Corpus
Web comme corpus

Béatrice Daille

Faculté des Sciences et Techniques de Nantes
Département Informatique

2018-2019

Corpus

- Corpus : exemples, définition, caractérisation
- Métadonnées : Dublin Core, TEI
- **Web comme corpus**
- Annotation des corpus
- Exploration des corpus : concordance, collocations

Web comme corpus

1. corpus construits à partir du web
2. construire un corpus à partir du web
    1. structure du web
    2. moissonnage
    3. post-traitements
    4. évaluation du corpus

Corpus du web

- WaCky initiative (Baroni et al. 2009)
  - ukWac : 2 billion mots, .uk domaine, mots-graines : mots de fréquence moyenne du BNC
  - frWac : 1.6 billion mots, .fr domaine, mots-graines : mots de fréquence moyenne du Monde diplomatique + listes de vocabulaire de base du français
  - deWac : 1.7 billion mots, .de domaine, mots-graines : mots de fréquence moyenne du SudDeutsche Zeitung + listes de vocabulaire de base de l'allemand
  - sdeWac : 0.88 billion mots
  - itWac : 2 billion mots
- GloWbE : Global Web-based English, 1.9 billion
- Corpus wikipedia

Collecter les données

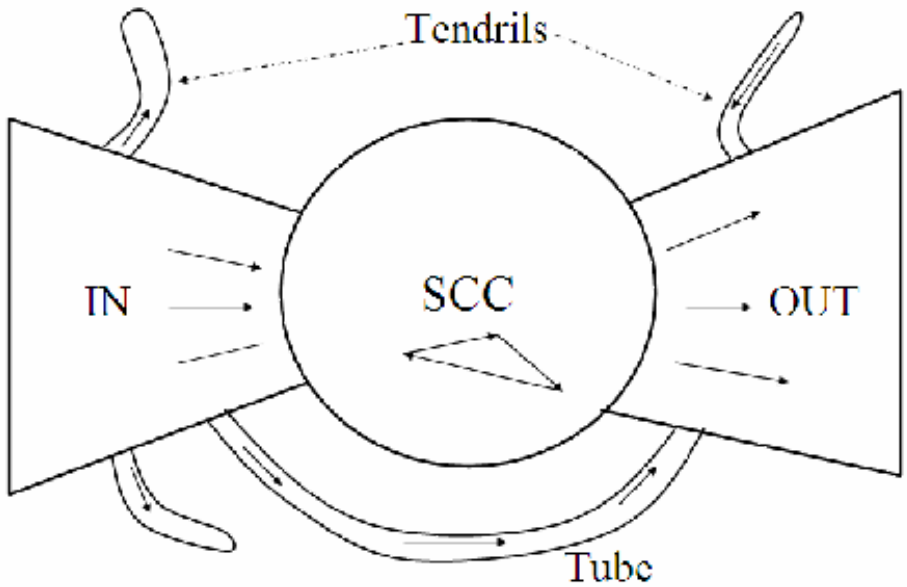
- Structure du web : liens, domaines
- Moissonnage du web

Liens

- Le web est constitué de pages et de liens les liant. Un lien est une URL (Uniform Resource Locator)
- Chaque page possède :
  - Ensemble de liens entrants $I(p)$, degré entrant : $ID(p) =\mid I(p) \mid$
  - Ensemble de liens sortants $O(p)$, degré sortant : $OD(p) =\mid O(p) \mid$
- le nombre de liens entrants est distribué selon une loi de puissance (réseau sans échelle) :
  $P(i) = \frac{1}{i^{2,1}} = i^{-2,1}$ avec i=degré entrant (Manning et al., 2009, 426.)
- Signification : peu de pages avec un degré entrant important et beaucoup de pages avec un degré entrant faible
- Pages avec un degré entrant important sont faciles à trouver, mais ne sont pas forcément les plus intéressantes pour construire un corpus.

La structure du web

Structure en nœud papillon du web (Broder et al. 2000) - Composants de même taille.



- IN $ID(p) = 0$ et $OD(p) > 0$
- OUT $ID(p) > 0$ et $OD(p) = 0$
- SCC $ID(p) > 0$ et $OD(p) > 0$
  Taille en 2014 WDC Hyperlink Graph : IN 49,61 %, SCC 18,95 %, OUT 1,59 %, TENDRIL et TUBE 21,16 %, Non connecté 8,69 %

Accès aux pages

- navigation nécessite une liste d'URL graines pour débuter le moissonnage
- pages inaccessibles (web profond)
  - $ID(p) = 0$
  - $ID(p) > 0$ mais trop éloignées des URL graines
  - pages interdites (interdiction explicites aux robots, leurre)
  - pages nécessitant un login (forum, portail documentaire, etc.)

Contenu statique et dynamique du web

- **statique** : page éditée manuellement
- **dynamique** : page générée à partir d'une base de données

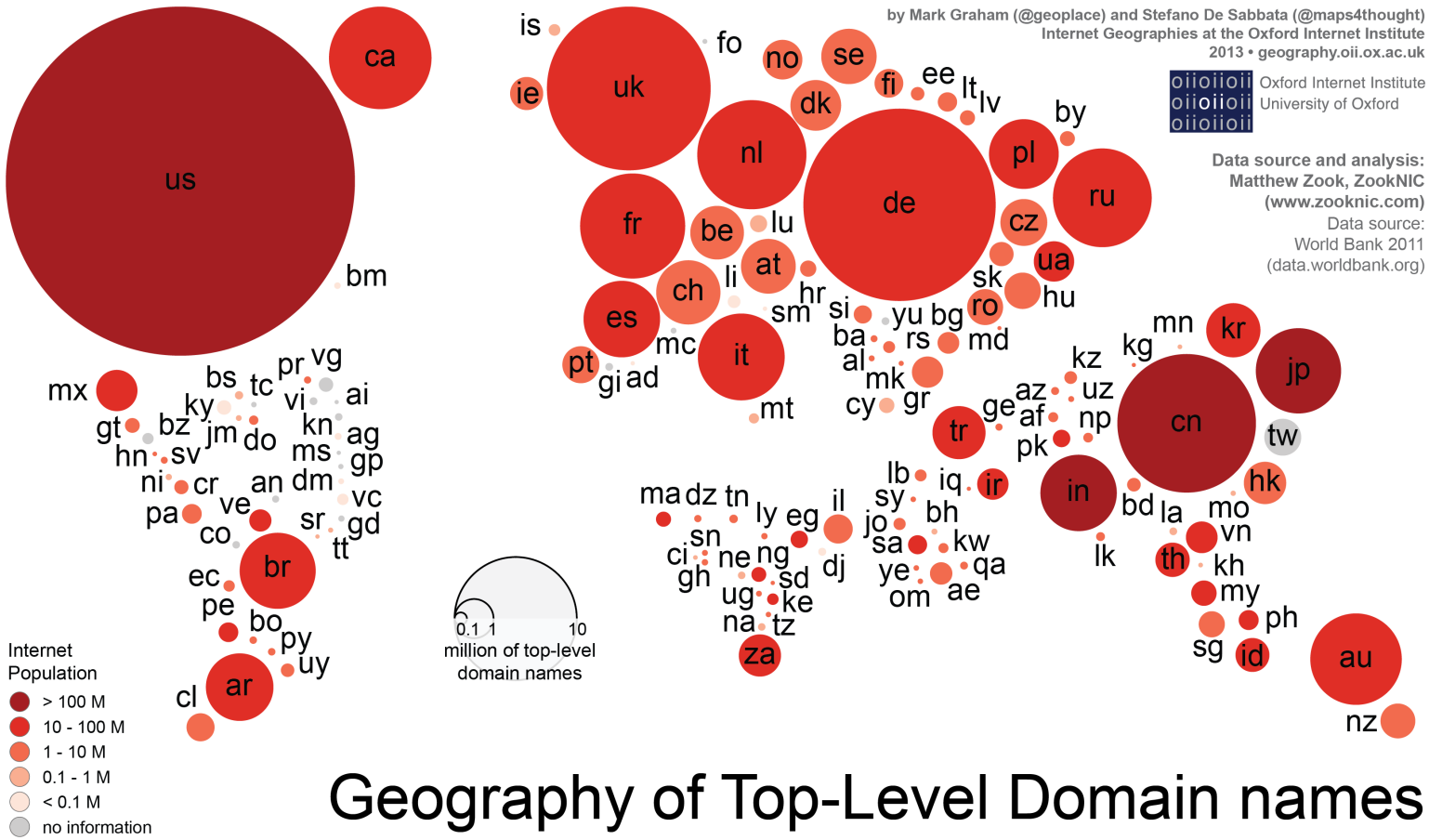
Majorité des pages sont dynamiques
Distinction statique/dynamique non intéressante pour la construction du corpus

Domaines de premier niveau et langues

10 niveaux de granularité pour caractériser le web :
bit < caractère < mot < bloc < page < sous-site < site < domaine < domaine de premier niveau < domaine national < web global
(Bjorneborn et Ingwersen 2004)

- caractère : identification d'une langue, distribution des caractères, n-grammes de caractères dans les documents
- mot : identification d'une langue, d'un registre, d'un domaine, etc.
- bloc (paragraphe) : identification du genre, de la structure du document
- domaine de premier niveau et domaine national : identification de textes d'une langue

Domaines de premier niveau et domaines nationaux



Geography of Top-Level Domain names

Moissonnage à l'aide du domaine de premier niveau

- Anglais : première langue de nombreux domaines nationaux
  afrique 75 %, espagne : 30 %, chine : 8 %
- .us : pas fiable. la majorité des entreprises n'utilisent pas le .us
- .ca et .uk : des variantes nationales de l'anglais
- .es : beaucoup de problèmes, catalan, castillan, galicien
  El Pais elpais.com, El mundo elmundo.es

Liens problématiques

- Taille infinie des sites dynamiques : impossibilité de mesurer une profondeur de lien.
- Ferme de liens créée artificiellement pour améliorer la réputation d'un site

Moissonnage à l'aide de moteur de recherche : Bootcat (Baroni et Bernadini 2004)

1. construction d'un lexique constitué manuellement ou à partir d'une liste de fréquence calculée sur un corpus
2. sélection d'une liste de mots-graines au sein de ce lexique possédant une fréquence moyenne
3. permutation des mots-graines pour créer des n-uplets aléatoires et uniques (disjonction et coordination, variation des fréquences de mots-graines) : les n-uplets graines
   habituellement triplets ou quadruplets
4. envoi des n-uplets graines comme requêtes à un moteur de recherche et récupération des URL
5. suppression des URL dupliquées
6. récupération des documents

Avantages de BootCat

- simple
- rapide
- peut être utilisé pour construire des corpus en langue de spécialité
- outil avec interface graphique http ://bootcat.dipintra.it/

Inconvénients de BootCat

- construction de petits corpus
- biais de l'ordonnancement renvoyé par le moteur de recherche : tous les documents ne sont pas renvoyés
- biais des requêtes : combinaisons très spécifiques. Pire des cas : lexiques ou dictionnaires
- arrêt de l'API Bing, impossibilité de lancer de nombreuses requêtes gratuitement
- Pour mémoire :
  - Google et Yahoo ont arrêté leur API
  - impossibilité d'utiliser CURL ou WGET car vous allez être repérés très rapidement

La commande WGET

wget [option]... [URL]...
`http://www.delafond.org/traducmanfr/man/man1/wget.1.html`

- -r recursion
- -l –level=profondeur (-l2 ou -l3)
- -D liste-domaines –domains=liste-domaines (-D .us, .uk)

Exemples de résultats obtenus avec BooTCat

10 000 4-uplets construits à partir de 5 000 mots extraits de la base lexicale CELEX ayant une fréquence moyenne calculée dans un corpus de référence
Résultats sur le hollandais en 2012 :

- 127 910 URL moissonnées
- 70 897 uniques (55,43 %)
- URLs apparaissant aux premiers rang sont des lexiques

Constats similaires : De 37 %, Es : 21 % Se : 16 %

Exemples de 4-uplets pour l'anglais

glorious discretion virtually unhappy
circles texts ingredients procurement
bothered eastern sponsorship monitored
attracted muslim part-time bars
enhance keys continued report
settings watch floors briefly
lucky seating fear sleeping
expectation participate please sectors
publication moves latter biological
height capable percent tricky

Exemples d'URL moissonnée

glorious discretion virtually unhappy

Dancing in the Glory of Monsters Jason Stearns[1]
Dancing in the Glory of Monsters Jason Stearns[1] - Free ebook download as ... and could
legislate by decree and change the constitution at his discretion ... a fanciful spiritual
order that sold banking licenses in the name of a virtual state ... was hospitalized in South
Africacand was obviously unhappy with the question.
www.scribd.com/... /Dancing-in-the-Glory-of-... Cachad

Exercise

Tester

1. la méthode bootcat sur le français en utilisant les mots de fréquence moyenne occurrant dans Jules Verne
2. la commande wget sur le domaine français

Moissonneur (crawler)

composants logiciels typiques :

1. **Récupérateur** télécharge les documents d'URLs listées dans un agenda
2. **Analyseur** récupère les URLs apparaissant sur la page web
3. **Filtreur d'URLs** élimination des URLs en double ou qui ne répondent pas à certains critères
4. **agenda** une structure de données qui stocke et ordonne les URLs et les communiquent au récupérateur

Avantages du Moissonneur

- corpus de taille illimitée
- ne dépend pas des moteurs de recherche
- permet de télécharger des textes entiers
- répond aux normes de courtoisie de la recherche sur internet

Domaine public

- Heritrix `https://github.com/internetarchive/heritrix3`
- Nutch `http://nutch.apache.org/` (cluster)

Désavantages du Moissonneur

- requière un ensemble d'URLs graines
- nécessite beaucoup de ressources, très lent
- configuration délicate
- difficile à contrôler

Nettoyage d'un corpus construit par moissonnage standard

Nombre de documents supprimés lors de la phase de nettoyage de DECOW 2012, un corpus de 9 M de mots en allemand crawlés sur le domaine .DE

Supression	Nb de documents	Pourcentage
petits textes	93 604 922	71,6 %
documents non textuels	16 882 377	12,9 %
documents identiques	3 179 884	2,43 %
documents presque identiques	9 175 335	7,03 %
Total	122 842 518	94,06 %

Préfiltrage des documents

Avant l'enregistrement de l'URL dans l'agenda

- restriction à un domaine ou à une liste de sites
- liste d'anti-sites (serveurs publicitaires) ou anti-URL
- URL contenant des mots-clés caractéristiques de sites non intéressants
- mauvais formats de fichiers (.avi, .flv, .pdf, etc...)
- normaliser les URL

Liste d'URL graines

- Combien d'URL sont nécessaires ?
- Comment évaluer la qualité d'une URL ?
- Doivent-elles être diversifiées ?

Constitution

- liste manuelle
- utilisation de Bootcat

Constitution d'une liste d'URL graines

Autre solution : Wikipedia

- Disponible pour de nombreuses langues
  - envoi de requêtes aléatoires
    `http://en.wikipedia.org/wiki/Special:Random`
  - Extraire tous les liens pointant sur des sites hors wikipedia et les utiliser comme URL graines
- avantages : pas de restriction, nombreux sujets
- désavantages : peu de liens sortants, des documents .pdf, de nombreuses bibliographies

Autres solutions

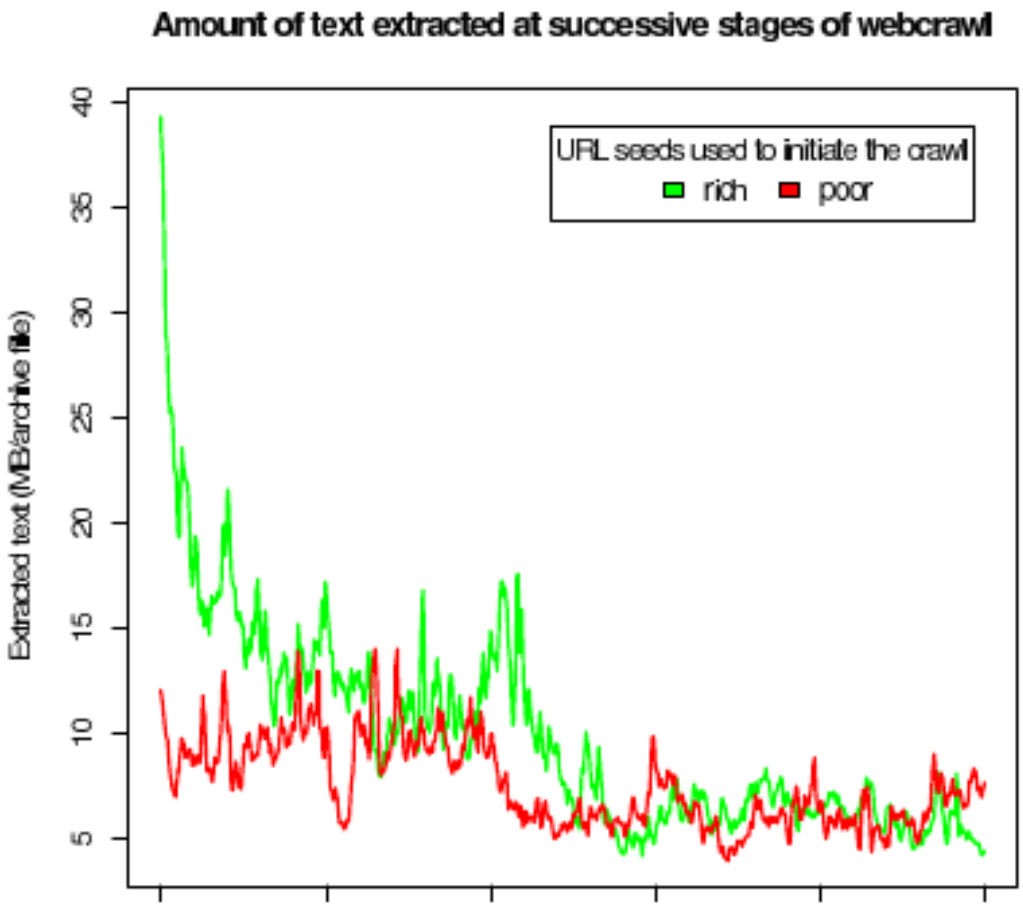
Trouver des listes d'URLs graines

- les sites les plus fréquentés : `http://www.alexa.com/topsites`
- les sites par pays et thématiques : `http://www.curlie.org/`

Récoltes des moissonages avec une petite ou une longue liste d'URL graines

Une petite liste d'URL graines n'implique pas une petite récolte de textes

poor : 10 URL graines rich : 10 000 URL graines



Estimer la couverture du moissonnage

Trouver les documents qui maximisent la mesure de couverture *WC* (Weight Coverage) d'un moissonnage à un temps t, avec :
$C(t)$: le nombre de pages moissonnées en t
w : une fonction qui estime l'adéquation de chaque page à la finalité du crawl
$WC = \sum_{p \in C(t)} w(p)$

PageRank

Introduit par Brin and Page (1995) Google
Voir Manning et al. (2009)
Idée : on mesure un dégré

- un degré de PageRank d'une page est haut si elle est liée à de nombreuses pages avec un dégré haut ;
- les pages avec quelques liens sortants contribuent à augmenter le degré de PageRank des pages référencées

PageRank

- Le degré de PageRank R d'une page p est la probalilité d'aller sur p lors d'une page de marche aléaloire sur le graphe du web
- Pour chaque page p, le marcheur suit un lien sortant à partir de p avec une probalilité fixé à $1 - d$, ou saute sur une page aléatoire avec une probalité d
- les valeurs considéréespour d soont entre 0,1à et 0,15
- le saut aléatoire évite l'accumumation de PageRank avec des pages (du OUT) sans redistribution de PageRank

PageRank

PageRank P d'une page p avec des probalité de saut fixe d où N est le total de pages web et $p_{1...k}$ les pages qui sont liées à p et $C(p)$ le nombre de liens sortants de p_n :
$R(p) = \frac{d}{n} + (1-d)\sum_{i=1}^{k} \frac{R(p_i)}{C(p_i)}$

Crawler à base de marche aléatoire

Post-traitements

Une page web

LE FIGARO.fr SANTÉ

NEWS | ENCYCLOPÉDIE SANTÉ | MIEUX-ÊTRE | SOCIAL | VOYAGES | COACHING

MON PROFIL SANTÉ | ANNUAIRE PRO | GUIDE DES MÉDICAMENTS | LE FIGARO.fr | Newsletter | Rechercher

Accueil > Actualité

Article précédent | TT | Envoyer | J'aime 15 | CONSERVER

La «fish pedicure» n'est pas sans risque

Par Delphine Chayet - le 25/04/2013

L'Agence nationale de sécurité sanitaire demande un encadrement de cette pratique à visée esthétique qui consiste à immerger ses pieds dans un bocal rempli de poissons.

Se laisser grignoter les peaux mortes des pieds par des petits poissons n'est pas dénué de risque, selon l'**Agence nationale de sécurité sanitaire** (Anses) qui recommande, dans un avis dévoilé ce jeudi, «un encadrement strict de cette pratique».

La «fish pedicure» est apparue en France en 2010.

Apparue en France en 2010, la «fish pedicure» n'est aujourd'hui soumise à aucune règle sanitaire spécifique. De plus en plus de curieux se laissent tenter par cette **expérience** de massage exfoliant et indolore. Selon l'Anses, plusieurs centaines d'instituts de beauté seraient équipés de ces grands bacs contenant une centaine de Garra rufa - des poissons sans dents, mais très gourmands en squames, mesurant

Webpage cleaning, *boilerplate removal* ou détourage

Distinguer le **contenu informatif** :

- en bleu dans le diagramme précédent, les segments faisant partie du contenu informatif : titre, chapeau, sous-titre et paragraphes ;
- en orange, les segments potentiellement intéressants : informations sur l'auteur, date de l'article et légende de la photographie.

Du **contenu non-informatif** (ou très faiblement informatif) :

- menus de navigation (en haut) ;
- articles connexes (en bas) ;
- image (à droite) ;
- publicité (en bas à droite).

En résumé pas d'opposition binaire mais une sorte de **continuum**, on doit donc formuler des hypothèses adaptées.

Un exemple d'attendu

Étiquette	Contenu
<h>	La fish pédicure n'est pas sans risque
<auteur>	Par Delphine Chayet
<date>	25/04/2013
<legende>	La fish pédicure est apparue en France en 2010.
<p>	L'Agence nationale de sécurité sanitaire demande un encadrement [...]
<p>	Se laisser grignoter les peaux mortes des pieds par des petits poissons [...]
<p>	Apparue en France en 2010, la fish pédicure n'est aujourd'hui [...]
<h>	Poissons d'élevage
<p>	Même si aucun cas documenté n'a pour l'instant été rapporté, [...]
<p>	Dans une eau qui ne peut par définition être désinfectée, [...]
<p>	Elle recommande aussi une information objective du public [...]

Définir la fonction de détourage : deux sous-tâches

- Le nettoyage :
  - Du code (*javascript*, feuille de style...) ;
  - Du squelette de page (menus, liens, entêtes et pieds de page...).
- L'annotation maîtrisée de la structure :
  - Titres (<h>) ;
  - Paragraphes (<p>) ;
  - Listes (
,) ;
  - Autres éléments ?

Proportion de blocs dans le **corpus de référence** *Cleaneval* :

- titre, sous-titre(s), chapeau et corps de texte (13 %) ;
- autres segments de l'article, par exemple les légendes (3 %) ;
- commentaires des lecteurs (1 %) ;
- contenu connexe, par exemple liens vers d'autres articles (4 %) ;

Leur conclusion : une tâche peu *sexy* mais capitale.

Traitements des textes

Nettoyage

- caractères et entités &gt ; ;
- codage des caractères ;
- suppression de la césure.

Identification de la langue

Supression des documents en double

- duplication totale
- duplication partielle (plagiat, citation)
- vues multiples du document
- duplication au sein d'un document

Les principaux indices disponibles

- Le site Web ;
  - utiliser les caractéristiques de différentes pages du même site ;
- l'image de page ;
  - observer le rendu de la page tel qu'il est donné par un ou plusieurs navigateurs ;
- le code HTML lui-même ;
  - exploiter les informations de hiérarchie entre les blocs ;
- le contenu des blocs HTML ;
  - détecter des candidats à partir de phrases, mots ou caractères ≪ attendus ≫.

Exemples d'outils de nettoyage de page HTML

- BoilerPipe `https://boilerpipe-web.appspot.com/`
- Readability
- NCleaner et autres participants de CleanEval
- Html2Text...

Exemples d'outils de nettoyage de documents pdf

- GROBID, PDF2XML, PDFALTO, PDFMINER, PYMUPDF, ...
- pour les articles scientifiques : 1) pdf2txt 2) PARSCIT (LREC 2008)
- pour les autres documents : PDFMINER : ajouts de balises similaires au HTML (p, h1, h2, Div, Figure)

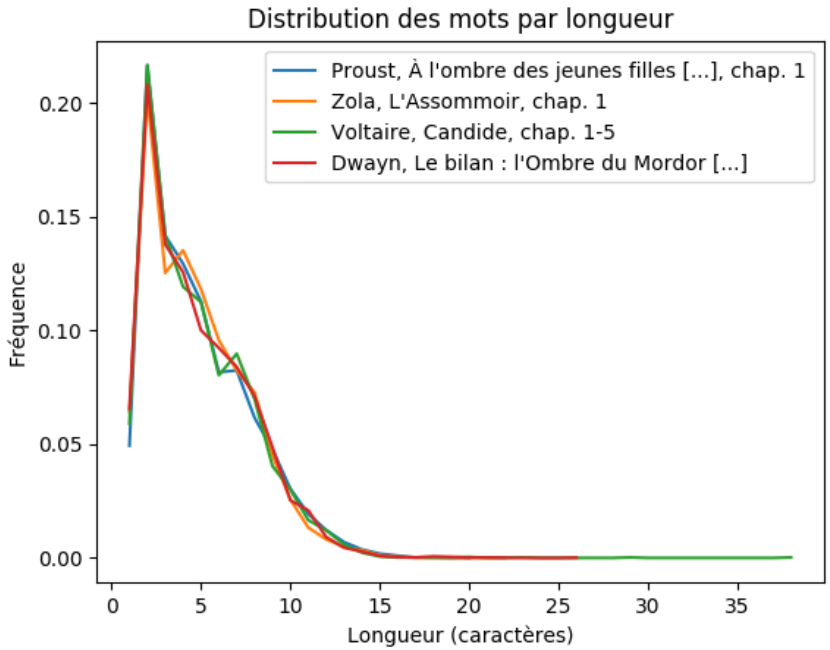
Évaluer la qualité du corpus

1. évaluation intraséque
   - vérification rapide de la qualité linguistique du corpus
   - analyse plus poussée comparativement à d'autres corpus
2. évaluation extraséque

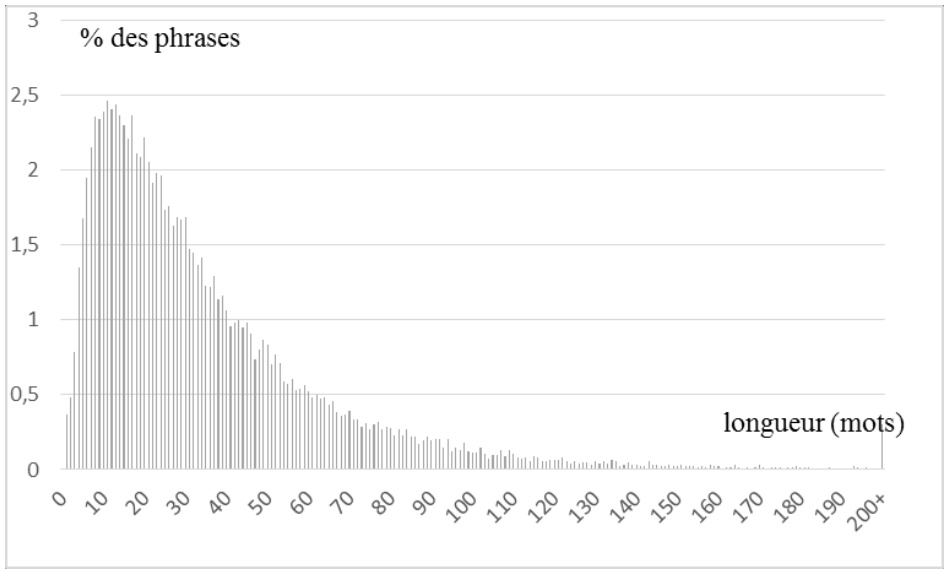
Vérification rapide de la qualité linguistique du corpus

1. distribution de la longueur des mots et des phrases
2. quantités de duplicats
3. 200 statistiques calculées sur les corpus de la collection Leipzig Corpora Collection
   `http://cls.informatik.uni-leipzig.de/en`

Distribution des mots



Distribution des phrases



Verification distributions de la longueur des mots et phrases

- vérifier si distributions similaires sur votre corpus
- si distributions identiques : vérification rapide que les listes de fréquences des mots de la taille la plus fréquente font sens (idem pour la longueur de phrase la plus fréquente)
- si différences de distribution : vérification des longueurs surreprésentées. Typiquement phrases de longueur 1

Mesurer la similarité entre corpus

1. Comparer les listes de fréquences
2. Test d'hypothèse $\mathcal{X}^2$ avec comme hypothèse nulle : "les deux corpus sont des échantillons d'une même population où la proportion des mots est stable ; la différence est due aux variations aléatoires"
3. corrélation de Spearman ρ : similaire $\mathcal{X}^2$ mais calculée sur les rangs.

Comparer les listes de fréquences

selectionner des caractéristiques (mots, lemmes) d'un rang X à Y et les comparer dans les deux corpus
Exemple : 14 premiers noms communs de chaque corpus

Rang	FRKOW (peu d'URL graines)	FRWAC (nombreuses URL graines)
1	*année*	*site*
2	*travail*	*an*
3	*temps*	*travail*
4	*an*	*jour*
5	*jour*	*année*
6	pays	*service*
7	monde	*temps*
8	*vie*	article
9	*personne*	*personne*
10	homme	projet
11	service	information
12	cas	entreprise
13	droit	recherche
14	essai	*vie*

Test d'hypothèse $\mathcal{X}^2$

	Fréquences	
mots	Corpus1	Corpus2
de	6781719	6802262
,	5627749	5633555
la	3613946	3614049
.	3574395	3579032
que	2963992	2956662
y	2642241	2653365
en	2562028	2564809
el	2450353	2446328
a	1885112	1882813
los	1597103	1603537
del	1173860	1172623
se	1139311	1143202
las	10554729	1054924
un	1001556	1000106

Test d'hypothèse $\mathcal{X}^2$

T1	Xa	Xb	
Ya	a	b	L1
Yb	c	d	L2
	C1	C2	N

T2	Xa	Xb	
Ya	a'	b'	L1
Yb	c'	d'	L2
	C1	C2	N

$$a' = \frac{C1 * L1}{N} \quad b' = \frac{C2 * L1}{N} \quad c' = \frac{C1 * L2}{N} \quad d' = \frac{C2 * L2}{N}$$

$$X^2 = Somme(\frac{(Observes - Theoriques)^2}{Theoriques}) = \frac{(a-a')^2}{a'} + \frac{(b-b')^2}{b'} + \frac{(c-c')^2}{c'} + \frac{(d-d')^2}{d'}$$

Test d'hypothèse $\mathcal{X}^2$

	Fréquences		X^2	**p**
mots	Corpus1	Corpus2		
de	6781719	6802262	32,99	<,001
,	5627749	5633555	3,12	0,077
la	3613946	3614049	0,001	0,975
.	3574395	3579032	3,08	0,079
que	2963992	2956662	9,36	<,010
y	2642241	2653365	23,88	<,001
en	2562028	2564809	1,53	0,217
el	2450353	2446328	3,40	0,065
a	1885112	1882813	1,44	0,230
los	1597103	1603537	13,09	<,001
del	1173860	1172623	0,67	0,415
se	1139311	1143202	6,68	<,010
las	10554729	1054924	0,02	0,896
un	1001556	1000106	1,07	0,302

	Corpus1	Corpus2	Total
de	6781719	6802262	13583981
¬ de	103292256	103272321	206564577
Total	110073975	110074583	220148558

- Différence significative pour 5 sur les 14 distributions
- Les 2 corpus sont des échantillons de la même population : pas de différence significative attendue

Test d'hypothèse $\mathcal{X}^2$

Attention, sur des gros corpus, ce n'est plus vrai

- les mots ne sont pas distribués aléatoirement
- avec des gros échantillons, des petites différences de distributions des mots vont être suffisantes pour rejeter l'hypothèse nulle

Corrélation de Spearman ρ

Le coefficient de corrélation de Spearman, appelé R ou ρ est similaire au X^2 mais est calculée sur les rangs.

$$\rho = 1 - \frac{6 \sum D^2)}{n(n^2-1)}$$

avec :
D : différence de rang d'un élément entre deux corpus
n : nombre de rangs considérés
hypothèse nulle : pas de corrélation (ou uniquement la chance) entre les deux listes de fréquences des deux corpus

Corrélation de Spearman ρ

	Fréquences	
mots	Corpus1	Corpus2
de	6781719	6802262
,	5627749	5633555
la	3613946	3614049
.	3574395	3579032
que	2963992	2956662
y	2642241	2653365
en	2562028	2564809
el	2450353	2446328
a	1885112	1882813
los	1597103	1603537
del	1173860	1172623
se	1139311	1143202
las	10554729	1054924
un	1001556	1000106

$\rho = 1 - \frac{6 \sum D^2)}{n(n^2-1)} = 1 - \frac{(6x0)}{14(14^2-1)} = 1 - 0 = 1$

$\rho = 1, p < 0,001$: parfaite corrélation entre les deux corpus, l'hypothèse nulle est démantie

Corrélation de Spearman ρ

- très difficile de trouver deux corpus sans corrélation
- tendance à rejeter l'hypothèse nulle même s'il existe des différences linguistiques significatives
- les différences entre les unités de bas rang ont le même impact que celles de haut rang

Rank-Biased Overlap *RBO*

Compare 2 listes ordonnées quelqueconques avec plus de poids sur les premiers rangs
(Webber et al. ACM transactions on Information Systemes 2010)

$RBO(A,B) = \frac{1-p}{p} \sum_{d=1}^{n} p^d \frac{A_{1:d} \cap B_{1:d}}{d}$
avec $A_{1:d}$ les d premiers rangs de A et $p = 0{,}98$

différence totale : 0 corpus identiques : 1

Faire un filtrage préalable sur les catégories grammaticales précises comme les noms, adjectifs

Tests statistiques

Kilgariff 2001

- Calculer les tests statistiques (X^2, Spearman, etc.) mais ne pas les utiliser pour tester l'hypothèse
- Interpréter les tests statistiques comme des mesures de similarité (même si ce n'en est pas)
- Tester sur les corpus où la similarité est connue, X^2 est plus performant que Spearman et de nombreuses autres mesures comme l'entropie.

Références

Baroni, Marco, Chantree, Francis, Kilgarrif, Adam and Sharof, Serge. 2008. CleanEval : A Competition for Cleaning Webpages. In Proceedings of LREC 2006 , pages 638-643, ELRA, Marrakech.
Councill, Giles Kan. 2008. ParsCit : An open-source CRF reference string parsing package. LREC 2008.
Kilgarrif, Adam. 2001. Comparing Corpora. International Journal of Corpus Linguistics 6(1), 97-133.