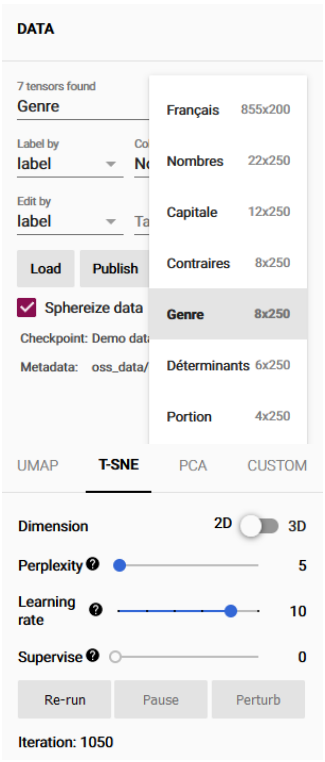
Plongements de mots

Lorsque la page charge, cliquer sur l'icône de lettre A pour afficher les mots.

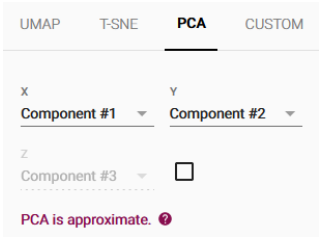
Plongements de mots. Pour générer du texte, l'ordinateur manipule les mots comme des points sur une carte, ou plus formellement un espace. Par défaut, on affiche les mille mots les plus fréquents en français. Les points sont placés de façon à conserver des relations syntaxiques et sémantiques entre eux.

Pour les matheux, les points sont des vecteurs d'un espace vectoriel, donc directement des nombres que l'ordinateur est capable de manipuler.



Dimensions. L'espace est normalement un espace à grande dimension (plusieurs centaines ou plusieurs milliers). Pour pouvoir le visualiser, on n'affiche qu'une projection en deux ou trois dimensions, sélectionnées arbitrairement par la machine. Il ne faut donc pas trop s'encombrer à comprendre la position exacte des mots. Ce qui compte surtout est leur position relative les uns par rapport aux autres.

Genre. Pour bien s'en rendre compte, dans le menu en haut à gauche, sélectionner le modèle « Genre ». Il s'agit d'une sélection de huit mots seulement. En bas à gauche, cliquer sur « T-SNE » et décochez le sélecteur « 2D/3D ». Attendre un millier d'itérations puis cliquer sur stop. Cette manipulation permet à l'ordinateur de trouver les meilleures dimensions sur lesquelles projeter cette sélection de mots. Observez maintenant le résultat : les mots sont placés en paires, toutes organisées de la même façon. Le chemin de « homme » à « femme » est le même que « roi » à « reine » ou que « masculin » à « féminin ».



Autres jeux de donnés. Il est possible d'essayer avec d'autres jeux de donnés, même si l'effet est moins clair. Par exemple, il est possible d'essayer avec « Capitale » (des paires pays-capitale) ou « Déterminants » (des paires d'articles définis et indéfinis). Pour ceux-là, utiliser la projection « PCA » et décocher le « Compenent #3 ».