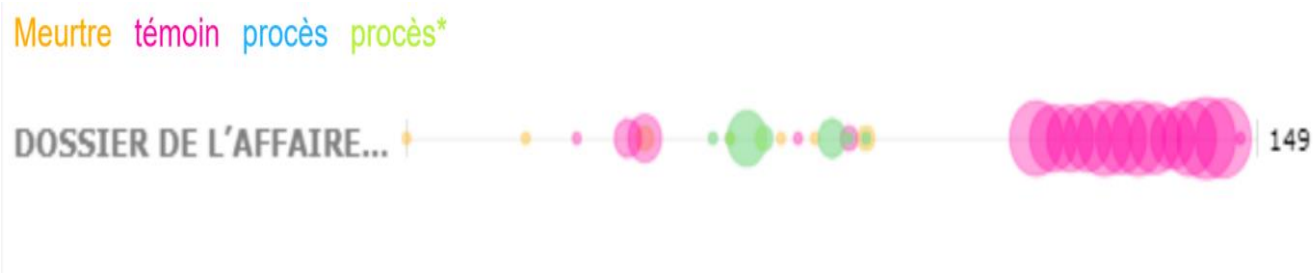
J'ai choisi la représentation Cirrus aussi appelé « nuage de mots » car je la trouve très visuelle, la forme et les couleurs utilisées permettent d'envisager les choses sous un nouvel angle. Ce sont les termes qui qualifient le mieux l'ensemble du dossier, un peu comme un brain storming. On remarque évidemment que le mot Menesclou est très récurrent ainsi que les notions de tribunal, instance, déposition, instruction, greffier, serment ainsi que d'autres correspondant au vocabulaire juridique. Je trouve cette disposition plutôt parlante et c'est un bon moyen pour quelqu'un ne connaissant pas du tout le dossier de s'en faire une bonne première idée.

Term Frequencies Chart



Cadet Aurélie
M2 HCS

Je trouve cette représentation très intéressante car selon moi une représentation graphique visuelle peut parfois être bien plus parlante qu'un document écrit. Il s'agit d'un « Term Frequencies Chart » qui nous donne une idée de la fréquence de certains termes en fonctions des parties du dossiers. Par exemple je remarque ici qu'on cite beaucoup le nom de l'accusé mais que le terme « coupable » n'est que très peu employé. Le mot « déposition » est très utilisé au milieu du document puis beaucoup moins par la suite et encore moins au début. On voit réellement l'évolution contrairement aux autres outils proposés par le logiciel.



Je pense que ce graphique est celui qui m'a certainement été le plus utile. En effet, cette représentation s'appelle Bubblelines, c'est un outil de visualisation de répartition de mots dans des documents. Il suffit d'entrer les mots qui nous intéresse et en fonction du nombre et de la taille des cercles on note à quel endroit il faut chercher pour trouver tel ou tel sujet. De plus on peut entrer les mots que l'on souhaite, effacer les précédents, isoler un seul terme. Dans la photo ci-dessus je remarque que tous les témoignages se trouvent à la fin du dossier, que le procès est évoqué plutôt au milieu et que le mot « meurtre » est très peu utilisé. C'est un outil très pratique pour s'orienter dans des dossiers volumineux.

Pour conclure je dirai que je trouve qu'il s'agit d'un très bon outil pour les professionnels comme pour les amateurs tout d'abord grâce à la simplicité de la plateforme et ensuite du fait que cet outil est gratuit. Les différents graphiques proposés sont très parlants et variés ce qui permet une bonne analyse de documents. Je l'ai trouvé bien plus accessible que d'autres outils que je voulais utiliser pour ce devoir.

Cependant j'ai rencontré quelques difficultés avec Voyant Tools, par exemple j'ai essayé de supprimer certains termes comme « j'ai » ou « qu'un » sans y parvenir. En effet, les termes étaient déjà enregistrés dans la liste des mots vides et même en variant l'écriture ils apparaissent toujours dans l'analyse ce qui rend plus difficile une bonne vision globale des documents. Je ne

sais pas si c'est moi qui n'y suis pas parvenue mais je n'ai pas réussi à sélectionner un graphique pour l'avoir en plein écran, les captures d'écran ci-dessus ont été obtenues grâce au zoom et au rognage des photos. Je trouve de ce fait, qu'il y a beaucoup d'informations sur une page, j'aurais aimé pouvoir sélectionner le graphique de mon choix pour pouvoir mieux l'utiliser et le visualiser.

Bibliographie

Site de Voyant Tools : https://voyant-tools.org/

Site Outils Froids expliquant l'outil : https://www.outilsfroids.net/2016/02/voyant-tools-un-puissant-service-de-text-mining-en-open-source/

Site EduTech Wiki expliquant l'outil : http://edutechwiki.unige.ch/fr/Voyant_Tools

Site de Alceste : https://www.image-zafar.com/Logiciel.html

Site de Tropes : http://www.tropes.fr/

Bénédicte Pincemin, Céline Guillot, Serge Heiden, Alexei Lavrentiev, Christiane Marchello-Nizia, « Usages linguistiques de la textométrie Analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex », Syntaxe et sémantique 2008/1 (N° 9), pages 87 à 110, URL : https://www.cairn.info/revue-syntaxe-et-semantique-2008-1-page-87.htm