WILLARD

Zeste de savoir

SpaceFox à t-il réellement écrit tous ces textes ?

21 décembre 2018

Table des matières

1. Introduction

? SpaceFox à t-il réellement écrit tous ces textes ?

Si je vous pose la question, vous vous doutez bien que la réponse n'est pas si simple.

Je me suis intéressé à la question d'un point d'un point de vue statistique, sans à priori sur les réponses, et les résultats que j'ai pu obtenir sont plutôt intéressants.

i TL ; DR

Vous n'aurez la réponse que lorsque vous aurez terminé de lire le billet.

2. Notre cadre de travail

Comme vous le savez-tous, durant le mois d'octobre, SpaceFox s'est lancé un défi personnel, celui d'écrire trente et une nouvelles , à raison d'une par jour pendant tout le mois d'octobre.

En début novembre il à fait un retour d'expérience/bilan de celui-ci. Il racontait dans son bilan que l'expérience avait été plutôt positive.

Néanmoins, on peut se demander comment c'est possible pour un ~~renard~~ être humain normalement constitué, qui doit certainement avoir un boulot à plein temps de rédiger un texte différent tous les jours.

Deux réponses sont possibles :

1. Il est fort ... très fort

2. Il triche !

C'est donc à la deuxième question, que nous allons tenter de répondre.

Pour répondre à cette question, nous allons essayer d'identifier le style d'écriture de @SpaceFox, puis comparer son style avec celui de ses textes produits et à partir de là, voir si le style est plutôt uniforme (il a gardé la même façon d'écrire) ou toujours différents (il a essayé de se renouveler chaque jour), ou les deux.

3. Cherchons le style @SpaceFox

Pour information, nous nous basons sur ses textes qui sont disponibles sur son site internet .

FIGURE 3. – Quel style ?

Pour trouver son style, nous allons, pour chacun de ses textes identifier les propriétés de chacun de ces textes. Pour cela, nous allons calculer plusieurs éléments :

— Le nombre de mots dans le texte
— Le nombre de verbes, noms, adverbes, ... utilisés dans le texte
— La taille moyenne des mots d'une phrase

On va s'aider d'une bibliothèque de traitement automatique du langage naturel entrainée sur des modèles de la langue française pour faire le travail de *parsing* du texte. On va donc séparer les phrases, faire de l'analyse lexicale, détecter la nature des mots, et calculer les indicateurs ci-dessus.

On obtient donc le tableau suivant :

Contenu masqué n°1

Pour chaque texte (les lignes de notre tableau), on a les différentes caractéristiques.

Il ne nous reste plus qu'à analyser ces résultats pour comprendre si elle se ressemble ou au contraire elles n'ont rien à voir.

4. Comparons les styles d'écriture

Statistiquement, nous sommes ici en possession de données avec des variables continues, qui ont été normalisées. On peut procéder à une analyse en composantes principales pour visualiser ces résultats sur deux dimensions, afin de juger par nos propres yeux (l'œil humain voit mieux en 2D qu'en 33D) si les textes de @SpaceFox suivent un style particulier.

Attention, l'ACP se contente de recréer 2 dimensions à partir de nos 33 dimensions, en s'assurant que les deux dimensions nouvelles gardent le plus d'information possible.

Pour cela, on va utiliser la bibliothèque scikit-learn couplée à pandas de python pour procéder à cette ACP. Mon code ressemble donc à ceci :

```python
import pandas as pd
from sklearn.decomposition import PCA
from sklearn.preprocessing import normalize

df = pd.read_csv('results.csv')
pca = PCA(n_components=2) # une ACP sur deux dimensions
pca.fit(df)

## notre ACP suit une loi normale
PCA(copy=True, n_components=2, whiten=False)

existing_2d = pca.transform(df)

existing_df_2d = pd.DataFrame(existing_2d)
existing_df_2d.index = df.index
existing_df_2d.columns = ['PC1','PC2']
existing_df_2d.head()

print(pca.explained_variance_ratio_)
```

Le résultat de ce morceau de code donne :

```
[0.98615272 0.00731169]
```

Ce qui nous dit que notre ACP a réussi à créer deux dimensions, dont la première (PC1) porte 98,6 % de l'information de notre tableau.

Autrement dit, la projection sur deux dimensions sera plutôt fidèle. On a perdu très peu d'information.

5. La réponse à la grande question

Trêve de plaisanterie, est-ce que tu peux nous dire quelle est la réponse ?

Observons les résultats ensemble :

Résultat de l'ACP

Ce que l'on observe ici c'est que (souvenez-vous que l'axe qui porte le plus d'information est l'axe **PC1**, celui des abscisses) : Le texte N°28 est très différent des autres du point de vue de son style d'écriture.

C'est étonnant, **très étonnant**.

Ce détachement ne semble pas lié à la longueur du texte, car même après l'application d'une opération de centrage et réduction des données, on retrouve toujours ce texte en électron libre.

6. Que s'est-il passé ?

Si on remonte quelques semaines plus tôt, le jour de la publication de ce texte, l'auteur @SpaceFox déclarait :

> Je ne vous ai pas oubliés, et je compte bien terminer les 31 textes! Celui du 28 est à la fois le plus long et celui qui a été **le plus difficile à écrire**, en plus d'être la suite directe du précédent. Place à Le secret le plus intime!
>
> *SpaceFox*

Pourquoi l'auteur déclare-t-il que ce texte a été le plus difficile à écrire ? Tout porte à croire que ce texte n'a pas été écrit comme les autres, ni par la même personne.

J'ai ma petite idée, mais je laisse l'auteur ou quelqu'un d'autre qui a compris le stratagème nous expliquer tout ça de lui-même.

Contenu masqué

Contenu masqué n°1

Numéro du texte	AVG_WORD_SIZE	COUNT_WORD	A	ADJ	ADJWH	ADV	ADVWH
1	3.992	658	0	55	0	28	0
2	3.970	955	0	62	1	74	1
3	3.938	1060	0	56	1	67	0
4	4.070	962	0	68	0	47	0
5	3.666	712	0	64	0	25	0
6	4.078	663	0	41	1	33	1
7	3.855	1070	0	63	2	85	2
8	3.952	1518	0	108	0	74	3
9	3.581	1268	0	69	0	71	2
10	4.089	949	0	71	0	57	1
11	3.924	1202	0	75	0	79	4
12	3.493	942	0	67	2	63	0
13	3.905	423	0	37	0	19	0
14	3.980	1187	0	89	0	60	4
15	3.699	1147	0	81	0	70	0
16	4.385	712	0	61	1	40	0
17	3.916	795	0	46	0	40	1
18	3.864	701	0	54	0	29	1
19	3.928	447	0	54	0	26	0
20	4.534	704	0	69	1	30	0
21	3.742	457	0	34	0	36	0
22	4.337	449	0	47	0	21	0
23	4.035	603	0	51	0	22	0

Contenu masqué

24	3.560	812	0	43	0	33	0	
25	4.011	651	0	43	0	46	0	
26	3.052	159	0	4	0	12	0	
27	3.581	1337	0	76	0	115	2	
28	3.953	1825	0	134	2	159	1	2
29	3.524	712	0	41	0	41	3	
30	3.590	409	0	18	0	24	0	
31	4.211	469	0	35	0	41	1	

Retourner au texte.

Contenu masqué n°2 : Résultat de l'ACP

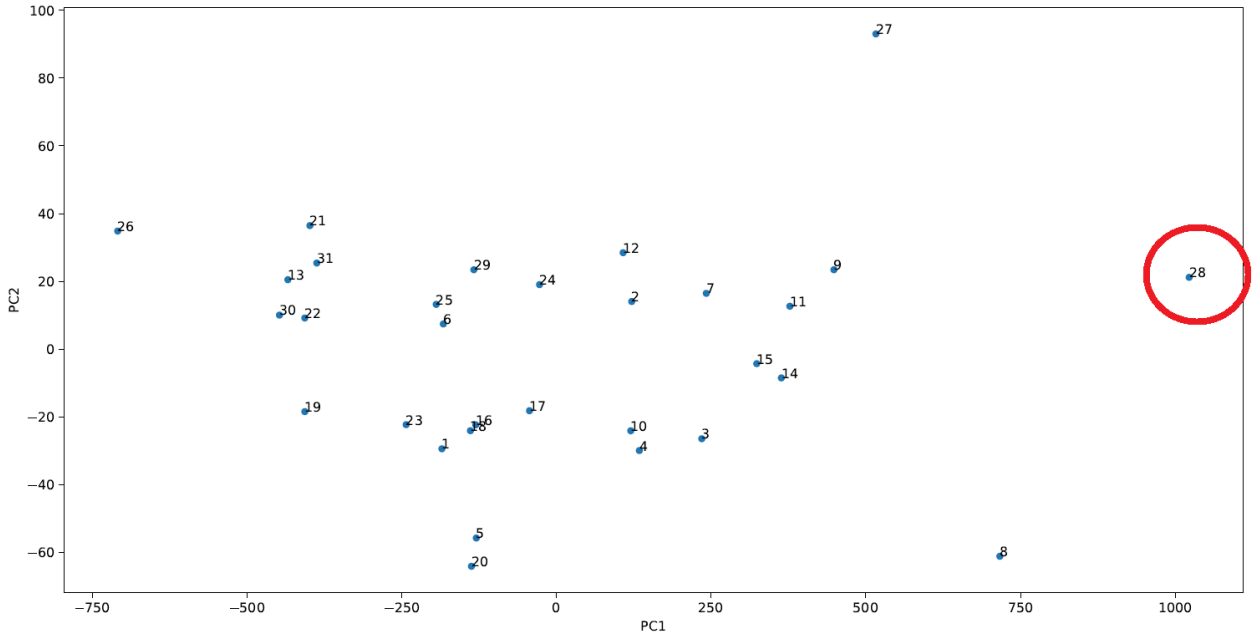

FIGURE 6. – résultat des analyses

Retourner au texte.

Liste des abréviations

ACP Analyse en Composante Principale. 3, 4, 6