

# Analyse statistique et empirique des modèles de Word Embedding sur Twitter

Kim Antunez, Romain Lesauvage, Alain Quartier-la-Tente  
sous l'encadrement de Benjamin Muller (Inria)

## 1 Évaluation du modèle implémenté

### 1.1 Comment évaluer le modèle ?

#### 1.1.1 Similarité cosinus

Afin de pouvoir évaluer le modèle, nous devons fixer un certain nombre de critères sur lesquels s'appuyer. L'un des enjeux principaux est de pouvoir estimer la proximité entre deux mots. En effet, le modèle doit capter grâce aux vecteurs ces notions de proximité entre mots. Il faut donc se fixer une distance afin de mesurer ce phénomène.

Il existe différents types de distances que nous pouvons classiquement utiliser pour mesurer la proximité entre deux vecteurs, parmi elles :

- la distance euclidienne  $d(\vec{u}, \vec{v}) = \|\vec{u} - \vec{v}\|_2$  ;
- la distance de Manhattan  $d(\vec{u}, \vec{v}) = |x_{\vec{u}} - x_{\vec{v}}| + |y_{\vec{u}} - y_{\vec{v}}|$  ;
- la similarité cosinus  $d(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\|_2 \|\vec{v}\|_2}$  .

Chacune de ces distances possède des propriétés intéressantes mais l'utilisation d'une d'entre elles en particulier dépend du problème sur lequel on est en train de travailler. Dans le cadre de *Word2Vec*, nous préférons ici la similarité cosinus car elle s'intéresse plutôt à l'angle formé entre deux vecteurs, tandis que les distances euclidiennes et de Manhattan s'intéressent aux valeurs du vecteur directement, or l'angle est un paramètre plus robuste dans notre cas.

Nous pouvons donc interpréter ici la similarité cosinus comme une mesure de l'angle formé entre deux vecteurs représentant deux mots différents. Aussi, une similarité proche de +1 signifiera que les mots sont très proches, une valeur proche de -1 que les mots sont corrélés négativement et une valeur proche de 0 une quasi-indépendance.

#### 1.1.2 Analyse en Composantes Principales

Une fois que le modèle *Word2Vec* est entraîné, nous obtenons des *word-embeddings* pour chacun de nos mots, représentés par des vecteurs de dimension pouvant être 20, 50 ou même 100. Dès lors, il est très compliqué d'avoir un outil de visualisation permettant de repérer les mots qui semblent proches. Pour parer à cela, nous pouvons alors utiliser une analyse en composantes principales. L'objectif premier de cette analyse est d'arriver à projeter un nuage de points sur un espace de dimension inférieure, tout en étant le plus proche de la réalité.

Considérons le cas où nous disposons de  $n$  individus et de  $p$  variables. On note  $X = (x_{ij})$  la matrice de taille  $(n, p)$  des données brutes, où  $x_{ij}$  représente la valeur de la  $j$ -ème variable pour le  $i$ -ème individu. On munit notre espace de la distance euclidienne, on peut alors mesurer la distance entre deux individus  $x_i$  et  $x_{i'}$  en utilisant :

$$d(x_i, x_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

On peut commencer par remarquer que si l'on centre et/ou réduit notre matrice de données, la distance en deux éléments restera inchangée mais que cela permet de donner le même poids à tous les individus. On notera dans la suite  $Y = (y_{ij})$  la matrice des données centrées et  $Z = (z_{ij})$  celle des données centrées et réduites.

La dispersion du nuage de points se mesure grâce à l'inertie, que l'on définit comme

$$I(X) = \frac{1}{n} \sum_{i=1}^n d^2(x_i, \bar{x})$$

. C'est une généralisation de la notion de variance dans le cadre multivarié et c'est cette quantité que l'ACP cherche à représenter au mieux sur les axes factoriels.

Considérons désormais la métrique  $N = \text{diag}(\frac{1}{n}, \dots, \frac{1}{n})$ , on peut alors définir le produit scalaire  $\langle x, y \rangle = x^t N y$ . En notant  $x^j$  la vecteur pour la  $j$ -ème variable, on a  $\text{var}(x^j) = \|y^j\|^2$  et de même  $\text{var}(z^j) = \|z^j\|^2$ .

Pour trouver les axes de l'ACP, on va désormais procéder par projection orthogonale. Rappelons que les coordonnées de projection des  $n$  individus sur une droite de vecteur directeur  $v_\alpha$  est  $f^\alpha = Z v_\alpha$ . Nous allons donc chercher un vecteur  $v_1$  vecteur directeur du premier axe tel que

$$v_1 = \underset{\|v\|=1}{\text{argmin}} \text{Var}(Zv) = \underset{\|v\|=1}{\text{argmin}} v^t R v$$

où  $R = \frac{1}{n} Z^t Z$  est la matrice des corrélations entre les  $p$  variables. On peut alors montrer que  $v_1$  est un vecteur propre associé à la première valeur propre  $\lambda_1$  de  $R$ . Il reste à trouver  $v_2$  orthogonal à  $v_1$  tel quel l'inertie soit maximisée :

$$v_2 = \underset{\|v\|=1, v \perp v_1}{\text{argmin}} \text{Var}(Zv)$$

On obtient alors que  $v_2$  est un vecteur propre associé à la deuxième valeur propre  $\lambda_2$  de  $R$  et, en procédant de manière séquentielle, on obtient  $q < r$  axes orthogonaux sur lesquels on effectue notre projection, où  $r$  est le rang de la matrice  $Z$ .

### 1.1.3 Algorithme t-distributed Stochastic Neighbor Embedding

Bien que l'ACP soit une première manière de résumer l'information contenue dans nos vecteurs, elle présente des limites, notamment dans les vecteurs aux trop grandes dimensions, pour lesquels l'inertie des premiers axes de l'ACP peut se révéler faible. Pour combler ces lacunes, un autre algorithme peut être utilisé, celui dit du t-distributed Stochastic Neighbor Embedding. Contrairement à l'ACP, cet algorithme est stochastique et favorise l'apparition de groupes de mots proches. L'idée reste cependant

la même, c'est-à-dire pouvoir représenter dans un espace l'ensemble de notre nuage de points de manière à repérer les mots proches.

L'algorithme commence par transformer les distances euclidiennes entre nos vecteurs  $(X_1, \dots, X_n)$  en probabilités conditionnelles qui représenteront également les similarités entre nos vecteurs. Pour cela, on note :

$$p_{j|i} = \frac{\exp - \frac{\|\vec{X}_i - \vec{X}_j\|^2}{2\sigma_i^2}}{\sum_{k \neq i} \exp - \frac{\|\vec{X}_i - \vec{X}_k\|^2}{2\sigma_i^2}}$$

La similarité entre les points  $X_i$  et  $X_j$  est alors la probabilité conditionnelle que  $X_i$  choisirait  $X_j$  comme voisin si ces derniers étaient choisis selon une loi gaussienne centrée en  $X_i$ . Ici,  $\sigma_i$  est la variance de cette gaussienne centrée en  $X_i$ , calculer de manière itérative de sorte à maximiser la perplexité du modèle (mesure du modèle de probabilités).

Dans une seconde étape, il faut réaliser les mêmes calculs mais dans l'espace de projection, où nous aurons nos points  $(Y_1, \dots, Y_n)$ . Cependant, puisqu'en dimension réduite, la gaussienne a tendance à concentrer les points, l'algorithme propose d'utiliser une distribution de *Student*, d'où le nom de l'algorithme, et ainsi de calculer :

$$q_{j|i} = \frac{(1 + \|\vec{Y}_i - \vec{Y}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\vec{Y}_i - \vec{Y}_k\|^2)^{-1}}$$

Il ne reste plus désormais qu'à déterminer les vecteurs  $(Y_1, \dots, Y_n)$  qui minimise l'écart entre les deux distributions de probabilité. Pour cela, l'algorithme se base sur la divergence de Kullback-Leibler entre les distributions P et Q :

$$KL(P, Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Avec

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

où  $n$  est le nombre de vecteurs.

Ensuite, on minimise grâce à une descente de gradient. On obtient alors les résultats de l'algorithme t-SNE, qu'il faut cependant analyser avec précaution : il ne s'agit pas d'un algorithme linéaire équivalent à l'ACP, on ne peut donc pas interpréter directement des éléments tels que la taille des clusters obtenus ou leur distance relative.

## 2 Pourquoi on utilise la cosinus distance et pas la distance euclidienne

<https://www.quora.com/Why-do-we-use-cosine-similarity-on-Word2Vec-instead-of-Euclidean-distance>  
Cosine distance + robuste ?

Measuring Word Significance using Distributed Representations of Words : <https://arxiv.org/pdf/1508.02297v1.pdf>

- La magnitude/longueur du vecteur est liée à la fréquence d'apparition des mots donc attention à l'usage de la distance euclidienne.
- During training, words that appear in similar contexts are grouped together in the **\*\*same direction\*\*** by this unsupervised learning algorithm. The distributed representation thus ultimately captures semantic similarities between words. This has been impressively demonstrated by a series of experiments in the original word2vec papers, where semantic similarity was measured by the dot product between normalized vectors.

<https://arxiv.org/pdf/1512.00765.pdf>