Projet de visualisation analytique Master 1 Ingénierie Statistique et Informatique

Mehdi Fellat

16 janvier 2019

Table des matières

Table des figures

1 Introduction

Les données constituent un support de travail très intéressant pour les entreprises. Une donnée est un point de départ à un raisonnement ayant pour but de déterminer une solution à un problème. Dans notre cas, nous devons donner du sens à nos données et donc de savoir qu'est ce qu'elles représentent. La visualisation de données est un outil qui permet de transformer les données, initialement à l'état brute, dans un état facilement analysable par un individu. Ceci est très fréquent dans les entreprises. Les experts du domaine demandent aux ingénieurs de rendre les données "lisible" afin d'en exploiter de l'information. Nos données seront issues des projets retenus par l'Agence Nationale de Recherche (ANR).

2 Présentation du problème

2.1 L'Agence Nationale de Recherche

L'Agence Nationale de Recherche est une agence de moyens créée le 7 février 2005, qui finance la recherche publique et la recherche partenariale en France. Initialement créée sous la forme d'un groupement d'intérêt public par le gouvernement de Jean-Pierre Raffarin, elle est dotée depuis le 1er janvier 2007 du statut d'établissement public à caractère administratif. . L'ANR s'est substituée aux dispositifs ministériels pré-existants de financement incitatif : le fonds national pour la science (FNS) et le fonds pour la recherche technologique (FRT). Elle finance directement les équipes de recherche publiques et privées, sous forme de contrats de recherche à durée déterminée.

2.2 Objectifs

L'Agence Nationale de Recherche finance les projets de recherche publiques et privés. Le recensement de ces projets s'est fait dans un fichier de données comportant plus de 10 000 projets recensés. Le but de notre projet est de récupérer ces données, les traiter, et les analyser afin de déceler certaines informations non visibles à premières vues. Pour cela nous allons utiliser les outils basés sur la visualisation de données. Les graphes ainsi que tout concept visuel aidant à comprendre les données seront utiles dans notre projet. En particulier, nous allons nous servir de :

— Python
— La librairie Pandas
— Le logiciel Tulip pour la visualisation de graphes
— L'outil de visualisation d3
— R

3 Acquisition des données

3.1 Détails sur les données

Les données à disposition sont les projets financés par l'ANR entre 2006 et 2016.

	Taille des données	Nombre de colonnes
Données sur les projets financés par l'ANR entre 2006 et 2016	13451 projets	23 variables

FIGURE 1 – Détails sur les données.

3.2 Données sur les projets

Nous avons reçu les codes des projets ainsi que plusieurs informations comme les titres des projets, le résumé, le montant financé par l'ANR et d'autres données. Voici la sémantique des données recueillies sur le site de l'ANR.

Colonne	Type	Sémantique
Code du projet	String	Code national du projet
Titre	String	Titre du projet
Acronyme	String	Acronyme du projet
Résumé	String	Résumé du projet
Année de financement	int	Année de financement du projet
Lien du projet	String	Lien Internet du projet
Code du programme	String	Code du programme
Programme	String	Programme du projet
Lien Programme	String	Lien du programme du projet
Perspectives	String	Perspectives du projet
Publications et brevet	String	Projets et publications des partenaires
Résultats	String	Résultats du projet
Résumé de la soumission	String	Résumé de la soumission
Montant	double	Montant financé par l'ANR
Date de début	int	Année de début du projet
Durée en mois	int	Durée du projet
Coordinateur du projet	String	Coordinateur du projet
Identifiant de partenaire	int	Identifiants national des partenaires
Type d'identifiant	String	Type d'identifiant
Libellé de partnaire	String	Libellé des partenaires du projet
Sigle de partnaire	String	Sigle des partenaires du projet
Code du type de partenaire	String	Code du type de partenaires
Type de partenaire	String	Type de partenaires

FIGURE 2 – Détails sur les projets

4 Pré-traitement des données

Le pré-traitement des données est la première étape dans le développement de notre système. Le pré-traitement implique des techniques d'exploration de données afin de les transformer dans le format requis et ensuite de pouvoir les exploiter.

4.1 Sélection des données

Nom de la colonne	Explication
Code du projet	Code du projet
Titre	Titre du projet
Résumé	Résumé du projet
Année de financement	Année de financement du projet
Date de début	Date de début du projet
Durée en mois	Durée du projet
Montant	Montant financé par l'ANR
Libellé de partenaire	Libellé des partenaires du projet

FIGURE 3 – Détails sur les données.

Voici un exemple de données sélectionnés

	Code du projet	Titre	Résumé	Année de financement	Date de début	Durée en mois	Montant	Libellé de partenaire
0	ANR-11-BSV8-0005	Assemblage moléculaire du pilus chez S. pneumo...	Cibles anti-bactériennes innovantesLes antibio...	2011	2012.0	36.0	348479.0	Commissariat à l'énergie atomique et aux énerg...
1	ANR-11-BSV8-0008	Exploration fonctionnelle et structurale d'une...	PRO-RNase PÉtude structure fonction d'un nouve...	2011	2011.0	48.0	379973.0	Institut de Biologie Moléculaire et Cellulaire...
2	ANR-11-BSV8-0013	Régulation de l'épissage alternatif par la mac...	Régulation de l'épissage alternatif par la mac...	2011	2011.0	36.0	550000.0	Centre national de la recherche scientifique;C...
3	ANR-11-BSV8-0014	Bases moléculaires de la reprogrammation post-...	Bases moléculaires de la reprogrammation post-...	2011	2012.0	48.0	350000.0	Institut national de la santé et de la recherc...
4	ANR-11-BSV8-0015	La machinerie Hsp90-R2TP et l'assemblage de co...	Architecture du système chaperon Hsp90-R2TP : ...	2011	NaN	48.0	550000.0	Institut Pluridisciplinaire Hubert Curien;Inst...
5	ANR-11-BSV8-0019	Structures, mecanismes, et conception d'inhibi...	Structure et mécanism des polymérases des Aren...	2011	2011.0	36.0	350000.0	Laboratoire public;Laboratoire public
6	ANR-11-BSV8-0020	Etude integrative de la fonction du Médiateur ...	Compréhension des mécanismes fondamentaux de l...	2011	2012.0	36.0	350000.0	Institut de Biologie et de Technologies de Saclay
7	ANR-11-CEPL-0005	Éléments trace métalliques Perturbations clima...	Éléments trace métalliques, Pollution Upwellin...	2011	2012.0	48.0	764217.0	Domaines Océaniques;Laboratoire des Sciences d...
8	ANR-11-CEPL-0011	Analyse de la résilience des nouvelles formes ...	Analyse de la vulnérabilité et de la capacité ...	2011	2012.0	48.0	798990.0	Bureau de recherches géologiques et minières;G...
9	ANR-11-CESA-0001	Effets et mécanismes d'action du bisphénol A (...	Au cours de la dernière décennie, les inquiétu...	2011	2012.0	36.0	420000.0	Commissariat à l'énergie atomique et aux énerg...

FIGURE 4 – Exemple de données sélectionnées.

4.2 Nettoyage et adaptation des données des données

Comme dis plus haut, nous avons en notre possession les données de 13450 projets. Cependant, toute ces données ne sont pas complète, en retirant les projet ou il manque des informations, il nous reste 7677 projets. Dans la colonne "Libellé de partenaire" on voit qu'il y a plusieurs partenaires pour un seul et même projet. On veut "décoller" cette variable. On voudrait mettre notre tableau de donnée de la forme : Un partenaire -> Le projet sur lequel il travaille. En faisant cela on se retrouve avec un tableau de donnée de 19692 lignes. Grâce à cela, nous pouvons dors et déjà compter le nombre total de partenaires, il y en a :

A partir de là, notre tableau de données va donner naissance à un sous-tableau.

4.3 Données sur les villes de France

Dans le libellé de partenaire, on voit que certains centre possèdent le nom d'appartenance de la ville, ainsi, nous allons récupérer tout les partenaires appartenant aux neuf villes de France suivantes :

- Paris
- Lyon
- Toulouse
- Marseille
- Montpellier
- Bordeaux
- Strasbourg
- Nantes
- Lille

On passe d'un tableau de 19692 lignes à un tableau de 2280 lignes. On a ajouter une colonnes pour préciser la ville d'appartenance du partenaire. On peut maintenant compter le nombre de projets par ville :

- Paris : 718 projets
- Lyon : 309 projets
- Toulouse : 299 projets
- Marseille : 182 projets
- Montpellier : 202 projets
- Bordeaux : 199 projets
- Strasbourg : 119 projets
- Nantes : 131 projets
- Lille : 121 projets

5 Extraction de mots

Il serait intéressant de savoir a quelle thématique appartiennent les projets, ainsi on pourrait ajouter une variable "domaine" qui prendrait comme valeur "informatique", "chimie", "physique". Pour cela nous allons extraire, à partir du titre et du résumé de nos 2280 projets, les mots ayant une signification particulière, c'est à dire les mots qui reviennent dans des projets en particuliers. Par exemple, dans les projets liés à l'informatique, nous verrons (sûrement) dans les titres le mot "algorithmes. Ainsi, si 2 projets partagent ce mot, c'est qu'ils sont proches, et seront classés dans la même catégorie, à savoir "informatique". Pour cela nous allons utiliser une mesure statistique appelée TF-IDF.

5.1 TF-IDF

Le TF-IDF (de l'anglais term frequency-inverse document frequency) est une méthode de pondération souvent utilisée en recherche d'information et en particulier dans la fouille de textes. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus. Des variantes de la formule originale sont souvent utilisées dans des moteurs de recherche pour apprécier la pertinence d'un document en fonction des critères de recherche de l'utilisateur.

5.1.1 Fréquence du terme

La fréquence « brute » d'un terme est simplement le nombre d'occurrences de ce terme dans le document considéré (on parle de « fréquence » par abus de langage). On peut choisir cette fréquence brute pour exprimer la fréquence d'un terme.

Des variantes ont été proposées. Un choix plus simple, dit « binaire », est de mettre 1 si le terme apparaît dans le document et 0 sinon. À l'opposé, on peut normaliser logarithmiquement la fréquence brute pour amortir les écarts. Une normalisation courante pour prendre en compte la longueur du document est de normaliser par la fréquence brute maximale du document.

5.1.2 Fréquence inverse de document

La fréquence inverse de document (inverse document frequency) est une mesure de l'importance du terme dans l'ensemble du corpus. Dans le schéma TF-IDF, elle vise à donner un poids plus important aux termes les moins fréquents, considérés comme plus discriminants. Elle consiste à calculer le logarithme (en base 10 ou en base 21) de l'inverse de la proportion de documents du corpus qui contiennent le terme :

$$\mathrm{idf_i} = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

ou :
$|D|$
le nombre total de document dans le corpus.

$|\{\mathrm{d_j} : \mathrm{t_i} \in \mathrm{d_j}\}|$
nombre de documents ou le terme apparaît

5.2 Mise en place de TF-IDF

5.2.1 Graphe termes - termes

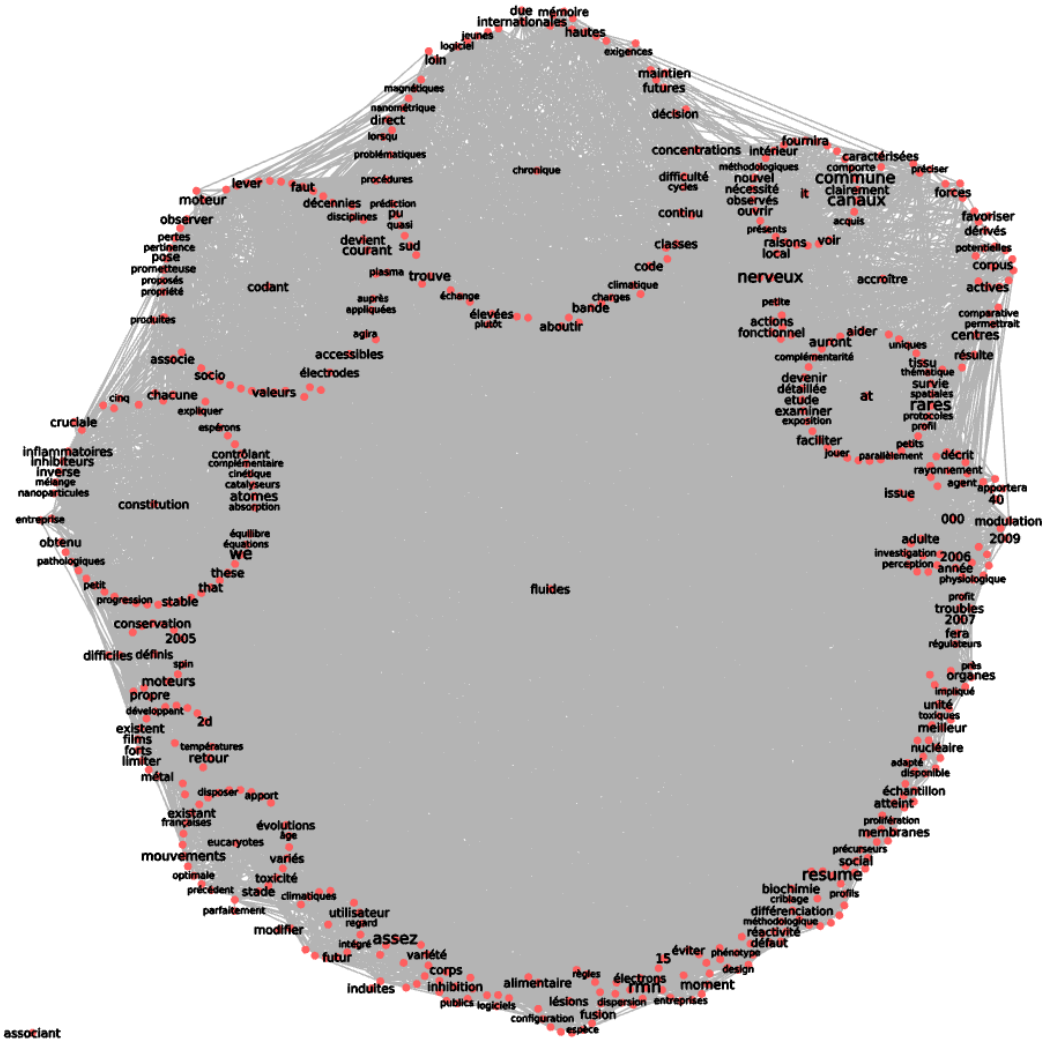

FIGURE 5 – Graphe de termes

Voici le graphe terme-terme dans lequel les sommets représentent les termes qui apparaissent avec une certaine fréquence dans un document. Cette fréquence est calculée avec un score tf-idf. Les termes sont reliés entre eux si ils apparaissent dans le même document. Le Bubble Tree de Tulip permet de représenter ces termes de sorte a ce que des termes apparaissant dans le même document soient dans la même bulle. Ainsi, les bulles peuvent représenter des sujets.

Les sujets de chimie sont représentés à gauche avec les termes 'inhibiteurs', 'atomes',' catalyseurs', 'équilibre',' nanoparticules', 'cinétique'.

En bas de cette bulle on a un petite bulle avec les termes 'métal' et 'températures', qui peuvent avoir un lien avec la chimie des matériaux.

En bas on peut lire les termes 'eucaryotes', 'évolutions', 'âge', 'toxicité', 'stade', qui peuvent avoir un lien avec la biologie.

En haut a droite on un une bulle avec les termes, 'étude', 'détaillé', 'examiner', 'exposition', 'protocole', 'thématique'.

Dans la plus grande bulle (en haut) on retrouve les termes 'internationales', 'climatique', 'prédiction', 'maintien', 'future', 'difficultés', qui peuvent êtres liés au problèmes dus au réchauffement climatique

La paramétrisation de notre fonction CountVectorizer s'est fait avec les paramètres

$$min_{df} = 0.02, max_{df} = 0.03$$

5.2.2 Graphe documents - documents

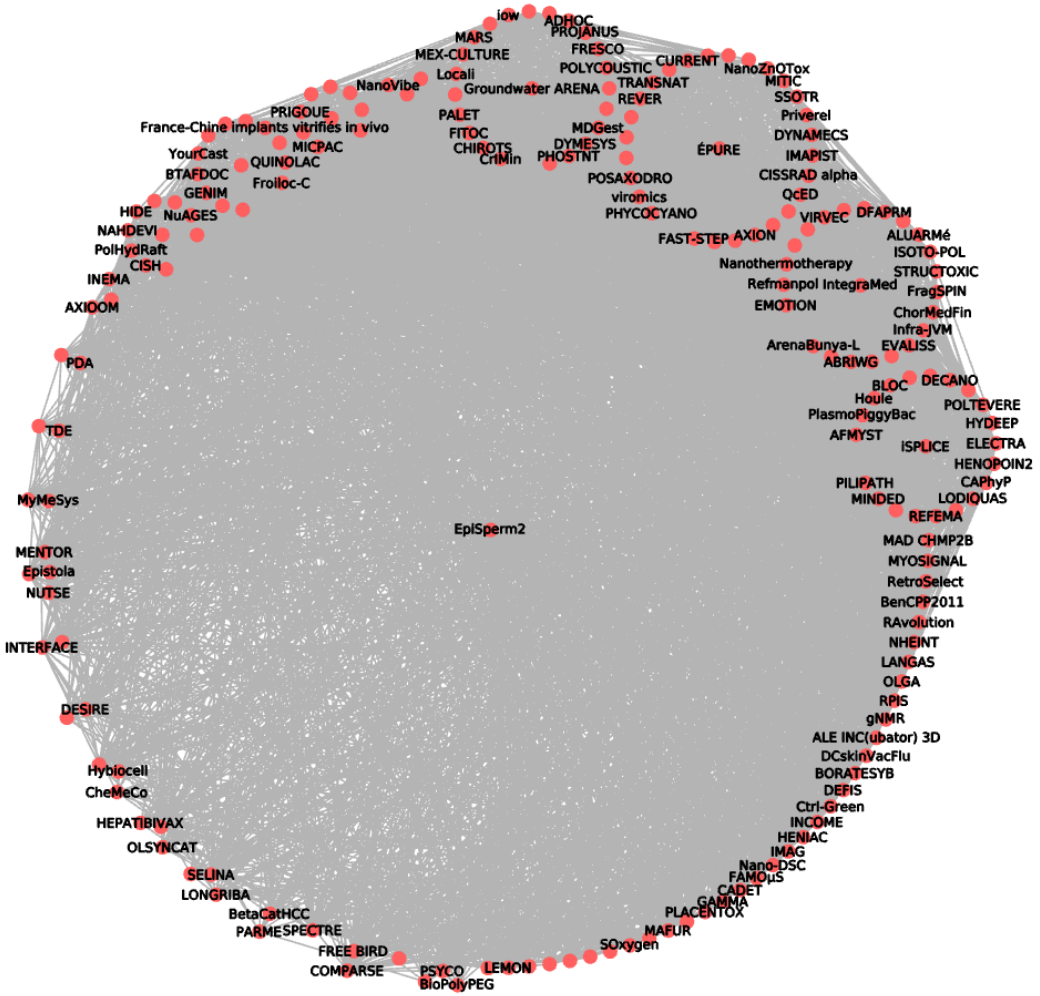

FIGURE 6 – Graphe de documents

Voici le graphe document document ou chaque sommet représente un document étiqueté par son nom de projet. Deux documents sont reliés par une arrêtes si ils possèdent des termes en communs. Ainsi, les projets apparaissant dans une même bulles on de forte chances de correspondre au même sujet.

Pour des soucis de complexité, nous avons choisis les 200 premiers documents, avec

$$min_{df} = 0.025$$

et

$$max_{df} = 0.03$$

5.2.3 Graphe termes - documents

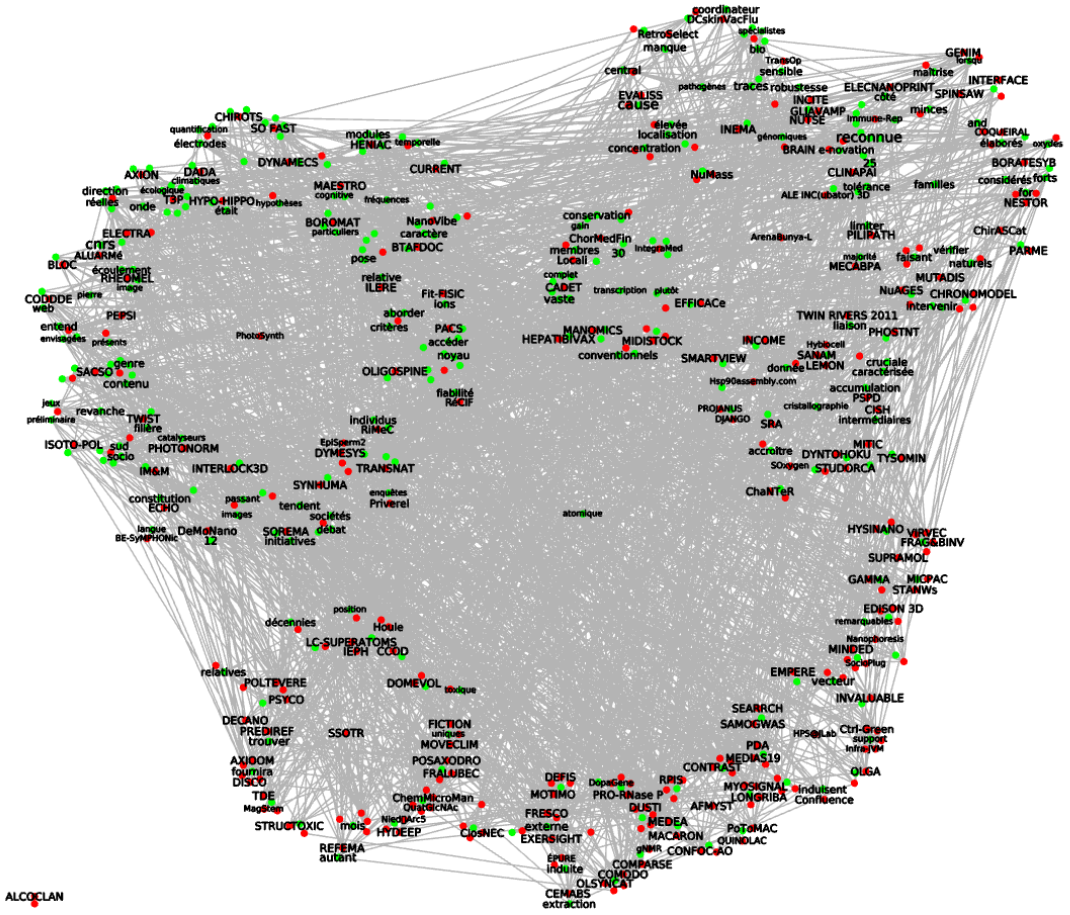

FIGURE 7 – Graphe terme - document

Les sommets en rouge représentent les documents, les sommets en verts représentent les termes. On comprend que si un terme apparaît plusieurs fois dans un document, alors ils seront proches dans le graphe.

Pour ce graphe nous avons pris les 300 premiers documents, la paramétrisation de CountVectorizer() s'est fait avec les mêmes paramètres que précédemment.

$$min_{df} = 0.025$$

et

$$max_{df} = 0.03$$

5.3 Thématiques

A partir des résultats obtenus via TF-IDF, nous avons relevé les mots suivants, auquel nous associons un thème donné.

Environnement	Informatique	Biologie	Sante	Mathématiques	Chimie	Physique
climat	données	bio	cognition	arithmétique	chimique	élec
eaux	algorithme	cellule	neuro	géométrie	molécule	energie
séisme	data	plante			carbone	thermo
sol	internet	terre			hydrogène	
toxicité		génome				
océan		vaccin				
écologie						

FIGURE 8 – Mots sélectionnés

6 Visualisation

6.1 Montant par année

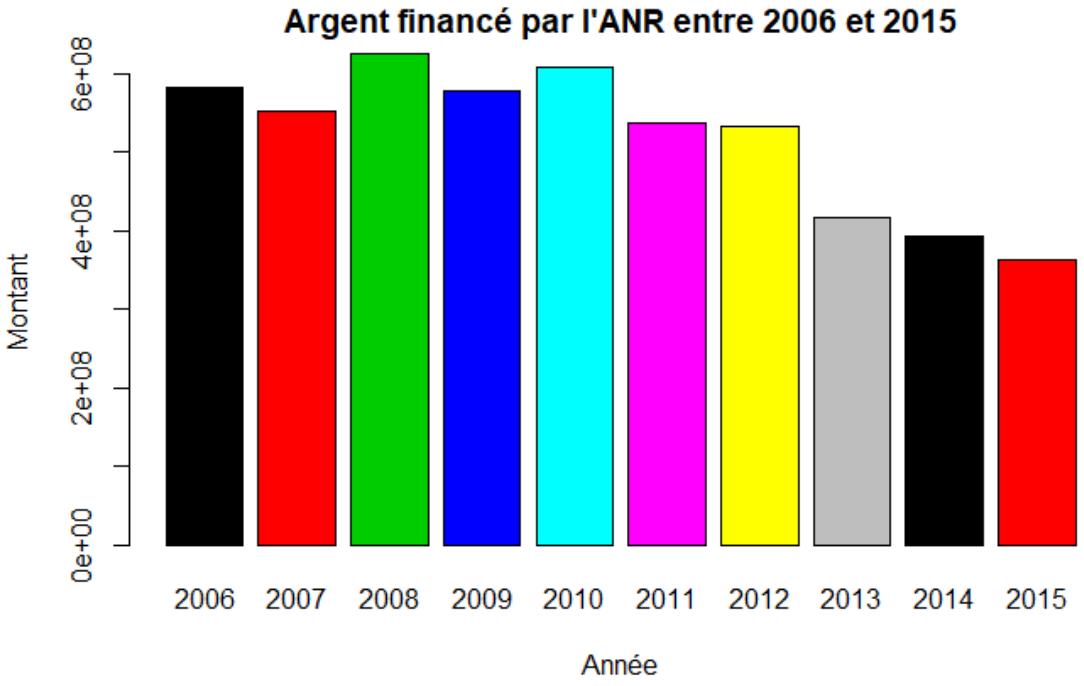

FIGURE 9 – Argent financé par l'ANR entre 2006 et 2015

On voit que l'argent financé par l'ANR décroît au fil des année. Entre 2006 et 2012. L'ANR fournissait environ 6 millions d'euros par an aux entreprises de recherche. Après 2012, ce montant est en nette baisse. En 2015, l'ANR fournira 4 millions d'euros. On aurait tendance à tirer des conclusions comme le fait que l'ANR, pour un projet, fournirait de moins en moins d'argent par simple précaution, or nous devons nous pencher sur le graphique suivant qui sera plus indicateur.

6.2 Nombre de projet par année

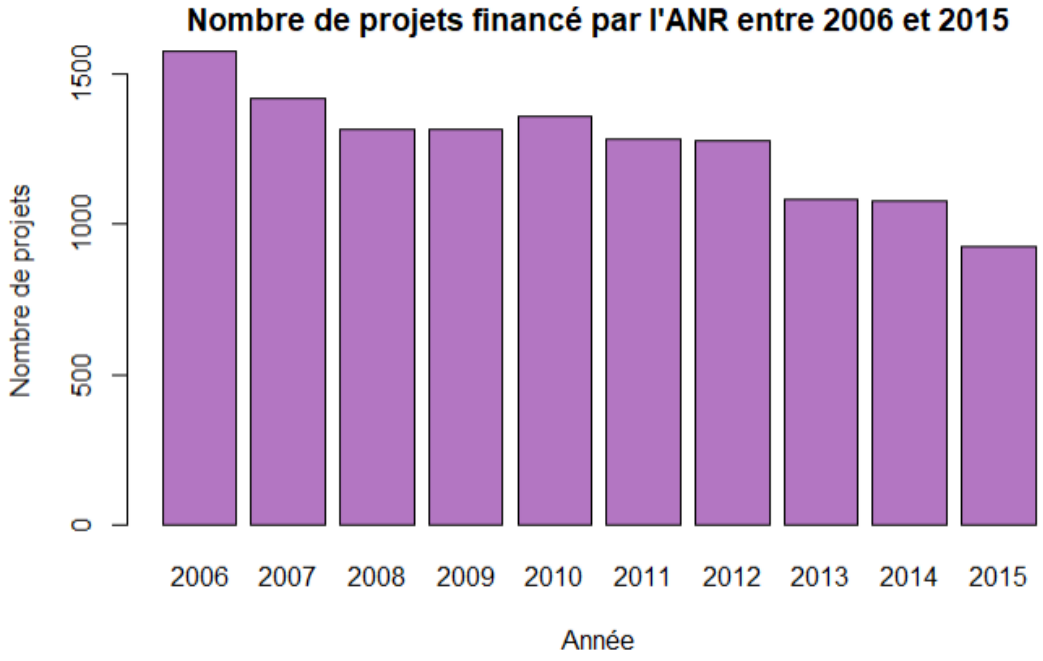

FIGURE 10 – Nombre de projets financés par l'ANR entre 2006 et 2015

Avec ce diagramme en barre, on voit que le nombre de projets financés décroît au fil des années. Ainsi, nous pouvons faire un lien avec le diagramme précédent, qui est que le montant financé par l'ANR est en baisse a cause du nombre de projets retenus qui est en nette baisse.

6.3 Montant par ville

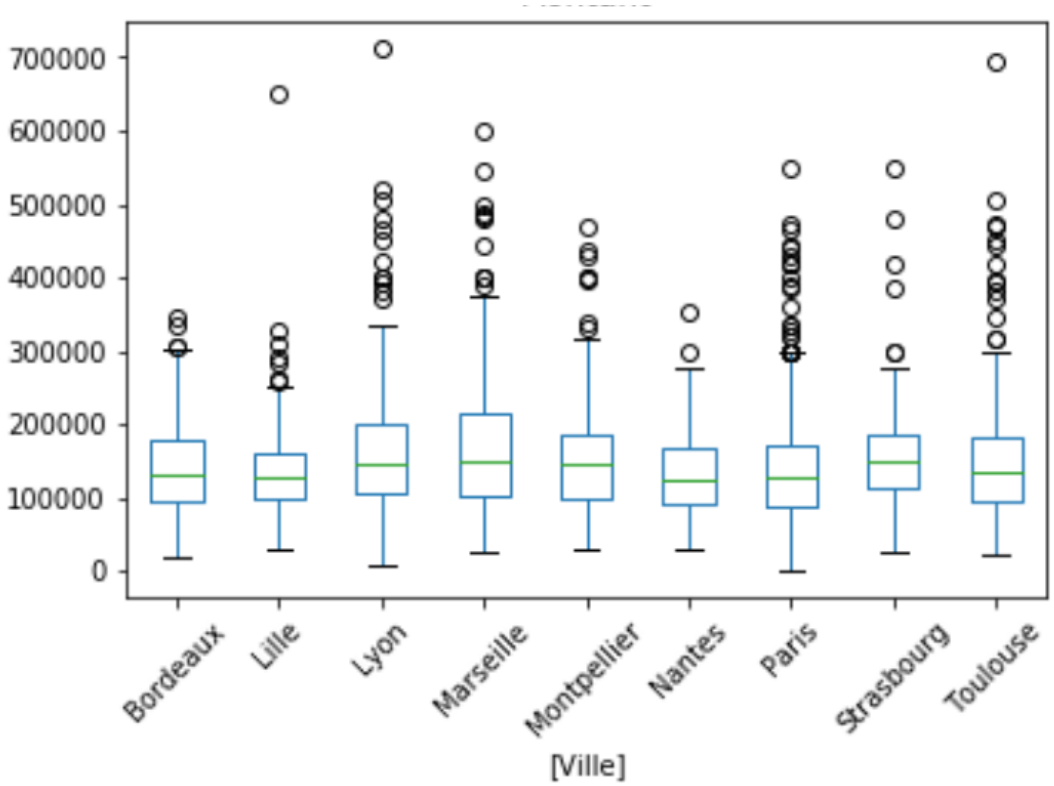

FIGURE 11 – Boxplot Montant par ville

En se basant sur nos données, qui bien sur ne sont qu'un échantillon, on peut voir que les projets qui ont reçu le plus grand financement, viennent de Lille, Lyon et Toulouse. D'autre part, Lyon, Marseille, Paris et Toulouse reçoivent un montant important sur certains projets.

6.4 Domaine et villes

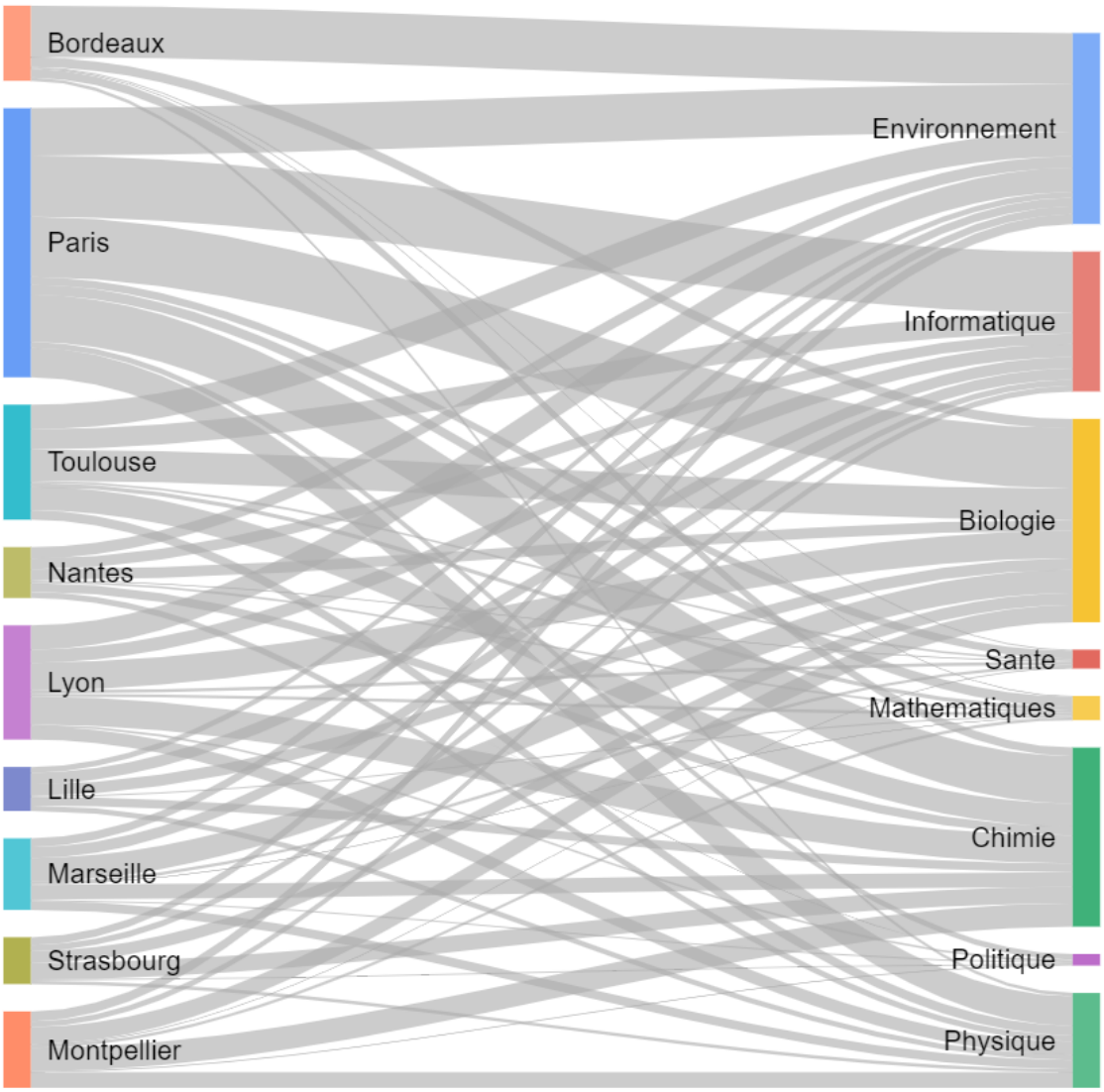

FIGURE 12 – Sankey Diagram

A l'aide de la classification des mots faites en page 16, nous avons le diagramme suivant

Ce diagramme en sankey nous montre que Paris est fortement présent dans tout les domaines. Bordeaux semble s'intéresser à l'environnement, sûrement grâce aux recherches via la biologie. La proximité de Bordeaux avec l'océan peut expliquer cette tendance. On rappelle que le mot "océan" faisait partie des mot retenus en page 16 et qu'il était relié a l'environnement.