

Représentations vectorielles de corpus collaboratifs sur la ville de Paris

Carmen Brando (EHESS), Catherine Dominguès (IGN)

Journées Analyse de données textuelles en conjonction avec EDA 2017
Lyon 3-5 mai

Contexte et objectifs

- ❖ Activités de gouvernance collaborative (ou démocratie participative) mise en place par les administrations publiques (villes, régions, ...)
- ❖ Analyser/valorisation des opinions/suggestions/attentes des usagers/citoyens
- ❖ Moyens techniques pour recueillir ces corpus

Contexte : Mairie de Paris

- ❖ Corpus **web** composé des **propositions** de Parisiens concernant des projets d'**aménagement urbain**
- ❖ Plateforme web *Madame la maire, j'ai une idée* : <https://idee.paris.fr/>
- ❖ Quatre thèmes proposés par la mairie de Paris :
 - **Reconquête des berges**, corpus : Berges
 - **Réinventons nos places**, corpus : Places
 - Italie, Bastille, Nation, Gambetta, Panthéon, Madeleine, Fêtes
 - **Construisons la métropole du Grand Paris**, corpus : Métropole
 - **Budget participatif**, corpus : Budget



CO-CONSTRUISONS
PARIS

VOS IDÉES

LES ATELIERS DE CO-
CONSTRUCTION

OÙ PARTICIPER ?



Rechercher



En discussion



Gambetta



Identité, patrimoine et paysage

Objectif de l'idée

Vegetaliser

Descriptif de l'idée proposée

Rendre plus verte et agréable pour les piétons la place Gambetta. Créer des îlots de verdure tout autour en rognant sur la route. Mettre des bancs au pieds des arbres/ arbustes. Permettre des terrasses plus grandes dans les rues autour de la place et Vegetaliser également ces rues et fleurir également!



Proposé par **Cleros6**

02.09.15

Rendre plus verte et agréable pour les piétons la place Gambetta. Créer des îlots de verdure tout autour en rognant sur la route. Mettre des bancs au pieds des arbres/arbustes. Permettre des terrasses plus grandes dans les rues autour de la place et Vegetaliser également ces rues et fleurir également!



AMÉNAGER LES BOULEVARDS DES MARECHAUX



En discussion



Culture et Identité



Métropole du Grand Paris

Objectif de l'idée

Effacer les frontières entre Paris et les villes limitrophes

Descriptif de l'idée proposée

Les boulevards des maréchaux marquent une frontière entre Paris et les villes de la banlieue proche et ne constituent actuellement que des axes de circulation périphérique à Paris. Les trottoirs et contre-allées qui les bordent sont larges mais ne sont absolument pas exploités, ni valorisés. Ils pourraient surement être mieux aménagés en y développant des axes piétonniers végétalisés (promenades vertes), des voies cyclistes, des zones de commerces ou lieux artistiques et culturels, afin d'en faire de véritables lieux de vie unissant et bénéficiant aux arrondissements périphériques de Paris ainsi qu' aux villes limitrophes.

Diagnostic / Inspiration / Exemples d'expérimentation passée

Les trottoirs et contre allées des boulevards des maréchaux, bien que larges, ne sont pas aménagés.

Les **boulevards des maréchaux** marquent une **frontière** entre **Paris** et les **villes de la banlieue proche** et ne constituent actuellement que des **axes de circulation périphérique** à Paris. Les **trottoirs** et **contre-allées** qui les bordent sont **larges** mais ne sont absolument pas **exploités**, ni **valorisés**. Ils pourraient sûrement être mieux **aménagés** en y développant des **axes piétonniers végétalisés** (promenades vertes), des **voies cyclistes**, des **zones de commerces** ou **lieux artistiques et culturels**, afin d'en faire de véritables **lieux de vie** unissant et bénéficiant aux **arrondissements périphériques** de **Paris** ainsi qu'aux **villes limitrophes**.

Noms de lieux

- ❖ Importance numérique des noms de lieux dans les corpus
- ❖ Définition fondée sur la définition d'ENS de (Ehrmann, 2008)
1 expression linguistique <-> 1 lieu géographique
- ❖ Définition élargie afin de prendre en compte :
 - ~~1 expression linguistique <-> 1 lieu géographique~~
 - expressions linguistiques <-> entité localisée
 - typologie fondée sur les noms propres/descriptions définies (Leroy 2004; Ehrman 2008) et sur les notions de noms génériques (Edwardes et al. 2007; Derungs et al. 2013; Nouvel et al. 2016)
 - subjectivité peut être exprimée dans les noms de lieux

Noms de lieux (exemples)

❖ “Nom propre” : *Champs-Élysées, Italie 2*

❖ Localisation génériques :

➤ nom “géographique” : *rivière, boulevard*

➤ nom “géographique” + “nom propre” : *rue de Lappe, boulevard Malesherbes, centre commercial Italie 2*

➤ entité localisée : *centrale solaire, banc, trottoir, escalier mécanique, mur anti-bruit, place de stationnement, marches de l’Opéra*

➤ localisations vagues :

un point névralgique de la métropole, lieu de création, en bas de mon immeuble

Objectifs et méthode

❖ Objectifs :

- Confirmation des concepts (mots) importants
- Construction du contexte sémantique de ces concepts importants
- Comparaison des contextes sémantiques des concepts dans les thématiques proposées par la ville
- Identification des questions communs et divergentes à tous les corpus

❖ Proposition de méthode :

- Recherche des mots (dont les noms de lieux) fréquents
- Construction du contexte sémantique de ces mots par calcul de leurs vecteurs de similarité
- Interprétation et comparaison entre corpus de ces vecteurs de similarité
- Construction des réseaux sémantiques de ces mots, par corpus (recherche de relation et typage de ces relations avec WordNet/BabelNet)

Recherche des mots fréquents dans les corpus

Description détaillée des quatre corpus

Corpus	# lemmes	# verbes	# adjectifs	# noms	# nr
Berges	18 384	1 691	991	3 834	203
Places	69 371	6 847	4 302	15 500	1 565
Métropole	59 788	5 961	4 100	13 665	1 952
Budget	991 881	97 172	69 801	249 247	31 419

Description détaillée des quatre corpus (2)

Corpus	# lemmes	# verbes	# adjectifs	# noms	# nr
Berges	18 384	1 691	991	3 834	203
Places	69 371	6 847	4 302	15 500	1 565
Métropole	59 788	5 961	4 100	13 665	1 952
Budget	991 881	97 172	69 801	249 247	31 419

+ 3%

Description détaillée des quatre corpus (3)

Corpus	# lemmes	# verbes	# adjectifs	# noms	# nr	# génériques
Berges	18 384	1 691	991	3 834	203	685
Places	69 371	6 847	4 302	15 500	1 565	3 377
Métropole	59 788	5 961	4 100	13 665	1 952	1 351
Budget	991 881	97 172	69 801	249 247	31 419	29 430

Modèle CRF entraîné sur nos corpus annotés (implémentation Stanford NER) : meilleur modèle pour annoter : 1-3 Berges + 1-3 Places (Brando et al 2016)

Description détaillée des quatre corpus (3)

Corpus	# lemmes	# verbes	# adjectifs	# noms	# nr	# génériques
Berges	18 384	1 691	991	3 834	203	685
Places	69 371	6 847	4 302	15 500	1 565	3 377
Métropole	59 788	5 961	4 100	13 665	1 952	1 351
Budget	991 881	97 172	69 801	249 247	31 419	29 430

**un nom sur 5 est un nom
de lieu générique**

Fréquence et répartition des mots

Regroupement de mots

espace, quartier, ville, centre, banlieue, parisien, métropole, GrandParis

vert, végétaliser, jardin

transport, circulation, voiture, vélo, bus, cyclable, piéton, covoiturage, métro, promenade

pollution, bruit, sonore, nuisance

école, scolaire, périscolaire

expérimenter, rendre, permettre, réduire

berge, Seine, quai

Construction du contexte
sémantique de ces mots par
calcul de leurs vecteurs de
similarité

Word embedding

- ❖ Technique qui permet de représenter les mots d'un dictionnaire, pour un corpus donné (en général de grande taille), par des vecteurs afin de **faciliter leur analyse sémantique et syntaxique**. Ainsi, **chaque mot sera représenté par un vecteur de réels** et les **mots apparaissant dans des contextes similaires auront des vecteurs plus proches que d'autres apparaissant dans des contextes différents**
- ❖ **Word2vec*** (Mikolov et al 2013) : implémentation d'un réseaux de neurones artificiels (à deux couches) pour construire les vecteurs de mots
 - **CBOW** utilise le contexte pour prédire le mot cible
 - **Skip-Gram** utilise un mot pour prédire le contexte cible

Word2vec (2)

- ❖ Détection d'analogies (opération de vecteurs), ex : $\text{vec}(\text{'Paris'}) - \text{vec}(\text{'France'}) + \text{vec}(\text{'Italy'}) = \text{vec}(\text{'Rome'})$
- ❖ Les représentations vectorielles sont employées dans plusieurs tâches de traitement automatique des langues et d'analyse textuelle comme :
 - caractérisation de corpus (Grayson et al 2016)
 - traduction automatique (Servan et al 2016)
 - extraction de relations temporelles (Tourille et al 2016)
 - constitution et enrichissement d'un lexique pour un domaine spécialisé (Muchemi et Grefenstette 2016)

Point de départ : mots d'intérêt

Regroupement de mots

espace, quartier, ville, centre, banlieue, parisien, métropole, GrandParis

vert, végétaliser, jardin

transport, circulation, voiture, vélo, bus, cyclable, piéton, covoiturage, métro, promenade

pollution, bruit, sonore, nuisance

école, scolaire, périscolaire

expérimenter, rendre, permettre, réduire

berge, Seine, quai

Mot :
 piéton
 l'écart de
 répartition
 est
 essentielle
 ment dû à
 un sur-
 emploi
 dans
 "Places"

Budget	Places	Métropole	Berges
cycliste_nom circulation_nom automobiliste_nom dangereux_adj traverser_ver trafic_nom voiture_nom _trottoirs__nom Risque_nom ...	circulation_nom Bastille_nam espace_nom _place__nom accès_nom Italie_nam place_nom réduire_ver _Place_de_la_Ré publique__nom ...	accès_nom permettre_ver local_nom partie_nom parisien_adj poubelle_nom énergie_nom art_nom faciliter_ver Réseau_nom ...	avoir_ver:cond accessible_adj hiver_nom ouvrir_ver participer_ver chose_nom piscine_nom cycliste_nom Impossible_adj ...

Mot : banlieue (gen.)

L'écart de répartition est essentiellement dû à un sur-emploi dans le corpus "Métropole"

Budget	Places	Métropole	Berges
périphérie_nom train_nom conserver_ver:ppre _lieux_publics__nom extérieur_nom retard_nom Allemagne_nam permanence_nom Angleterre_nam pierre_nom continu_nom reconstitution_nom _espaces_verts__nom ...	sale_adj aller_ver:infi utiliser_ver:pper jour_nom rendre_ver passer_ver architecte_nom lieu_nom haut_adj large_adj important_adj laisser_ver ...	vélo_nom stationnement_nom petit_nom place_nom émission_nom projet_nom instar_nom échelle_nom deux_num consultation_nom moyen_nom groupe_nom ...	-

Mot : transpo

rt
L'écart de répartition est essentiellement dû à un sur-emploi dans le corpus "Métropole" et dans une moindre mesure dans "Berges"

Budget	Places	Métropole	Berges
commun_nom marchandise_nom déplacement_nom rapide_adj mode_nom gratuité_nom hybride_adj taxi_nom fluvial_adj électrique_adj covoiturage_nom ..	faciliter_ver sol_nom mode_nom face_nom vitesse_nom déplacement_nom permettre_ver RER_abr _centre__nom permettre_ver _passages_piétons__nom ...	commun_nom mise_nom vélo_nom pied_nom itinéraire_nom libre_adj particulier_nom meilleur_adj immeuble_nom situer_ver Émergence_nom ...	développement_nom soir_nom _tunnels__nom péage_nom nombre_nom site_nom culturel_adj _capitale__nom permettre_ver réduire_ver: _lieu__nom _places__nom ...

Visualisation de données de très haute dimensionnalité

- ❖ L'algorithme **T-SNE** (van der Maaten et Hinton 2008) est utilisé pour **réduire** les données ayant une **très haute dimensionnalité** à deux ou trois dimensions à fins de visualisation
- ❖ Une fonction mesure les similarités et les différences parmi les objets de grande dimensions et projette ces objets sur un espace 2D ou 3D en préservant la **structure locale des données**
- ❖ Des objets similaires seront représentés par des **points proches** dans l'espace et des objets différents seront projetés par des **points éloignés**

Métropole

Berges

Places

Discussion & conclusion

- ❖ Vrai intérêt des représentations vectorielles pour l'analyse sémantique d'un corpus textuel
- ❖ Les limites de l'approche vis-à-vis de :
 - la taille des corpus,
 - la paramétrisation “volatile” du modèle,
 - l'influence des variations orthographiques (orthographe très variable, mots étrangers, ...),
 - la prise en compte de mots composés (ex : place de stationnement),
 - la création de méthodes de validation
- ❖ Possibilité de combiner des modèles de vecteurs de mots (Bernier-Colborne et Drouin 2016) existants pour un domaine similaire (taille vs. domaine)

Perspectives

- ❖ Construction des réseaux sémantiques des mots fréquents des corpus : typage de relations pertinentes (à partir de WordNet/BabelNet)
- ❖ Recherche des jugements/opinions sur les lieux (souvent portés par les modificateurs) : *Grand Paris écologique, Ile-de-France rurale, capitale européenne du klaxon*
- ❖ détection de comparaisons entre les lieux (Mpouli 2016)
- ❖ Prise en compte (et discussion) des métadonnées définies par la Mairie de Paris
- ❖ Comparer, dans le contexte de différentes villes, les thèmes proposés communs

Merci

carmen.brando@ehess.fr
catherine.domingues@ign.fr