Aris Xanthos
Unil, SLI

Import de données XML et création de matrices documents-termes

I. Objectifs

- importer un corpus encodé en XML et extraire les unités qui le composent avec Textable
- mettre en pratique le principe de construction d'une matrice documents–termes à partir d'une double segmentation
- mettre en pratique le principe de substitution du contenu des segments textuels par des annotations associés aux segments
- exporter une matrice documents–termes pour la traiter avec un autre logiciel

II. Prérequis

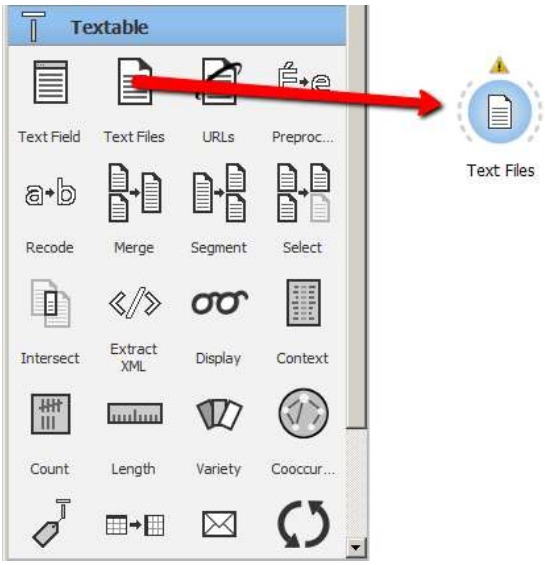
Les étapes suivantes doivent être effectuées au préalable:

- Installer Orange Canvas 3.7 (librement disponible sur https://orange.biolab.si/download/)
- Installer l'extension Textable (libre, voir instructions sur http://textable.io/get-started/)
- Télécharger le corpus des discours inauguraux américains depuis la page du cours-bloc (http://unil.ch/llist/cours-bloc/).[1]
- Il est utile également d'ouvrir le corpus dans un éditeur de texte pour prendre connaissance de sa structure: en particulier, notez au moyen de quels éléments et attributs XML les discours et les mots sont délimités et catégorisés.

III. Importer un fichier texte dans Orange Canvas

1. Exécutez *Orange Canvas*.

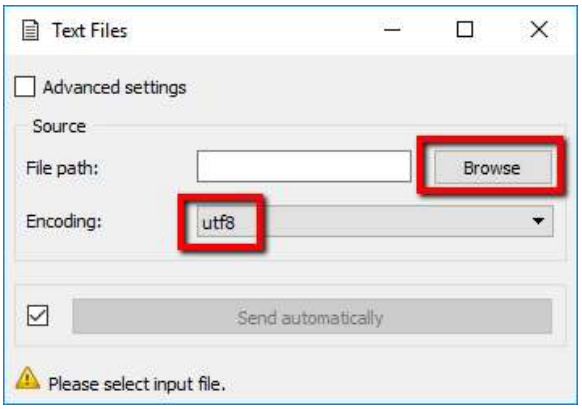
2. Cliquez sur l'onglet **Textable** à gauche de la fenêtre pour l'ouvrir, puis créez une copie du widget **Text Files** (soit en cliquant sur son icône, soit par glisser-déposer sur le canevas).



NB: vous pouvez créer plusieurs copies d'un widget, les déplacer ou les supprimer avec le menu contextuel; il vous permet également d'accéder à la documentation du widget.

[1] Merci à Lucien Chapuisat pour sa contribution à la préparation de ce corpus.

3. Double-cliquez sur le nouveau widget pour ouvrir sa configuration, puis cliquez sur **Browse** pour sélectionner le fichier *inaugural_speeches_tagged.xml* et réglez **Encoding** sur *utf8*:

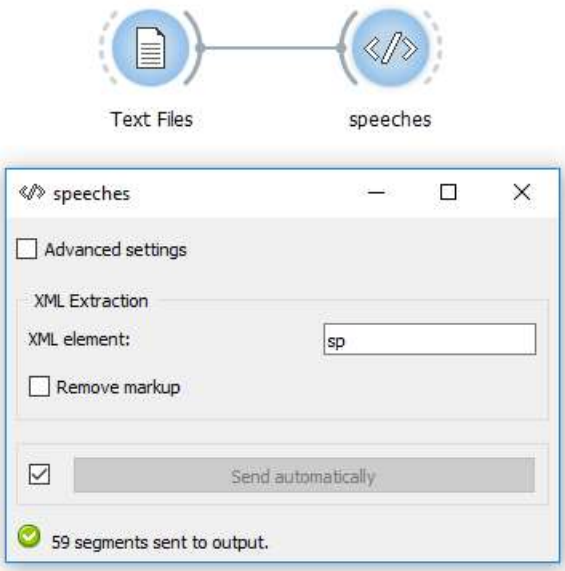


Si l'option **Send automatically** est cochée, le widget s'exécutera directement et affichera le nombre de caractères dans le texte importé, à savoir 5 805 276. Dans le cas contraire, vous pouvez cocher **Send automatically** ou cliquer sur **Send** pour lancer le processus manuellement.

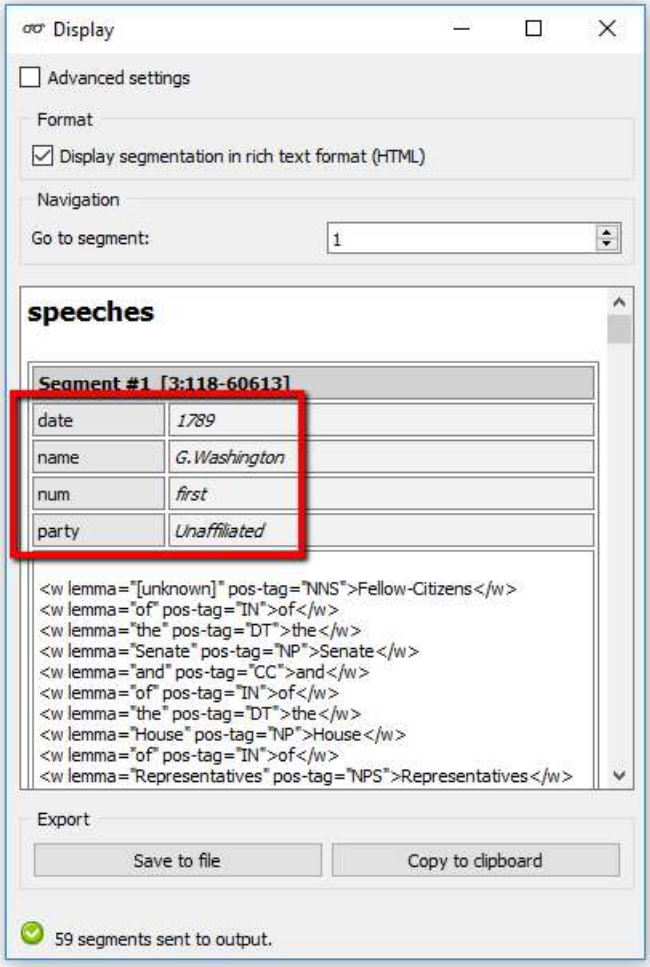
IV. Segmentation en discours inauguraux (=contextes) et mots (=unités)

1. A ce stade, le texte est encore traité comme un objet unitaire. Pour le diviser en unités textuelles de plus bas niveau, comme les discours inauguraux individuels, créez une nouvelle copie du widget **Textable > Extract XML**, double-cliquez pour ouvrir sa configuration et saisissez *sp* sous **XML Element**. Ensuite, connectez la sortie (à droite) de **Text Files** à l'entrée (à gauche) d'**Extract XML**. Cela devrait entraîner l'extraction de 59 segments (voir figure ci-dessous).

2. C'est une bonne pratique de donner aux nouveaux widgets des noms plus significatif; à cet effet, faites un clic droit sur la copie d'**Extract XML**, sélectionnez **Renommer** et saisissez par exemple *speeches*:

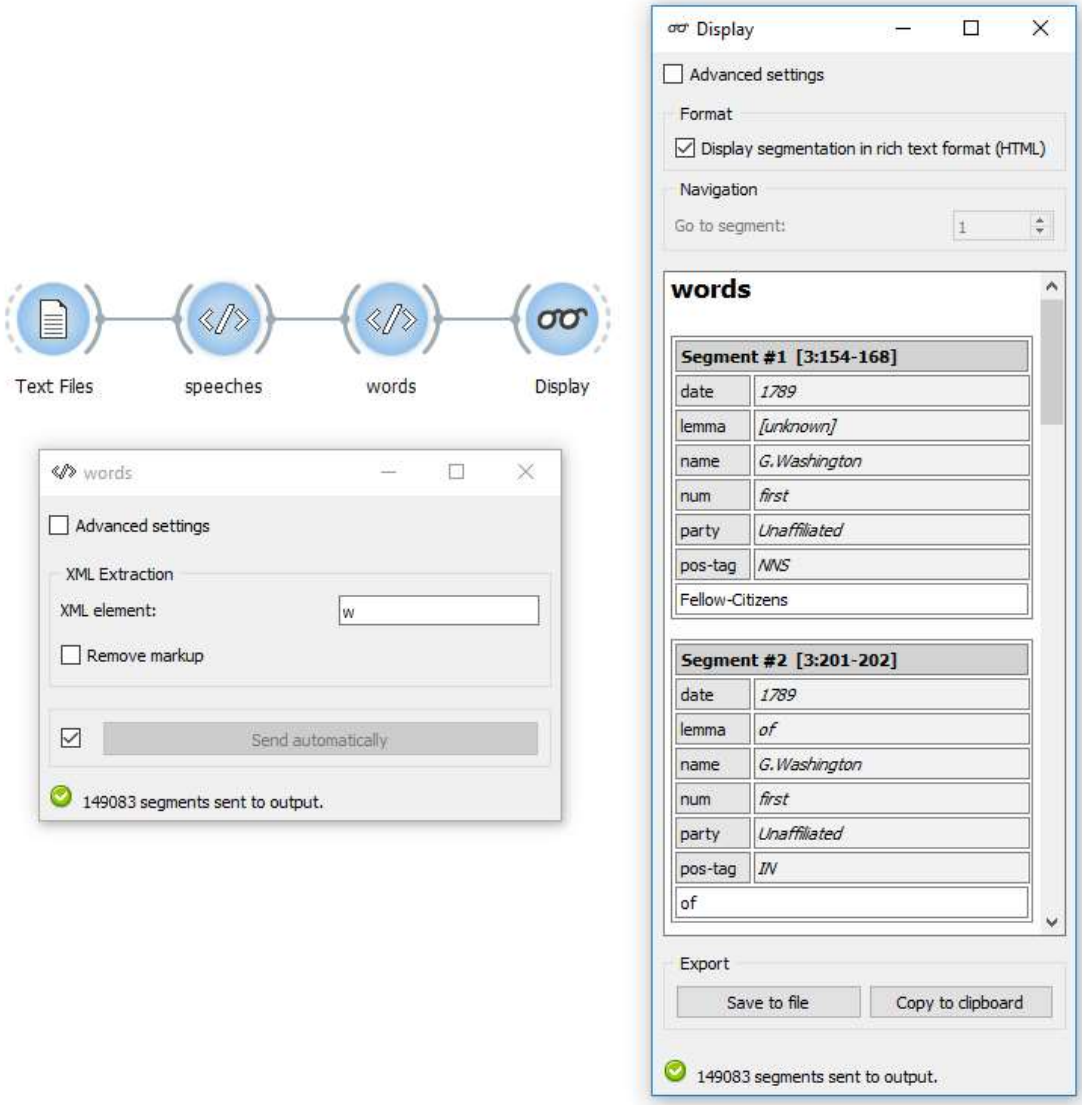


3. Le widget **Textable > Display** peut être utilisé pour visualiser les résultats de cette opération:



Notez que les attributs XML *date, name*, *num* et *party* ont été automatiquement extraits et convertis en annotations associées à chaque discours.

4. La même logique s'applique pour segmenter les discours inauguraux (à la sortie de **Speeches**) en mots: créez une copie d'**Extract XML**, renommez-la *words* p.ex., saisissez *w* sous **XML element** et connectez le nouveau widget à la sortie de celui précédemment renommé *speeches*:

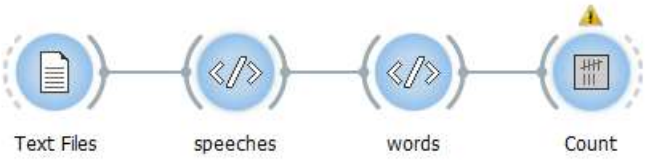


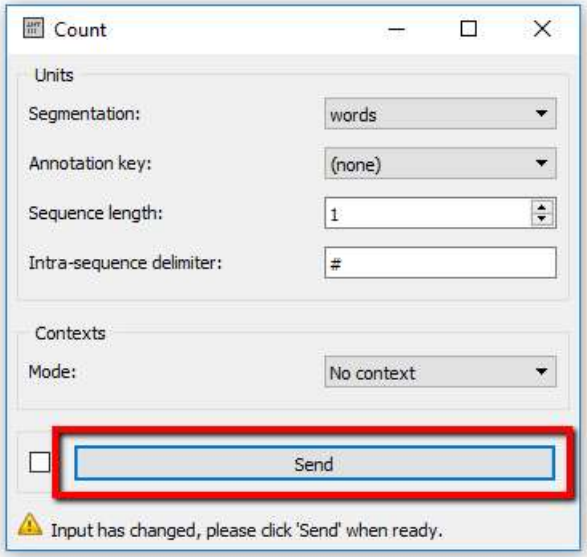
Notez qu'en plus des nouvelles annotations *lemma* et *pos-tag*, chacun des 149 083 tokens extraits hérite des annotations *date, name*, *num* et *party* du discours dont il est tiré.

V. Distribution des mots

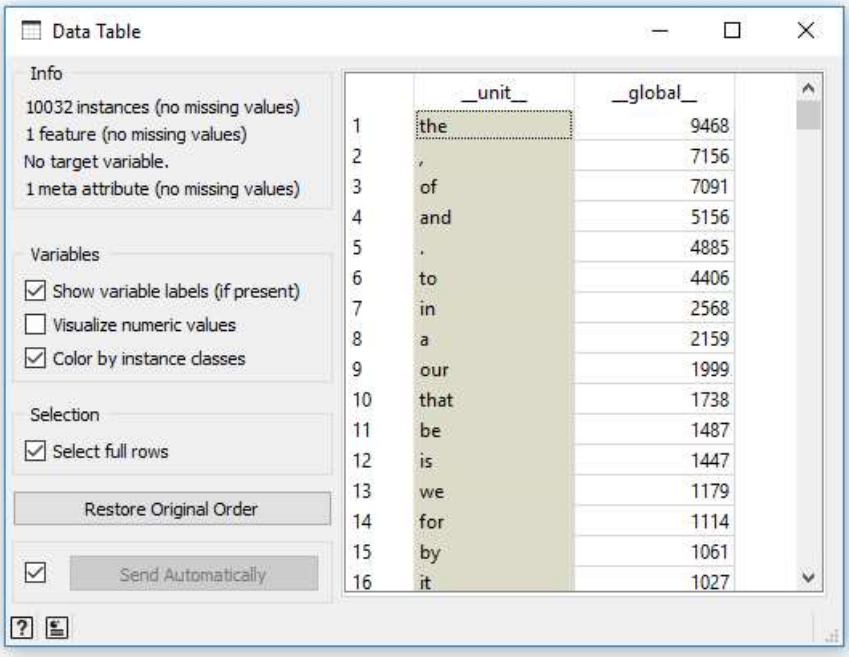
1. Pour afficher la distribution des mots, connectez le widget renommé *words* (ou le widget **Display**) à une nouvelle copie de **Textable > Count**. Par défaut, ce widget est configuré pour compter simplement les segments entrants (ici les mots), donc il peut être directement exécuté en cliquant sur **Send** (voir figure page suivante[2]).

[2] Notez qu'un triangle jaune au-dessus d'un widget signale en général qu'une action est attendue de l'utilisateur, en l'occurrence cliquer sur **Send**.





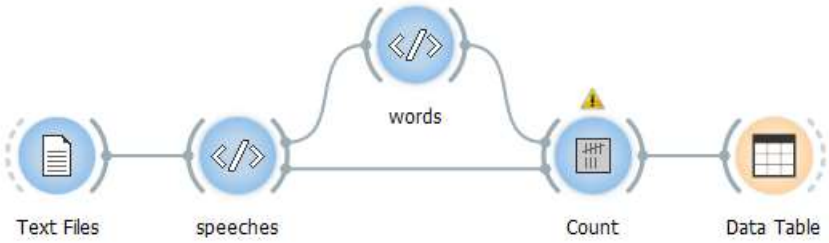
2. Connectez **Count** à une nouvelle copie de **Data > Data Table** pour afficher la distribution des "mots" (qui incluent aussi des signes de ponctuation et autres symboles).



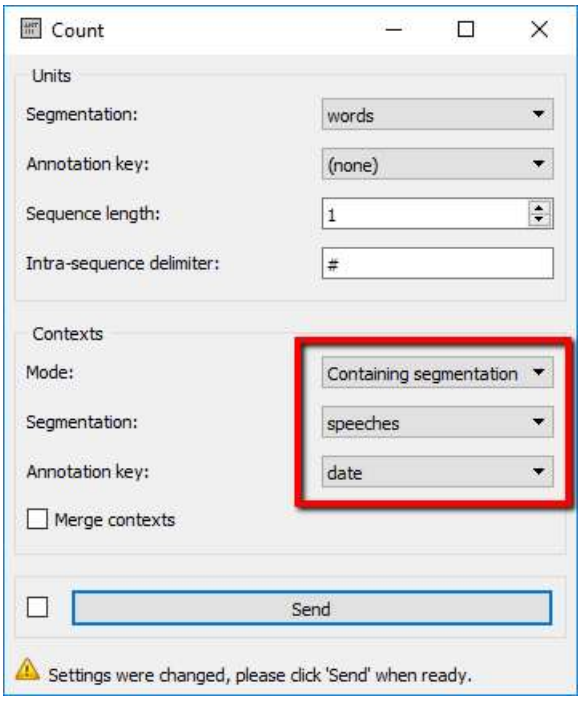
Notez que le tableau peut être trié par ordre alphabétique en cliquant sur l'en-tête de colonne __*unit*__. Relevez également que le nombre de types figure au début de la section **Info** (soit ici 10 032).

VI. Construction de matrice documents–termes

1. Le widget **Count** permet également de construire une matrice documents–termes. A cet effet, il faut lui transmettre non seulement la segmentation en unités (comme c'est déjà le cas) mais aussi la segmentation en contextes qui contiennent ces unités. En l'occurrence, nous pouvons utiliser les discours inauguraux comme contextes, et à cet effet connecter le widget qui produit cette segmentation, *speeches*, à **Count**:



2. La configuration de **Count** doit être modifiée pour indiquer que le comptage des unités doit maintenant être effectué séparément dans chaque contexte (discours). Pour cela, sélectionnez dans la section **Contexts** de l'interface **Mode: Containing segmentation** et **Segmentation:** *speeches***:**

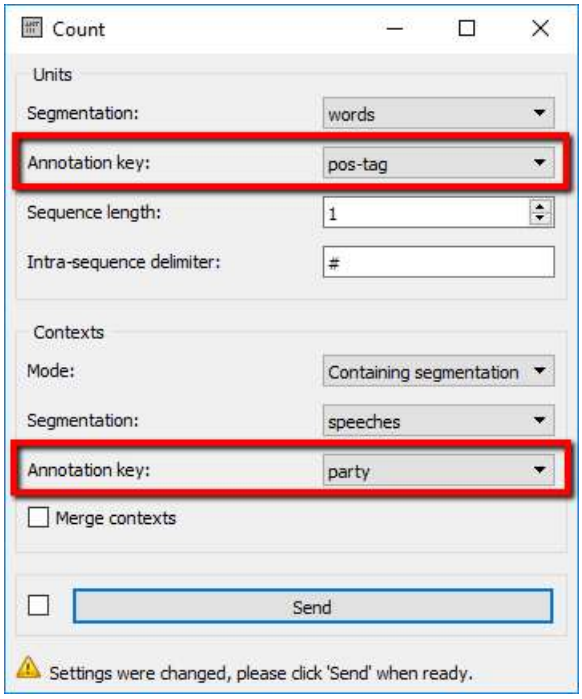


Il faut également sélectionner l'annotation qui servira d'en-tête de ligne dans la matrice documents-termes; pour cet exemple, nous choisissons *date*, qui a le mérite d'être unique pour chaque discours.

3. Lancez le comptage en cliquant sur **Send** et consultez le résultat dans **Data Table**:

context	Fellow-Citizens	of	the	Senate	and
1789	1	71	116	1	48
1793	0	11	13	0	2
1797	0	140	158	1	128
1801	1	104	128	0	79
1805	0	101	138	0	93
1809	0	68	102	0	43
1813	0	65	95	0	42
1817	0	162	264	0	120
1821	1	196	335	0	141
1825	0	244	287	0	116
1829	1	71	88	0	47
1833	1	76	95	0	53
1837	1	198	240	0	150
1841	0	601	795	1	228
1845	1	298	382	0	190

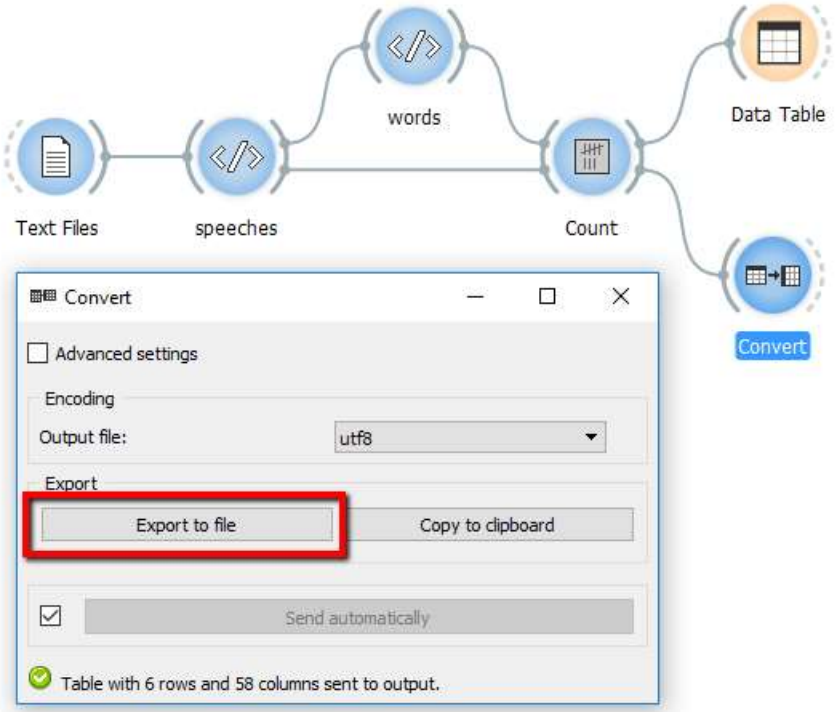
4. Faites l'expérience de modifier l'annotation utilisée pour la définition des contextes (p.ex. *party*) et d'en utiliser une également pour les unités (p.ex. *pos-tag*):



context	NNS	IN	DT	NP	CC	NPS	:	NN
Unaffiliated	92	235	198	18	61	5	11	208
Federalist	166	370	295	44	164	7	18	387
Democratic-Republican	1154	2475	2167	236	742	60	141	2476
Democratic	2699	5502	4962	710	2191	131	409	6515
Whig	627	1427	1362	219	378	26	43	1321
Republican	3678	8151	7402	1416	3119	178	499	9886

VII. Exportation de la matrice documents–termes

1. Pour exporter la matrice en format texte brut (tab-delimited) en vue de l'importer dans *R* ou *Iramuteq* p.ex., connectez **Count** à une nouvelle copie de **Textable > Convert**: puis cliquez sur **Export to file** (ou **Copy to clipboard**) dans l'interface de **Convert:**



2. Cocher la case **Advanced Settings** de **Convert** donne accès à des options supplémentaires, notamment la transposition de la matrice (**Transpose**), nécessaire pour l'importation dans Iramuteq, ou encore la conversion en format "creux pondéré" (**Reformat to sparse crosstab**):