

Colloque « Vulgarisation et médiation scientifique »
Fondation Maison des sciences de l'homme – Paris – 05/12/2014

***De l'arbre des espèces à l'arbre de mots,
un outil d'analyse textuelle
né sur un blog de vulgarisation***

Philippe Gambette

LIGM

Université Paris-Est
Marne-la-Vallée



La vulgarisation par les blogs

Les blogs de science

- développement des plateformes de blog en 1999

Pierre Mounier, Le blogging scientifique,
<http://fr.slideshare.net/revuesorg/le-blogging-scientifique>

- développement des blogs de science en France à partir de 2003

Antoine Blanchard, Petite histoire des blogs de science en français,
<http://www.enroweb.com/blogsciences/index.php?post/2014/09/08/Petite-histoire-des-blogs-de-science>

- tenus par des chercheurs, étudiants, journalistes scientifiques, amateurs passionnés, etc.
- **recherches en cours** (carnets de recherche) ou **recherches vulgarisées**
→ ou les deux !

Le « blogging académique »

Intérêts selon André Gunthert

André Gunthert, *Le blogging académique, entre art et science*
<http://culturevisuelle.org/icones/2820>

- formalisation (des idées et perspectives de recherche)
- conversation (publique, interlocuteurs variés)
- itération et expérimentation (distribuée ou collective)
- reproductibilité (mise à disposition des données et outils)

+ réactivité

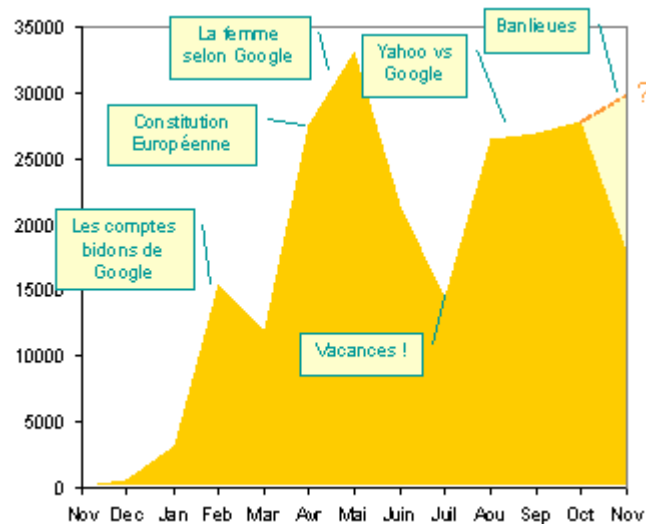
+ archivage

→ « **publication préliminaire** »

Un des pionniers : Jean Véronis

Blog Aixtal

- <http://aixtal.blogspot.com>
- pour ses étudiants en Traitement Automatique des Langues à Aix (professeur à l'Université de Provence)
- premiers billets fin 2004



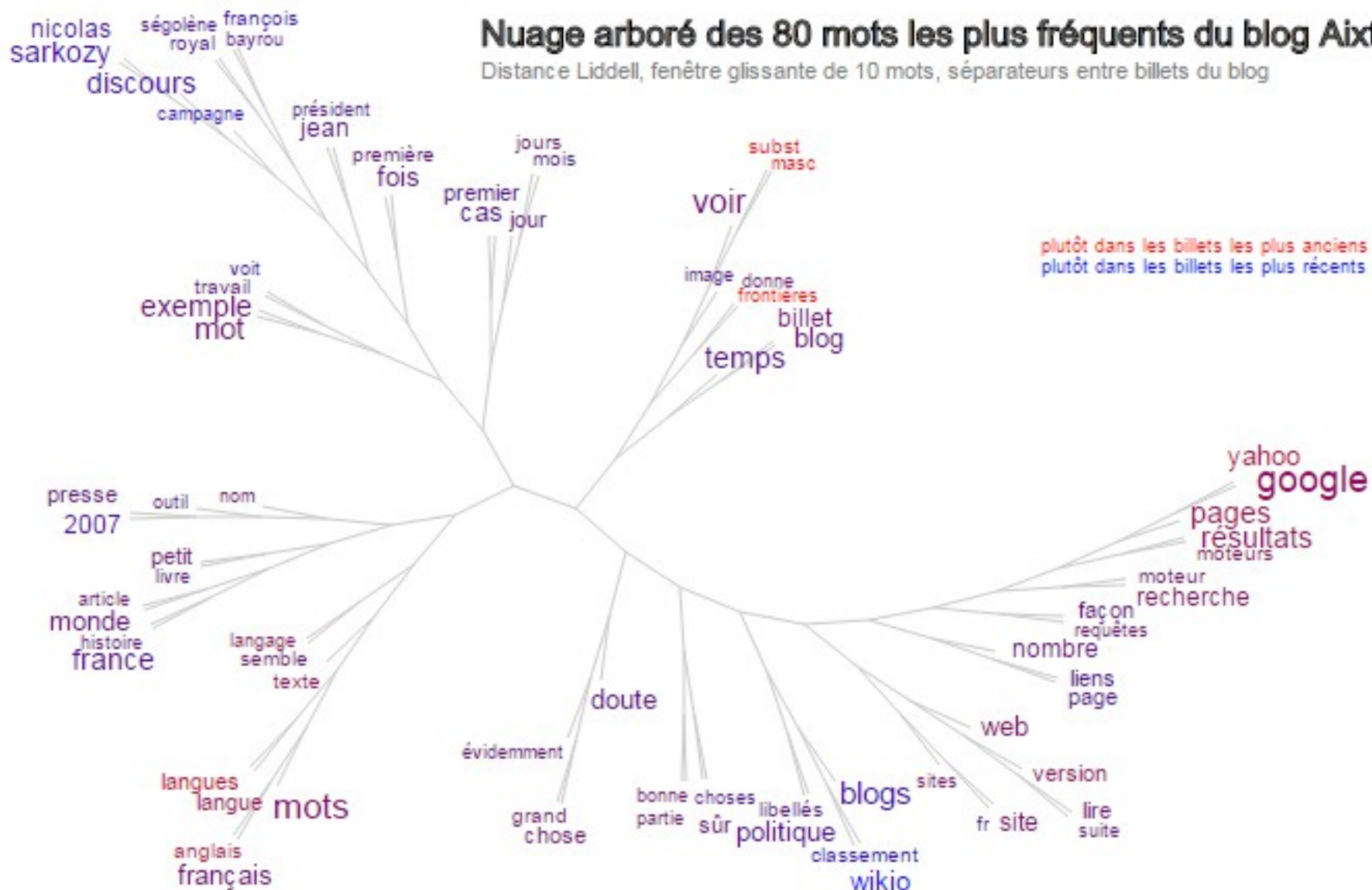
Benoît Raphaël, Jean Véronis : l'adieu et l'héritage

<http://benoitraphael.com/jean-v%C3%A9ronis-l-adieu-et-l-h%C3%A9ritage>

Aixtal en résumé

Nuage arboré des 80 mots les plus fréquents du blog Aixtal

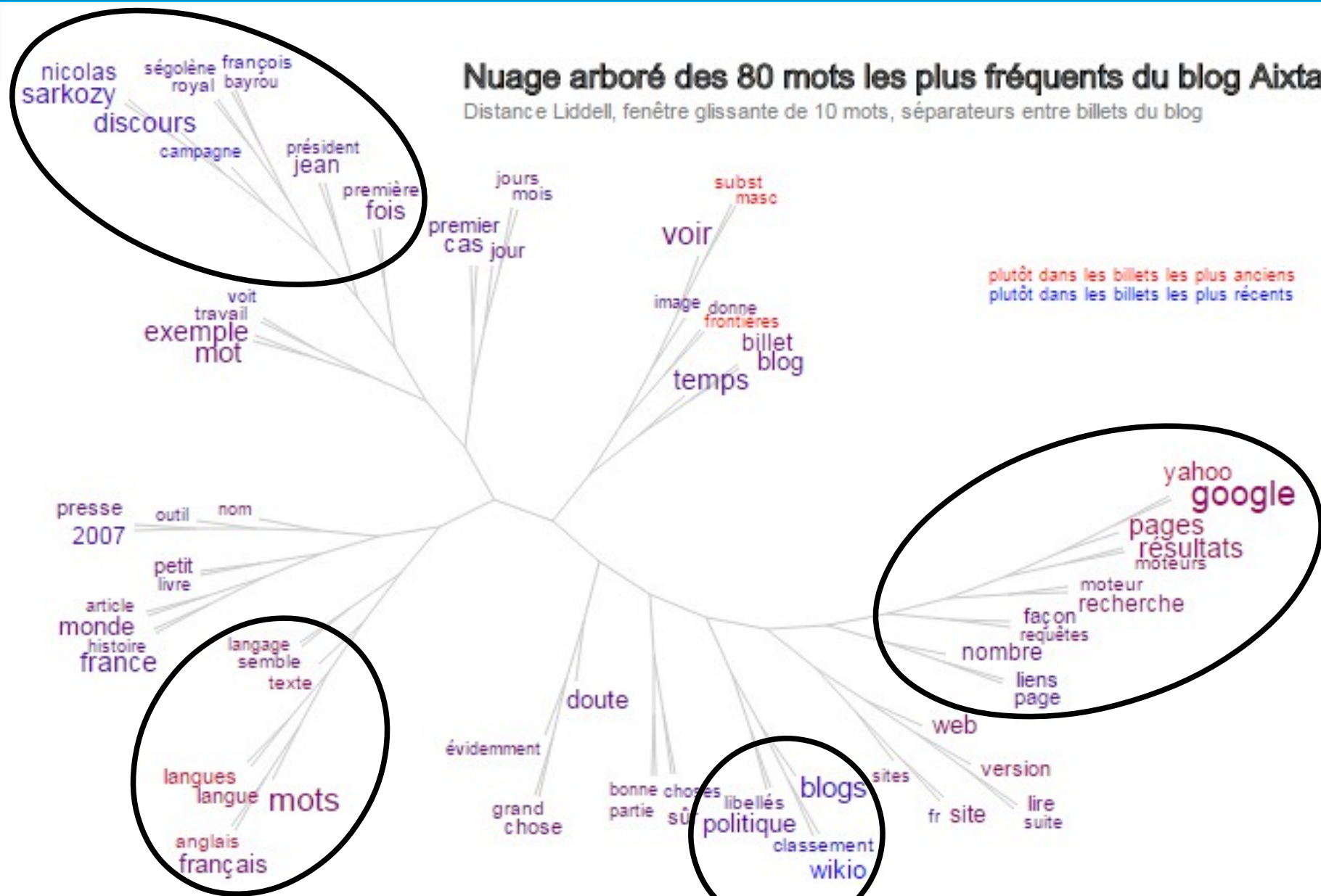
Distance e Liddell, fenêtre glissante de 10 mots, séparateurs entre billets du blog



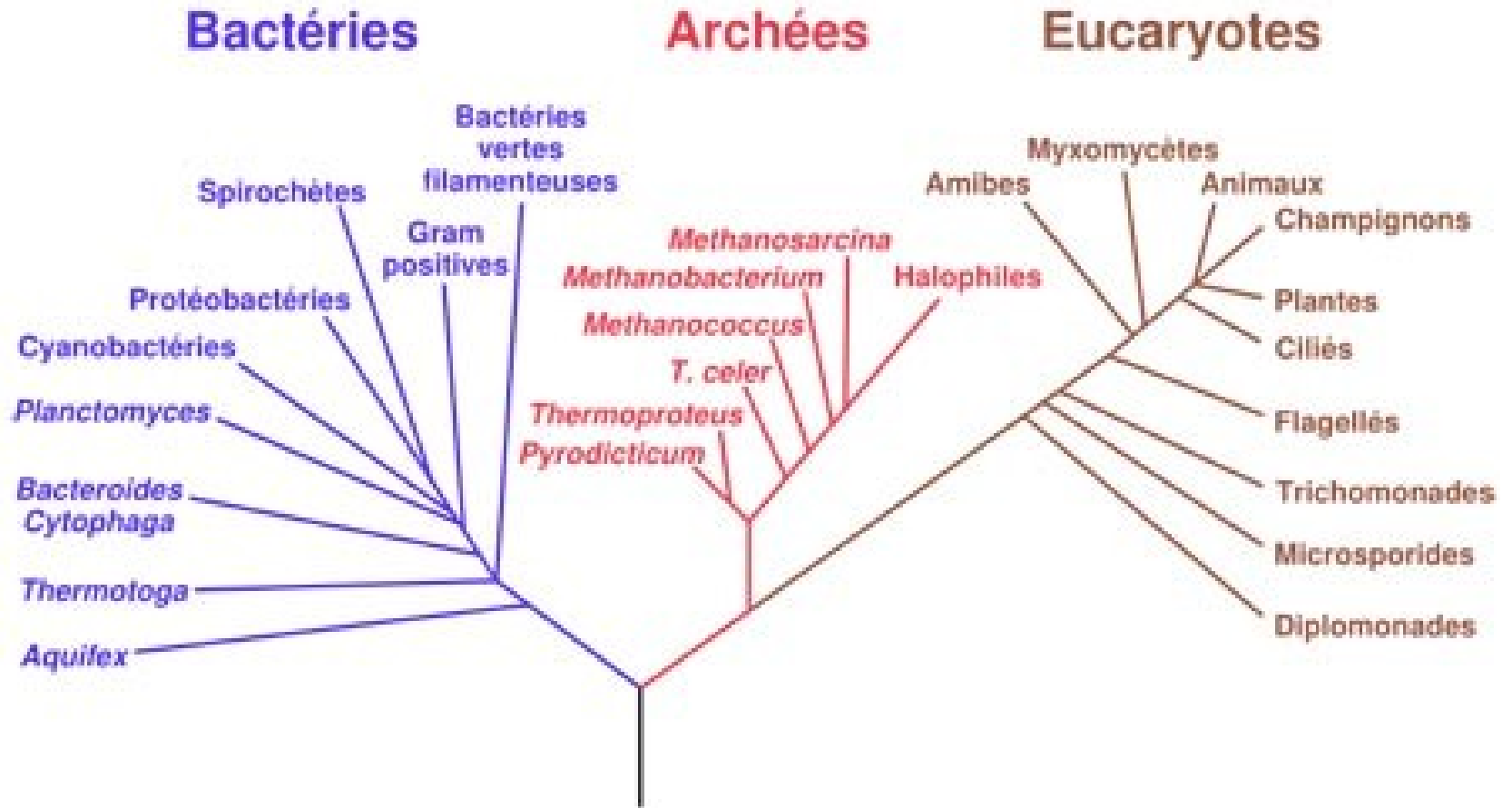
Aixtal en résumé

Nuage arboré des 80 mots les plus fréquents du blog Aixtal

Distance Liddell, fenêtre glissante de 10 mots, séparateurs entre billets du blog



Arbre phylogénétique de la vie



Wikipedia, d'après Woese, Kandler, Wheelis (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proceedings of the National Academy of Sciences*, 87(12), 4576–4579 (1990)

Arbre phylogénétique de la vie

Arbre phylogénétique d'un ensemble d'espèces :

- Les **classer** en fonction de caractères communs
- Décrire leur **évolution**

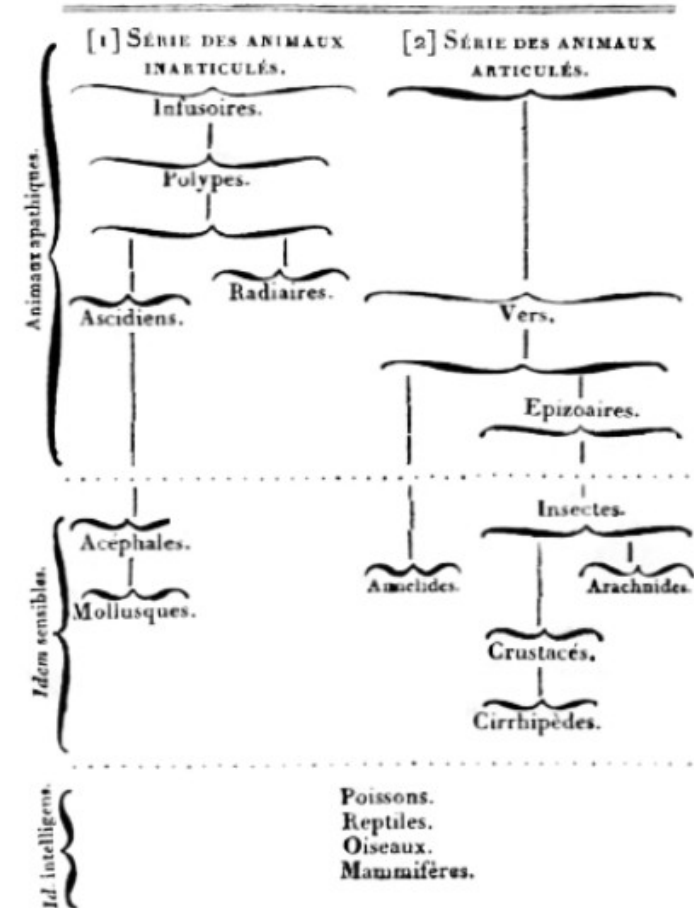
Arbre phylogénétique de la vie

Arbre phylogénétique d'un ensemble d'espèces :

- Les **classer** en fonction de caractères communs
- Décrire leur **évolution**

D'après Lamarck
(1815) *Histoire
naturelle des
animaux sans
vertèbres*

*ORDRE présumé de la formation des Animaux ,
offrant 2 séries séparées , subrameuses.*

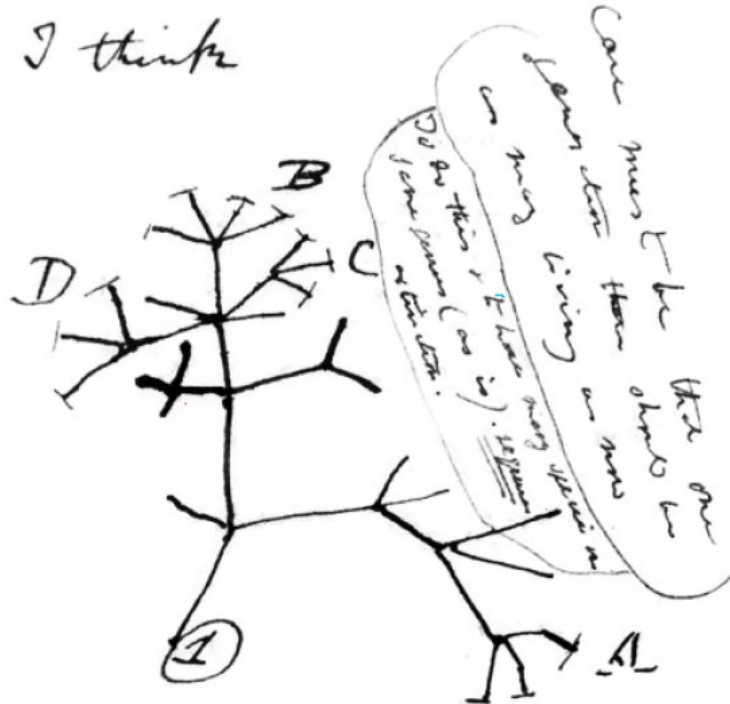


Arbre phylogénétique de la vie

Arbre phylogénétique d'un ensemble d'espèces :

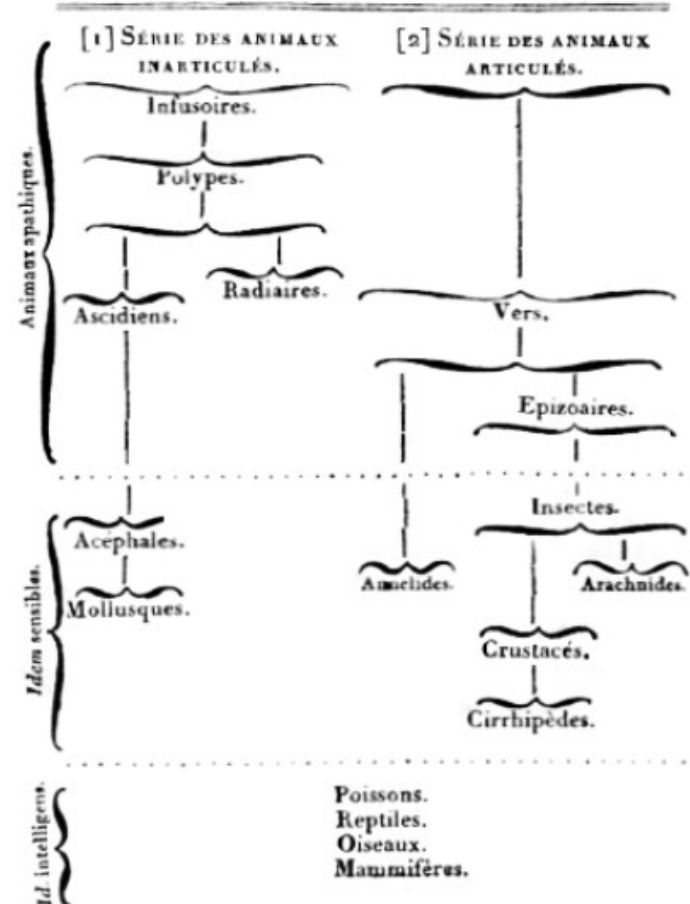
- Les **classer** en fonction de caractères communs
- Décrire leur **évolution**

Darwin (1837)
Carnet B



D'après Lamarck
(1815) *Histoire naturelle des animaux sans vertèbres*

ORDRE présumé de la formation des Animaux, offrant 2 séries séparées, subrameuses.



Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

Données sur
les feuilles

MOTS

Position des mots

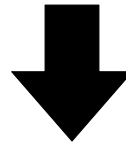
Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0

MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

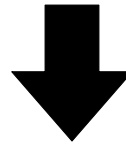
Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

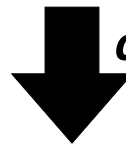
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0



classification hiérarchique ascendante

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

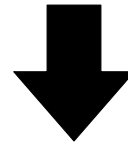
Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

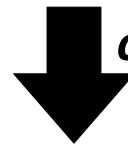
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



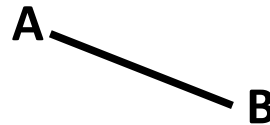
Distances entre les feuilles

	A+B	C	D
A+B	0	5	6
C	5	0	3
D	6	3	0



classification hiérarchique ascendante

Arbre



MOTS

Position des mots

Distances fondées sur la cooccurrence entre les deux mots

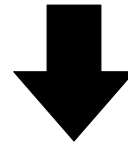
Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A+B	C	D
A+B	0	5	6
C	5	0	3
D	6	3	0

MOTS

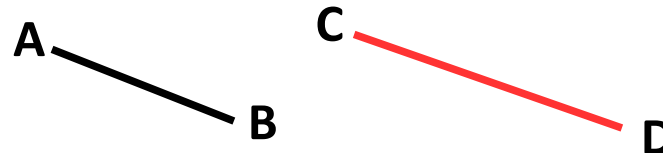
Position des mots

Distances fondées sur la cooccurrence entre les deux mots



classification hiérarchique ascendante

Arbre



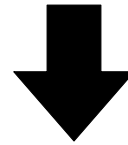
Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A+B	C+D
A+B	0	5,5
C+D	5,5	0

MOTS

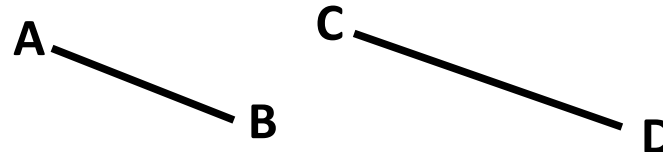
Position des mots

Distances fondées sur la cooccurrence entre les deux mots



classification hiérarchique ascendante

Arbre



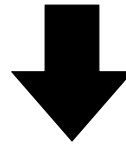
Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

Données sur les feuilles



Distances entre les feuilles

	A+B	C+D
A+B	0	5,5
C+D	5,5	0

MOTS

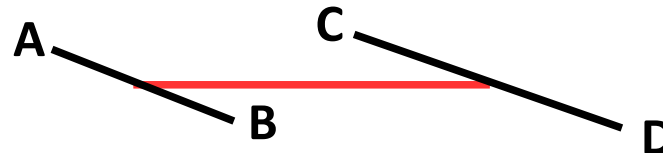
Position des mots

Distances fondées sur la cooccurrence entre les deux mots



classification hiérarchique ascendante

Arbre



Méthodes de construction à partir de distances

ESPÈCES

Séquences ADN

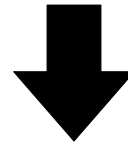
Distances fondées sur la différence entre les deux séquences (mutations, insertions, délétions)

MOTS

Position des mots

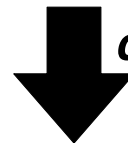
Distances fondées sur la cooccurrence entre les deux mots

Données sur les feuilles



Distances entre les feuilles

	A	B	C	D
A	0	2	5	6
B	2	0	5	6
C	5	5	0	3
D	6	6	3	0



classification hiérarchique ascendante

Arbre



Un premier arbre sur Aixtal

BILLETS RÉCENTS

PRISM: Orwell en a rêvé, les US l'ont fait

Podcast: Qu'est-ce que le traitement automatique des langues ?

e-Reputation: Droit à l'oubli et autres réflexions

Expertises vocales: Lettre à Jérôme Cahuzac

Conf: Big Data et Technologie du Langage

Trendsboard: L'app mobile #PepsiBuzz est sortie

Trendsboard: Analyse d'un buzz cochon

Trendsboard: Analyse du buzz #geonpi

Trendsboard: La version US c'est parti !

Twitter: Analyse du buzz Charlie Hebdo

Ayrault: Un discours sans surprise

Outil: Un demi-siècle de discours de politique générale

Législatives: Philippe et Catherine, les prénoms des candidats

Législatives: Carte de France de la (non)-parité

Appli: France 2012 - Législatives

Présidentielle: La présence des candidats sur le Web entre les deux tours

Débat: Moi, François Hollande

Google: Fichier juif ?

Présidentielle: Le Web a fait mieux que les sondeurs

Présidentielle: La présence des candidats sur le Web

... et plus

ARCHIVES

Par date

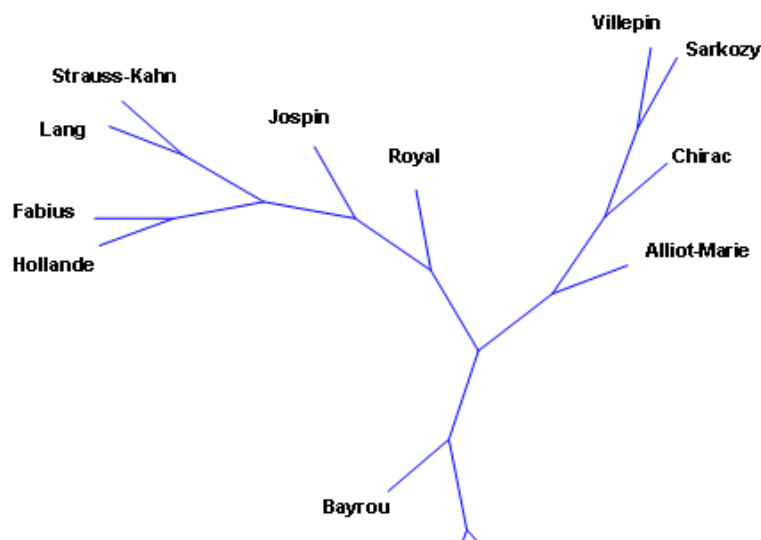
Par catégories

LUNDI, AVRIL 24, 2006

2007: L'arbre des prétendants

Avez-vous remarqué que lorsqu'on parle de Villepin, on parle souvent de Sarkozy? Que lorsqu'on parle de de Villiers, on parle souvent de Le Pen? Et ainsi de suite: Voynet/Cochet, Besancenot/Laguiller, etc. On peut **quantifier ces parentés** sur le Web à l'aide d'un simple moteur de recherche. Il suffit de regarder pour chacun des couples possibles **combien de fois les deux noms apparaissent ensemble dans une même page et d'appliquer des méthodes du type de celles qu'utilisent les biologistes pour représenter les parentés entre organismes vivants à partir des séquences d'ADN (arbre phylogénétique)**. C'est ce que j'ai fait pour 18 des principaux prétendants à l'Elysée, grâce au moteur Dir.com (il vaut mieux éviter Google pour ce genre de calcul, pour des raisons que j'ai déjà largement évoquées).

Voici l'arbre des prétendants:



MES LIVRES

LES MOTS DE NICOLAS SARKOZY

Je veux être le Président de la réconciliation. Je veux être le Président de la valeur travail. Je veux être le Président du pouvoir d'achat. Je veux être le Président qui fera toutes les libertés. Je veux être le Président qui fera présider la justice. Je veux être le Président qui assurera la capitalisme

Louis-Jean Calvet & Jean Veronis | SEUIL



LOUIS-JEAN CALVET
JEAN VERONIS
**Combat
Pour
l'Elysée**
PAROLES DE PRÉTENDANTS

SEUIL

Inspiration & expérimentation

Concours de l'Eurovision : 20 mai 2006

Je véronise...

Petits travaux ludico-informatiques

Archives du blog

- ▶ 2013 (2)
- ▶ 2011 (3)
- ▶ 2010 (10)
- ▶ 2009 (9)
- ▶ 2008 (19)
- ▶ 2007 (26)
- ▼ 2006 (18)
 - ▶ novembre (3)
 - ▶ octobre (4)
 - ▶ septembre (2)
 - ▶ août (1)
 - ▶ juillet (1)
 - ▶ juin (4)
 - ▼ mai (3)

Arbres phylogénétiques, le making-of...
GoogleFight pour l'orthographe
[Eurovision et géopolitique](#)



Liens

[Maître Véronis](#)
[TreeCloud](#)
[English version of this blog](#)

[Ma page pro](#)
[Quelques projets de programmation](#)
[L'encyclopédie des réseaux phylogénétiques](#)
[Redoc \(réseau doctoral Paris-Est\)](#)

[Ma page perso](#)
[Lisbonne par Pessoa](#)
[Barcelone par Mendoza](#)
[Le Démocheur](#)

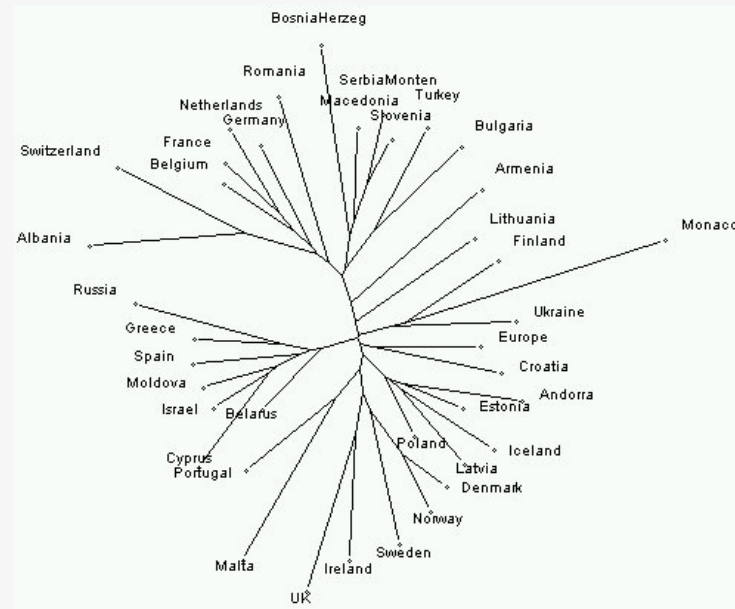
[Freecorp](#)
[Freecorp TagCloud Builder](#)
[Freecorp FuryPopularity](#)
[Freecorp CaptuCourbe](#)
[Algorithms and Permutations 2012](#)
[Doctorants au cinéma](#)

22 mai 2006

Eurovision et géopolitique


Vous aussi vous êtes persuadé que le système de vote à l'Eurovision favorise les pays d'Europe de l'Est ou les pays nordiques qui ne font rien qu'à mettre en place des magouilles pour tous voter pareil et faire gagner le prestigieux concours à un de leurs copains ?

Et bien il est vrai que la proximité géographique transparait un peu dans les votes attribués par les pays, comme on le voit dans l'arbre phylogénétique ci-dessous qui rapproche dans un même sous-arbre des pays qui ont voté de façon similaire.



Tous les détails sur cet arbre sur [cette page de construction de l'arbre phylogénétique selon la "distance Eurovision 2006"](#).

Partage d'outils en commentaires

 LDS A ÉCRIT...

Si je peux me permettre, il serait nécessaire de connaître le temps passé par chaque internaute sur le site pour vraiment parler de succès. Ceci dit, le contenu invite à y rester un moment, c'est certain.

Sinon, j'attends toujours des info sur la manière précise de construire les arbres à partir des couples de mots.


Merci à celui qui pourra me répondre.

22 MAI, 2006 16:00

 FREECORP A ÉCRIT...

lds> Jean avait cité l'article de référence sur le calcul de la distance (NGD, page 3) entre les mots à partir de leur fréquence dans un moteur de recherche. Une fois que ce calcul est effectué, il proposait de construire l'arbre phylogénétique avec la méthode ADDTREE. Je préfère personnellement utiliser le convivial Splitstree qui prend en entrée du format Nexus (en gros la matrice de distances dans un fichier texte avec une certaine syntaxe). En illustration, l'arbre phylogénétique des pays européens selon la "Distance Eurovision 2006".

Jean> Bravo pour ces 500 000 visites ! A quand le bandeau AdSense pour rentabiliser un peu l'inventivité et la qualité toujours au rendez-vous ?

23 MAI, 2006 01:39 

Un premier nuage arboré

Technologies du Langage

ACTUALITÉS - COMMENTAIRES - RÉFLEXIONS

ECHOS

11-10 Sept-Dix de France Inter
08-10 Le "détail" de Fillon chez Morandini
25-09 Sarkoverdose sur "20 minutes"
21-09 "20 minutes" a remarqué "remarquable"
22-07 Docu Hollande dans Le Monde
22-07 Stratégies a aimé
04-07 Sur Google dans la Tribune
27-06 Echo en couleurs
12-06 Coup de coeur dans Marianne
16-05 Rupture linguistique
... et plus

BILLETS RÉCENTS

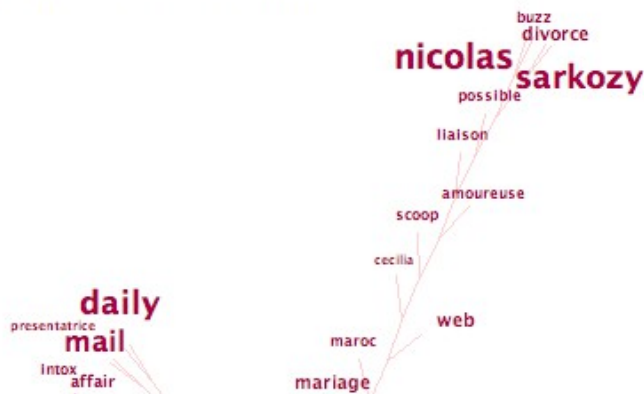
Video: La campagne sur Dailymotion (2)
Video: La campagne sur Dailymotion (1)
Télé: La politique et le sport
Sarko: Moi, je (3)
Sarko: Moi, je (2)
Sourds: La Marseillaise en LSF
Sarko : Moi, je (1)
Télé: Déshabillons-les
Sarko: Grand chef à plumes (3)
Sarko : Grand chef à plumes (2)
Sarko: Grand chef à plumes (1)
Lexique: Sénehitiles

LUNDI, DÉCEMBRE 10, 2007

Actu: Une Ferrari dans un arbre

Vous aurez compris que je ne vous parle pas d'un accident de la route... Je me suis intéressé à la folle rumeur qui parcourt le Web ces jours-ci, et je me suis demandé si, en appliquant mes petits outils, on pouvait en donner une image. J'ai analysé les résultats de Wikio (un superbe moteur, d'ailleurs, même si ce n'est peut-être pas immédiatement apparent — j'y reviendrai certainement dans les prochains jours parce que j'ai été littéralement bluffé par la technicité qui s'y cache). J'ai passé à la moulinette tous les résultats retournés par Wikio sur [Laurence Ferrari](#) depuis le 25/11, et je les ai transformés en **arbre** à l'aide des techniques dont j'ai déjà parlé sur ce blog (par exemple [ici](#) ou [ici](#)).

Voici ce que ça donne (la taille reflète l'importance du mot-clé, et la proximité dans l'arbre correspond à la proximité dans les articles) :



A PROPOS DE L'AUTEUR

JEAN VÉRONIS
AIX-EN-PROVENCE
(FRANCE)



MES LIVRES



Tentative de reproduction

Je véronise...

Petits travaux ludico-informatiques

Archives du blog

- ▶ 2013 (2)
- ▶ 2011 (3)
- ▶ 2010 (10)
- ▶ 2009 (9)
- ▶ 2008 (19)
- ▼ 2007 (26)
 - ▼ décembre (1)
 - tag cloud + tag tree = nuage arboré (1)
 - ▶ octobre (3)
 - ▶ juillet (1)
 - ▶ juin (2)
 - ▶ mai (4)
 - ▶ avril (4)
 - ▶ mars (2)
 - ▶ février (5)
 - ▶ janvier (4)
- ▶ 2006 (18)

Liens

Maître Véronis
TreeCloud
English version of this blog

Ma page pro
Quelques projets de programmation
L'encyclopédie des réseaux phylogénétiques
Redoc (réseau doctoral Paris-Est)

Ma page perso
Lisbonne par Pessoa
Barcelone par Mendoza
Le Démonieur

Freecorp
Freecorp TagCloud Builder
Freecorp FuryPopularity
Freecorp CaptuCourbe
Algorithms and Permutations 2012
Doctorants au cinéma



12 décembre 2007

tag cloud + tag tree = nuage arboré (1)

Vous avez déjà travaillé avec deux objets, pour vous rendre compte que combinés ils fonctionnaient vachement mieux ? C'est la vraie révélation du [dernier article de Jean Véronis sur Aixtal](#) !

Après lecture de [son blog](#) ou [du mien](#), vous êtes convaincus qu'un nuage de mots dont la taille (et la couleur, j'apprécie beaucoup les [teintes rouges, oranges, et bleues du Nébuloscope](#), [introduites en partie](#) chez TagCloud) reflète la fréquence, c'est très utile pour donner un aperçu rapide d'un texte ou d'un corpus. Vous êtes tout autant convaincus qu'un [arbre phylogénétique](#) peut donner un aperçu rapide de [relations entre des mots-clés](#). Et quand on mixe les deux ? Je cite : "c'est marrant les arbres... ils peuvent nous raconter des histoires". Et voilà, un nouvel outil de visualisation d'histoires !

Dans [l'exemple dédié à Laurence Ferrari](#), ce sont les résultats Wikio qui sont "passés à la moulinette". Mais de la même façon qu'on peut effectuer un nuage de mots depuis un simple texte (avec [TagCloudBuilder](#) et une [échelle de coloriage logarithmique bien sûr](#) !), j'ai l'impression qu'un [nuage arboré \(tree-cloud](#) pour nos amis anglophones) est encore un meilleur moyen d'avoir un aperçu rapide d'un texte. L'ordre alphabétique complètement artificiel des tags dans le nuage est remplacé par une disposition hiérarchique intuitive et informatrice !

Alors maintenant, quelle distance entre mots choisir pour reconstituer l'arbre, sur un simple texte ? J'ai fait l'essai suivant : pour chaque paire de mots, leur distance mutuelle est égale au log du nombre de mots minimal qui les sépare. J'avais déjà tenté de commencer à [justifier théoriquement l'introduction du log pour certaines reconstructions "phylogénétiques"](#) (la seconde partie du billet est toujours dans les cartons), là je l'introduis a priori seulement pour éviter quelques trop longues branches. J'ai prévu de tester quelques améliorations de cette idée : nombre de mots moyen séparant la paire de mots, éventuellement seulement sur les occurrences les plus rapprochées de la paire...

Je garde ces préoccupations pour un éventuel [billet suivant](#), et je livre un exemple d'application de ce principe, tiré du premier texte de longueur convenable que j'avais sous la main, une [interview en anglais de Tom Sharpe](#). L'heure tardive ne me permet pas de mixer arbre et nuage de mots, les voici donc, à mixer mentalement (pour le moment) :

```
20 admirers after alone also always anymore author
barcelona bear because been before book books
brava britain britain' british calls cambridge can'
catalonia character costa crime daughters decorated
depressing desk don' down english everything
famous feel figure filled friends from gbp gentleman
get go health hear here hideaway him himself home
hotel house language latest learn lecturer life llafranc
long many meat money months most nancy nine novel
nowhere other out own photographs popular price
regularly rules says sea see set sharpe sharpe'
she so some spain spanish spending such that they
thing this three time tom two up villa village villas
want was we what where which white who wife wilt
with work worked writer year years
```

9h plus tard...



[Jean Véronis](#) a dit...

Très intéressante discussion, Philippe. Oui, c'est exactement ça: un mix entre les nuages et les arbres. Je n'ai pas ajouté d'information de couleur dans l'arbre parce que je trouve que c'est trop. La redondance aide dans les nuages, mais on n'y a qu'une info. Ici on a en deux : la fréquence et la cooccurrence. Une variation de couleur redondante avec la seule fréquence me semble difficile à lire.

Mon idée est de lui faire porter une autre information, je te le donne en primeur, parce que je n'ai pas encore posté là-dessus : la fraîcheur de l'info. **Le plus vif correspondra aux dernières news, le plus pastels aux plus anciennes.** J'ai fait des essais c'est très parlant (mais il me reste quelques réglages, notamment d'échelle: fait-on du log dans le temps? etc.).

En passant: les couleurs du Nébuloscope n'étaient pas inspirées de Yahoo mais de TagClouds. J'ai toutefois ajouté un niveau maximal, le rouge, qui n'était pas présent dans les couleurs initiales et qui me paraît important pour une lecture instantanée du "topic". J'ai aussi ajouté (mais je ne le fais apparaître que sur certains nuages) un niveau 0, qui est un gris minuscule à peine lisible (histoire de signaler : ne vous fatiguez pas à lire, mais il y a encore plein d'autres mots).

Par ailleurs, oui, **on peut faire ça sur n'importe quel texte.** C'est ce que je fais sur les discours politiques par exemple. J'ai un poste en préparation, que je n'ai jamais eu le temps de finir, sur l'arbre du travail chez Ségo et chez Sarko.

Petite remarque sur la cooccurrence, et le calcul d'une distance entre deux mots présents dans un texte. On peut utiliser une fonction qui reflète cette distance, mais **j'ai fait des tonnes d'essais et c'est inutilement compliqué.** J'utilise une bonne vieille fenêtre de x mots (ça dépend de la taille globale du corpus et ce qu'on veut visualiser, généralement 10 à 30. Ce qui éclaircit par contre drôlement la forêt, c'est **l'élimination des mots-outils (articles, prépositions, etc.)** qui polluent pas mal la situation sur tes exemples.

Au fait, ce nuage dynamique des cooccurrences de mots dans une fenêtre, c'est déjà exactement ce que je faisais avec la fonction Voisins dans Discours 2007.

Exemple:

[Travail \(Ségo\)](#)

[Travail \(Sarko\)](#)

Un grand merci pour ta réaction passionnante !

[12/12/2007 08:36](#)

Mise à disposition de TreeCloud

Archives du blog

- ▶ 2013 (2)
- ▶ 2011 (3)
- ▶ 2010 (10)
- ▶ 2009 (9)
- ▼ 2008 (19)
 - ▶ décembre (2)
 - ▶ novembre (2)
 - ▶ août (1)
 - ▶ juillet (2)

1 janvier 2008

tag cloud + tag tree = nuage arboré (2) Les vœux présidentiels pour 2008

De quoi nous a parlé notre Président [dans ses vœux](#) hier soir ? On avait eu droit l'an dernier aux [nuages de mots des vœux des présidentiables](#) pour nous en donner de jolies synthèses ; cette année, évolution technologique oblige, on va faire le **nuage arboré** de ce discours de 9 minutes. Pour ceux qui n'auraient pas suivi l'[épisode précédent](#), ou le [billet initial sur Aixtal](#), un **nuage arboré**, c'est le pouvoir de visualisation du nuage de mots, associé à celui de la classification hiérarchique en un arbre binaire non orienté !

La preuve en images pour tous les mots prononcés plus de deux fois :

Freecorp
Freecorp TagCloud Builder
Freecorp FuryPopularity
Freecorp CaptuCourbe
Algorithms and Permutations 2012
Doctorants au cinéma



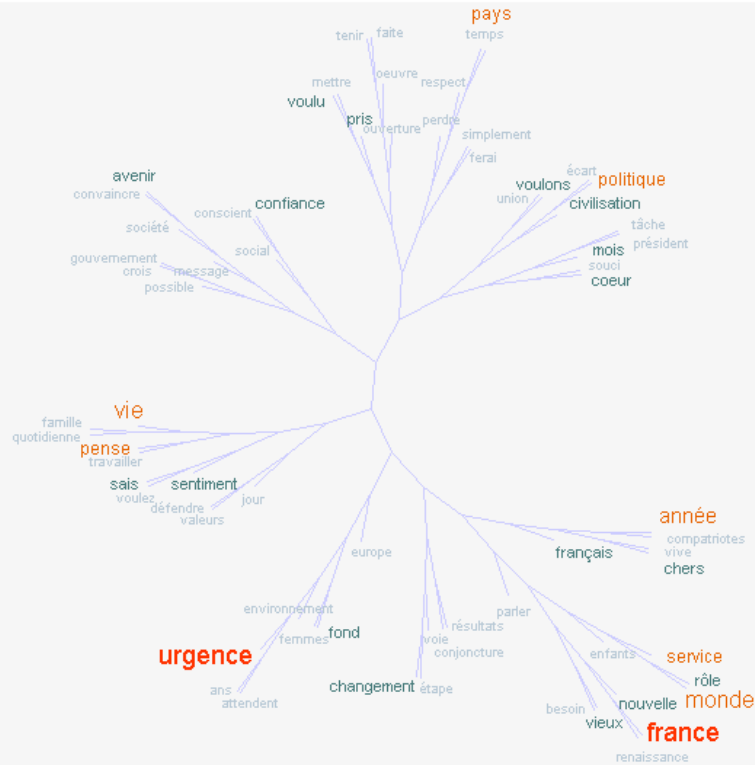
Ils véronisent :

Baptiste Coulmont
El JJ
Arthur Charpentier
François Briatte et Joël Gombin
...



Tags

alimentation Allociné analyse factorielle anaphore APEC BD biodiversité bioinformatique blogosphère
BlogPulse blogs BlogScope Blogsearch Britney Spears bug
buzz CaptuCourbe chemin chronologique cinéma clavier
clustering cognition combinatoire CompareIt
complexité corrections corrélation cuisine Densidées
diachronie DiffDoc Dir doctorat enveloppe convexe Exalead
Flickr FuryPopularity gastronomie GLPK **Google**
Google Docs Google Maps Google Trends graphe
graphique géométrie algorithmique HTML Map igraphe
langage Le Figaro Lexico3 LinkedIn Linternaute livre
logiciel loi de puissance loi exponentielle Longest Path
problem mail messagerie instantanée **moteurs de**
recherche MSN musique même nuage arboré
nuage de mots nébuloscope optimisation
combinatoire orthographe phylogénie



La bonne nouvelle, c'est que si vous auriez vos souhaits créer votre même de tels nuages arborés c'est possible en utilisant la nouvelle version **0.2** de TreeCloud Builder, associée au logiciel SplitsTree (introduction rapide à SplitsTree ici en français). La meilleure nouvelle, c'est que la prochaine version, actuellement en cours de codage et qui sera publiée en même temps que la troisième partie de ce billet, ne nécessitera plus d'utiliser SplitsTree.

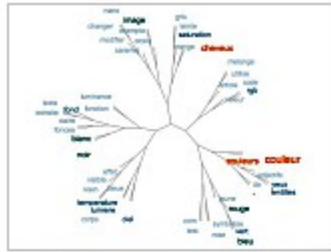
Questions de recherche

- Meilleure méthode pour construire l'arbre de mots ?
 - méthodes provenant de la phylogénie
 - méthodes provenant de la classification de données
- Qualité de la méthode ?
 - robustesse
 - comparaison par rapport à une classification « manuelle »
- Liens avec les autres outils de visualisation de textes ?
 - logiciels commerciaux
 - communauté scientifique de la textométrie
- Applications & utilisations de la visualisation ?
 - plusieurs articles et présentations depuis 2010
- Améliorations de la visualisation ?
 - longueurs de branche (Gambette, Nala & Nasr 2012)
 - dynamique (en cours)

Du « blogging » aux travaux de recherche

- 26/04/2006 : arbre de mots sur *Aixtal*
- 20/05/2006 : arbre des pays de l'Eurovision sur *Je véronise*
- 23/05/2006 : partage d'outil en commentaires sur *Aixtal*
→ **expérimentations méthodologiques sur les arbres**
- 10/12/2007 : utilisation optimisée de l'outil sur *Aixtal*
- 12/12/2007 : formalisation de la visualisation sur *Je véronise*
- 01/01/2008 : mise à disposition de TreeCloud sur *Je véronise*
→ **expérimentations méthodologiques sur les nuages arborés**
- 19/08/2008 : discussions en vue d'un article
→ **état de l'art**
- 03/11/2008 : soumission d'un résumé à IFCS 2009
→ **tests de robustesse des méthodes d'arbres**
- 17/03/2009 : présentation de TreeCloud à IFCS 2009
→ **application à la comparaison de deux pièces de Corneille**
- 11/06/2010 : présentation de TreeCloud aux JADT 2010

Outils et applications pour les nuages arborés



(source)



(source)



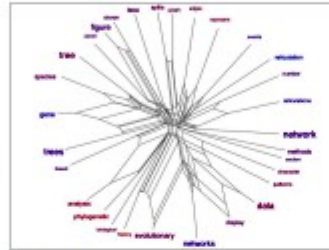
(source)



(source)



(source)



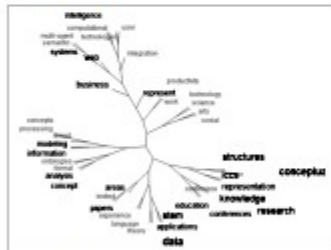
(source)



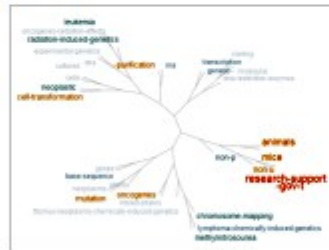
(source)



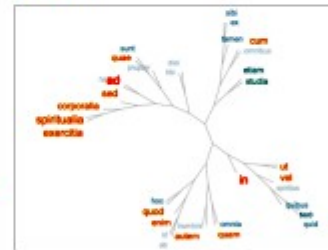
(source)



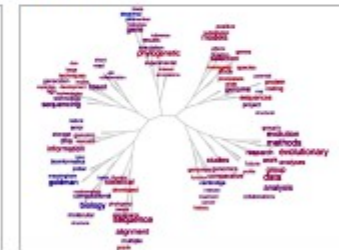
(source)



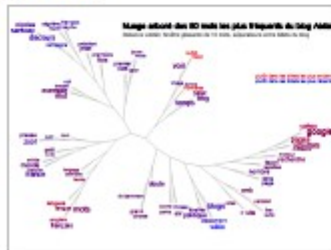
(source)



(source)



(source)



(source)



(source)



(source)

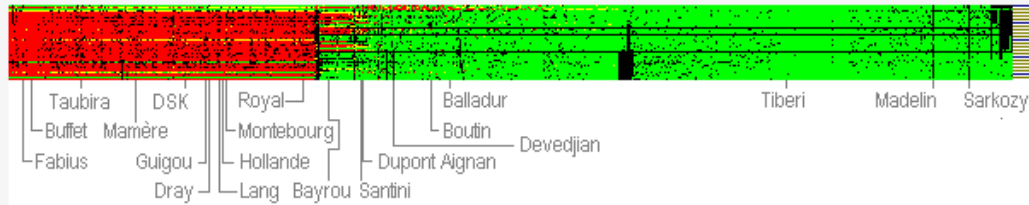
3 implémentations de la visualisation par d'autres programmeurs

D'autres visualisations inspirées par la biologie

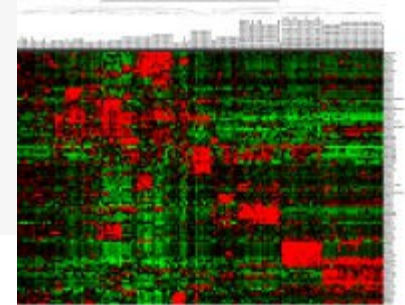
22 février 2007

La puce à ADN des députés

Ma [synthèse des votes des députés sous forme d'arbre](#) avait été bien appréciée, mais j'étais un peu frustré du fait que l'analyse précise de l'arbre soit limitée par de petits [ennuis techniques concernant sa construction](#). Et comme [j'aime bien les visualisations originales](#), en voici une nouvelle pour les votes des députés français (les 46 derniers scrutins publics), basée sur l'idée [de puce à ADN](#) :



<http://gambette.blogspot.fr/2007/02/la-puce-adn-des-dputs.html>

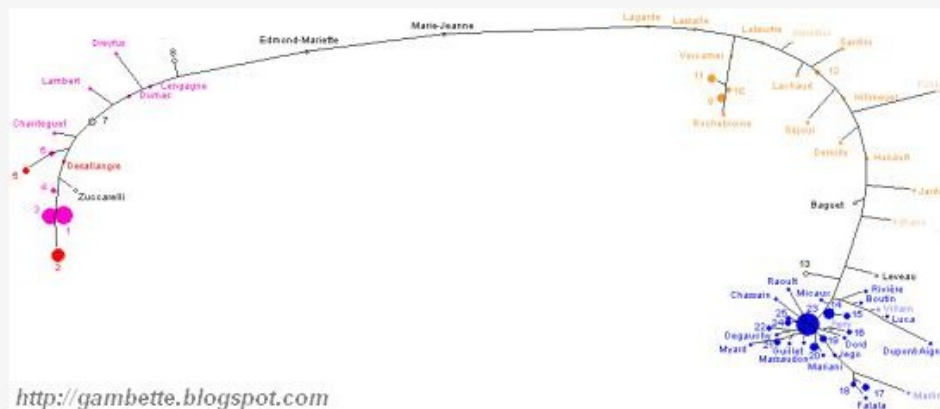


4 janvier 2007

Arbre phylogénétique des députés

Bonne année 2007 ! 2007, qui sera l'année des élections présidentielles, même [le Nébuloscope est au courant](#). Mais c'est aussi l'année des élections législatives, et donc l'occasion de revenir un peu sur les députés et leurs votes à l'Assemblée Nationale.

Je l'avais promis [en juin](#), le voilà enfin après plusieurs jours d'efforts pour colorier ses feuilles : [l'arbre phylogénétique](#) des députés qui rapproche ceux qui votent de façon similaire.



<http://gambette.blogspot.com>

<http://gambette.blogspot.fr/2007/01/arbre-phylogntique-des-dputs.html>

Références (*treecloud.org*)

Philippe Gambette, Jean Véronis (2009)

Visualising a Text with a Tree Cloud, *IFCS'09, Studies in Classification, Data Analysis, and Knowledge Organization* 40, p. 561-570

<http://www.slideshare.net/PhilippeGambette/visualising-a-text-with-a-tree-cloud>

Delphine Amstutz & Philippe Gambette (2010)

Utilisation de la visualisation en nuage arboré pour l'analyse littéraire, JADT'10 (Proceedings of the 10th International Conference on statistical analysis of textual data), Statistical Analysis of Textual Data, p. 227-238

<http://www.slideshare.net/PhilippeGambette/utilisation-de-la-visualisation-en-nuage-arbor-pour-lanalyse-littraire>

Philippe Gambette, Nuria Gala & Alexis Nasr (2012)

Longueur de branches et arbres de mots, *Corpus* 11:129-146

<http://www.slideshare.net/PhilippeGambette/longueur-de-branches-et-arbres-de-mots>

William Martinez & Philippe Gambette (2013)

L'affaire du Médiateur au prisme de la textométrie, *Texto!* XVIII(4)

<http://www.revue-texto.net/index.php?id=3318>

Philippe Gambette, Hilde Eggermont & Xavier Le Roux (2014)

Temporal and geographical trends in the type of biodiversity research funded on a competitive basis in European countries, *rapport BiodivERsa*

<http://www.biodiversa.org/700/download>

Co-auteurs des travaux en cours :

- Edna Hernandez : méthodologie d'utilisation de TreeCloud pour les analyses exploratoires
- Claude Martineau : intégration de prétraitements Unitex dans TreeCloud
- Deepak Srinivas : implémentation de l'algorithme de Barthélemy & Luong, visualisation avec bibliothèque d3.js
- Yu Zheng : visualisation avec bibliothèque d3.js