GUIDE D'ANNOTATION

Ce guide d'annotation vous aidera à attribuer les bons labels à chaque observation à l'aide d'une grille de lecture qui fonctionne de la même façon qu'une ampoule à décanter:

- certaines observations se rapprocheront plus des labels liées aux **modifications sur la forme**, décrites en **début** de chaîne (haut du tableau) ;
- d'autres observations se rapprocheront plus des labels liés aux **modifications de fonds**, en **fin** de chaîne (bas du tableau).

Certaines observations peuvent être étiquetées avec plusieurs labels différents. Afin d'isoler le plus possibles les modifications liées à la sémantique, on mettra de préférence des labels de **fonds** (semant_simil, semant_diff, nbr_wr, nbr_rw, nbr_x) une fois avoir vérifié que les labels de **forme** (typo_ortho_gram, wiki_formatting) ne peuvent être utilisés ou ne peuvent être appliqués.

Comment ça marche?
Pour chaque observation, consulter chaque ligne de haut en bas:

- si la modification correspond aux critères décrits, appliquez le label correspondant
- sinon, passez à la ligne suivante jusqu'à atteindre la dernière ligne.

Le but de cette démarche est de s'assurer que les modifications d'ordre sémantique ne soient pas "polluées" par des labels de **FORME** (de mise en page, ie. ortho_gram_typo, wiki_formatting...) et qu'elles portent (dans la mesure du possible) uniquement un label de **FONDS** (d'ordre sémantique, ie. nbr_wr, nbr_rw, nbr_x, semant_simil, semant_diff). Par exemple, si une modification a juste consisté en une modification orthographique sans changement du sens du texte, contentez-vous de mettre **ortho_gram_typo**, mais ne mettez pas le label **semant_simil**. De même, si une modification a causé des traces de vandalisme dans un texte (e.g. "Jules César était un empereur romain QUI VIOLAIT DES ENFANTS"), mettez le label **vandal**, mais ne mettez pas le label **semant_diff**. Il est possible que plusieurs labels labels coexistent (sauf contre-indication), il est cependant préférable de voir si un label est "plus prononcé" que les autres afin de ne pas avoir trop de bruit statistique.

FORME

- mauvaise épellation ("aurtheaugrâfe", "gramère") - violation d'une règle grammaticale sur un ensemble de mots (conjugaison, accords des temps, accord sujet-verbe) - faute de frappe mineure - ajout ou suppression d'espaces blancs - caractères inversés ou manquant au sein d'un mot	>>	**ortho_gram_typo**
- "bonnes pratiques de rédaction" Wikipédia - modifications concernant la mise en page de l'article (ex: ==Titre==, [titre de l'article du lien hypertexte (invisible au lecteur) | texte dans l'article (visible au lecteur)], [[lien hypertexte]], ''' texte mis en gras ''') - ajout de titres, en-têtes ou liens vers d'autres versions de l'article dans d'autres langue, liens dans un paragraphes vers un autre article (mot clicable), ajout de références dans le texte entre les balises <ref> (<ref> référence </ref>), édition de la partie gauche d'un lien hypertexte (entre les signes [[et |). **voir aussi cette liste des autres types de mise en forme:** https://upload.wikimedia.org/wikipedia/commons/1/12/Guide_de_la_syntaxe_Wiki.pdf?uselang=fr	>>	**wiki_formatting**
- effacement de paragraphes entiers, voire de l'article entier - graffitis ("salu sa va", "WIKIPEDIA SUCKS"), allusions sexuelles ("nik ta mère", "bande de salopes")	>>	**vandal**
- changement de la structure des paragraphes ou de la positions des textes sans modification de fonds du contenu lui-même (ex: inversion de la position de deux paragraphes sans modification de leur contenu)	>>	**reordering**
- restauration d'une version antérieure pour retirer des traces de vandalisme (les commentaires contiennent généralement "révocation", "revert", "retour à la version précédente")	>>	**revert_vandal** **non-compatible avec revert_other, remove**
- restauration d'une version antérieure pour toute autre raison que le vandalisme (les commentaires mentionnent généralement "révocation", "revert", "retour à la version précédente")	>>	**revert_other** **non-compatible avec revert_vandal, remove**
- simple suppression de quelques phrases ou éléments sur la version antérieure - se n'applique pas si les suppressions concernent des espaces blancs ou des modifications d'ordre orthographique. Il faudra utiliser ortho_gram_typo dans ce cas.	>>	**content_remove** **non-compatible avec revert_vandal, revert_other**
- ajout de nouveaux paragraphes ou nouvelles lignes avec des informations inédites - ajout de modifications mineures à un contenu/texte existant avec des informations mises à jour ou rectifiées (mandat politique d'une personne encore en vie, événement en cours au moment de la rédaction comme grève/attentat/guerre/élection)	>>	**content_add** **non-compatible avec revert_vandal, revert_other**

FONDS

- rectification ou correction de nombres (chiffres/dates) qui étaient faux - "wr" > Wrong to Right, la version précédente est fausse (valeur incorrecte), la version suivante est vraie (valeur correcte) - "L'humanité est arrivé sur la Lune en 1810" > "L'humanité est arrivé sur la Lune en 1969"	>>	**nbr_wr** **non-compatible avec nbr_rw, nbr_x, semant_simil, semant_diff**
- nombre(s) incorrect(s) (chiffres/dates) incrit(s) à la place d'une valeur correcte - "rw" > Right to Wrong, la version précédente est vraie (valeur correcte), la version suivante est fausse (valeur incorrecte) - "L'humanité est arrivé sur la Lune en 1969" > "L'humanité est arrivé sur la Lune en 1810"	>>	**nbr_rw** **non-compatible avec nbr_wr, nbr_x, semant_simil, semant_diff**
- un ou des nombres ont été modifiés, mais impossible de savoir quelle version est correcte	>>	**nbr_x** **non-compatible avec nbr_wr, nbr_rw, semant_simil, semant_diff**
- reformulation avec conservation du sens général de la phrase d'une version à l'autre - remplacement d'un terme par un autre avec une signification plus précise, mais le sens général de la phrase est préservé - modification sur le style ou la forme du texte sans modification du fonds du contenu par rapport à la version précédente - vérifier avant utilisation si semant_simil n'est pas "écrasé" par un label de forme (exemple: le label wiki_formatting écrase semant_simil si l'éditeur s'est contenté de changer d'ajouter des crochets [[ ]] sans changer le sens du texte) - ne s'applique pas si les modifications portent uniquement sur de la mise en page (ajout de [ ] { } == == , ajout de références entre les balises <ref> ... </ref>)	>>	**semant_simil** **non-compatible avec nbr_wr, nbr_rw, nbr_x, semant_diff** **non-compatible avec content_add, content_remove**
- substitution de mots qui changent le sens d'une phrase ("Jeanne d'Arc est morte au combat") - discours avec un sens différent par rapport à la version précédente ("Les Arméniens ont une musique originale..." > "Les Arméniens ont volé leur musique" - ne s'applique pas si les modifications portent uniquement sur de la mise en page (ajout de [ ] { } == == , ajout de références entre les balises <ref> ... </ref>)	>>	**semant_diff** **non-compatible avec nbr_wr, nbr_rw, nbr_x, semant_simil, vandal** **non-compatible avec content_add, content_remove**
- modification qui ne concerne pas de contenu textuel (ie. Image, formule de maths), la page ne contient pas de paragraphes ou phrases élaborées qui puissent être exploitables pour le projet DIT - modifications par un bot, on peut les trouver facilement par le nom d'utilisateur (HerculesBot, Machinbot) ou par le commentaire ("modification robot: ajout de liens hypertextes") - concerne les modifications sur des pages n'étant pas des articles Wikipédia classiques (ex: 'Discussion: XXX' , 'Projet:XXX', 'Modèle: XXX') destinés à un usage interne pour les éditeurs et non pour les lecteurs - concerne les modifications qui n'ont pas trouvé leur place dans les labels ci-dessus	>>	**other**

Certains articles font l'objet de "guerres d'édition", en particulier lorsque les modifications concernent un sujet sensible: Arménie (génocide), Islam, Israël-Palestine, politique, conflits en cours. Ces cas peuvent être intéressants comme objets sémantiques en veillant à bien séparer la forme du fonds.
exemple: "*Ce résistant du FLN a été assassiné*" > "*Ce terroriste du FLN a été abattu*" >> la forme diffère, mais le fonds reste le même (un homme est mort) >> **semant_simil**
exemple: une intifada en Palestine peut être expliquée de différentes manières d'une version à l'autre. Version 1: elle a été causée par le blocus d'Israël qui a poussé les habitants à se révolter. Version 2: elle a été causée par des bandes terroristes en Palestine qui incitent les habitants à se battre >> la forme diffère ainsi que le fonds car différentes explications sont apportées >> **semant_diff**

Comment procéder

Vérifiez dans cet ordre pour chaque observation:

1 > username (nom d'utilisateur)
est-ce que l'utilisateur est enregistré ? si oui, un pseudonyme apparaît, sinon, son adresse IP est affiché. Cette information peut donner un premier indice sur l'éditeur: est-il plutôt un éditeur régulier (enregistré) ou un intermittent (non-enregistré, avec peut-être des intentions malveillantes ou non) ?

2 > comment (commentaire)
lire le commentaire permet de savoir dans quelle section (exemple: /* Histoire */) de l'article on se trouve. Elle peut indiquer l'intention de l'éditeur :

- corriger un faute
- annuler les modifs d'un utilisateur précédent ("révocation", "annulation", "restauration d'une version antérieure", ces expressions sont utiles pour prédire un revert_vandal ou revert_other)
- parfois ce champ est vide

3 > title (titre)
lire le titre permet de savoir de quoi devrait traiter le texte en question

4 > texte wikipédia
> CE N'EST PAS UNE COURSE
> PRENEZ LE TEMPS DE LIRE

Cliquez sur le lien pour accéder aux modifications Wikipédia à travers le comparateur dédié. Analysez les blocs de textes en surbrillance jaune (blocs supprimés de la version précédente) et bleu (blocs ajoutés dans la version suivante) en vous aidant de la grille de lecture:

> quels sont les termes ou caractères qui sont supprimés / ajoutés ?
> est-ce que les modifs portent sur des éléments de mise en forme ([[texte]] {{texte}} [[texte|texte]] ==titre== *), sur des caractères (ajout de "s" pour accord au pluriel, changement de temps imparfait > présent…) ou des expressions entières ?
> est-ce que certains extraits de paragraphes ont été déplacés d'une section à l'autre sans que leur contenu n'ait été modifié ? (ce cas de reordering peut-être plus difficile à détecter car les extraits en commun ne sont pas forcément en surbrillance)

Si un doute persiste sur les intentions de l'éditeur, vous pouvez cliquer sur ***Modification précédente*** ou ***Modification suivante*** en haut au-dessus des blocs de textes pour voir par exemple si ses modifications ont été contestés par d'autres éditeurs ultérieurement (exemple: un éditeur qui annule les modifications observées car vandalisme ou ajout d'informations erronées).

Comme on cherche ici à isoler au maximum les modifications d'ordre sémantique (semant_*, nbr_*), il faut essayer le plus possible de ne garder que le label semant_* ou nbr_* sur l'observation à annoter ou du moins ne pas trop cumuler semant_* ou nbr_* avec d'autres labels. En cumulant plusieurs labels sur avec semant_*/nbr_*, notre modèle prédictif sera difficilement en mesure de déterminer précisément les caractéristiques du texte en lien avec des modifications sémantiques.