LETTRES SORBONNE UNIVERSITÉ

Le 5 mai 2019

Analyse des sentiments sur Weibo

Analyse des avis négatifs à partir du scandale de ZHAI Tianlin

Projet Linguistique de corpus

SUN Chen et GAO Shuai
SORBONNE UNIVERSITÉ

Table des matières

1. Introduction

« C'est quoi le site CNKI ? »

ZHAI Tianlin, jeune acteur célèbre chinois, titulaire d'un diplôme doctoral, chercheur postdoctoral chez l'École de management de Guanghua affiliée à l'Université de Pékin, se demanda-t-il lors d'une retransmission en direct le 8 février. Deux jours plus tard, ce nom est devenu beaucoup plus connu que jamais, pas en tant qu'acteur mais pour son comportement frauduleux dans le milieu de la recherche. Pourquoi un propos au hasard entraîne des conséquences tellement graves ?

Son propos a été en effet très suspect car toute recherche en Chine ne peut jamais se passer du site qu'il ne connaissait pas. CNKI ou CNKI.net (China National Knowledge Infrastructure, Infrastructure de connaissance nationale de Chine) est un site initié par l'Université Tsinghua dans le but d'englober toutes les productions de connaissance telles que les journaux, les thèses doctorales, les mémoires de master et même de licence. Tout étudiant, qui a rédigé n'importe quel texte de recherche, doit se référer sur ce site est censé de le connaître. Il est donc surprenant que ZHAI ne le connaissait pas et la qualité de son diplôme se comprend de soi. D'ailleurs, sa thèse n'a jamais été publiée.

Ce propos n'est qu'un déclencheur. Il a été ensuite accusé d'avoir plagié dans son mémoire de Master, où il y a un taux de plagiat à 36,2 %, ce qui ne le permettrait sûrement pas d'avoir eu le droit de passer la soutenance. L'opinion publique en effervescence, l'Université de Pékin l'a virée de son poste de chercheur postdoctoral le 16 février. Trois jours plus tard, l'Université de cinéma de Pékin a décidé de retirer son diplôme de doctorat.

1.1. Raisons du choix

Bien que coupable, ZHAI Tianlin n'est, d'après certains, qu'un symptôme du système académique « corrompu ». Le système académique reste également responsable de ce scandale. Selon l'Annuaire des statistiques de l'éducation de Chine, plus de la moitié de doctorants n'arrivent pas à être diplômés à l'heure eu égard à la difficulté de rédaction de thèse, même s'ils y consacrent tout leur temps. Cependant, ZHAI a réussi à être diplômé tout en jouant des rôles dans pas mal de séries pour faire fortune pendant son doctorat.

On n'arrête pas à se demander comment cela a marché ? Comment il a pu être embauché comme chercheur postdoctoral dans une école illustre avec sa « thèse » de mauvaise qualité ? Est-ce à dire que c'est le jury qui n'a pas bien examiné strictement ses documents et a truqué le processus d'admission ? Toutes ces questions nous amènent à étudier à travers Weibo, le plus grand réseau social chinois, l'opinion publique sur le fort contraste entre la facilité du doctorat pour ZHAI et la difficulté du doctorat aux yeux du public.

1.2. Hypothèse

Nous supposons que le public ait tendance à critiquer le système académique au lieu d'individu car on n'aurait pas dû l'autoriser à terminer des études. En d'autres termes, ce mauvais environnement mérite plus des critiques que l'individu en question.

2. Intermède : rendez-vous avec M. Debailly

Un conseil nous a été donné lors de la première présentation. Eu égard au caractère sociologique de notre sujet, il nous convenait de nous renseigner auprès d'un sociologue. Nous avons rencontré donc M. Debailly, sociologue de notre faculté, qui s'intéressait à notre sujet et qui a élargi nos pistes de réflexion.

Il nous a parlé de l'ancienneté des données. Afin d'avoir des résultats plus fiables, il vaut mieux que les données recueillies soient les plus proches possibles de l'événement. Tous les sites ont, d'une certaine mesure, une profondeur historique. Si nous ne pouvions recueillir que les données des deux dernières semaines à cause de certaines restreintes du site, et que l'événement s'était passé deux mois avant, les résultats obtenus de notre analyse s'avéraient futiles.

C'était justement un mois après l'événement que nous avons pu commencer à recueillir les données une fois l'outil de *web scraping* prêt. D'où notre inquiétude de la question de l'ancienneté. On est parvenu à obtenir des données avec une profondeur historique, les commentaires les plus anciens datent du 14 février, quelques jours après l'événement. Mais on ne sait pas si certains commentaires importants ont déjà été supprimés pour des raisons politiques, comme ce qui se passe très souvent en Chine.

3. Méthodes

Notre projet se situe dans le cadre de l'analyse des sentiments où la notion de polarité serait a priori primordiale. Il convient normalement de prendre en compte ce que les gens disent de cet événement pour déterminer leurs sentiments exprimés. Néanmoins, à la différence de l'analyse des sentiments proprement dite, notre projet ne concerne pratiquement pas la polarité.

Force est de constater que tous les commentaires ne soient pas les critiques car il existe bien des soutiens ou des encouragements de la part des fans de l'individu, nous ne pouvons toujours pas les considérer comme non-négatifs. Même si les fans disent des choses plus ou moins positives pour réconforter ZHAI, individu submergé des critiques à l'heure actuelle, voire insultes, leurs paroles ne peuvent jamais nier le fait que son acte de plagiat est scandaleux. Aussi admettent-ils avant tout leurs sentiments négatifs ayant trait à son plagiat. En conséquence, peu importent leurs affinités avec l'individu. En ne nous fixant que sur l'événement en soi, nous pouvons affirmer donc que les sentiments exprimés restent cent pour cent négatifs par rapport à son plagiat. En l'occurrence, nous ne pouvons donc pas prendre en considération le sens explicite des commentaires ni leurs sentiments extérieurement exprimés car il s'agit tout à fait d'un scandale.

Notre hypothèse pourrait alors se justifier par la détection des entités nommées. Les occurrences détectées pourraient représenter l'attitude du public sur cette affaire. Si dans un corpus constitué des commentaires il existe plus d'occurrences des entités nommées désignant l'individu, nous dirions que le public reproche plutôt ZHAI que l'environnement et que ZHAI n'est qu'un cas isolé ; et s'il existe plus d'occurrences des entités nommées désignant l'environnement, nous dirions que le public pense que ZHAI est probablement un bout de l'iceberg et que c'est le système d'éducation qui mérite plus de critiques.

3.1. Constitution des dictionnaires

Pour ce faire, deux dictionnaires, qui désignent respectivement l'individu ZHAI Tianlin et l'environnement académique, nous sont nécessaires. Nous constituons ces deux dictionnaires à partir de ce que nous constatons en parcourant les commentaires et y ajoutons quelques éléments anticipés. Outre cela, nous prenons en considération également les commentaires qui ne sont pas écrits en chinois simplifié. Nous dupliquons

les éléments des dictionnaires en version traditionnelle car les utilisateurs de Hong Kong et de Taiwan écrivent en chinois traditionnel sur Weibo.

Voici les dictionnaires avec la traduction.

3.1.1. Individu

翟天临 (ZHAI Tianlin), 天临 (Tianlin), 老翟 (ZHAI avec un préfixe familier), 翟 (ZHAI), 您 (vous), 你 (toi), 他 (lui), ztl (abréviation du nom utilisé par certains internautes), 翟同学 (Camarade ZHAI), 翟天臨 (ZHAI Tianlin en chinois traditionnel), 天臨 (Tianlin en chinois traditionnel), 妳 (toi en chinois traditionnel), 翟同學 (Camarade ZHAI en chinois traditionnel)

3.1.2. Environnement

北大 (abréviation de l'Université de Pékin), 北京大学 (Université de Pékin), 北京大学光华管理学院 (École de management de Guanghua affiliée à l'Université de Pékin), 光华 (Guanghua), 光华管理学院 (École de management de Guanghua), 北电 (abréviation de l'Université de cinéma de Pékin), 北京电影学院 (Université de cinéma de Pékin), 电影学院 (Université de cinéma), 大学 (université), 学术圈 (monde intellectuel), 学校 (école), 导师 (directeur), 老师 (enseignant), 博导 (directeur de thèses), 陈浥 (CHEN Yi), 院长 (doyen de la faculté), 委员会 (commission), 答辩委员会 (commission de thèses), 论文审核组 (jury de thèse), 北京大學 (Université de Pékin en chinois traditionnel), 北京大學光華管理學院 (École de management de Guanghua affiliée à l'Université de Pékin en chinois traditionnel), 光華 (Guanghua en chinois traditionnel), 光華管理學院 (École de management de Guanghua en chinois traditionnel), 北電 (abréviation de l'Université de cinéma de Pékin en chinois traditionnel) , 北京電影學院 (Université de cinéma de Pékin en chinois traditionnel), 電影學院 (Université de cinéma en chinois traditionnel), 大學 (université en chinois traditionnel), 學術圈 (monde intellectuel en chinois traditionnel), 學校 (école en chinois traditionnel), 導師 (directeur en chinois traditionnel), 老師 (enseignant en chinois traditionnel), 博導 (directeur de thèses en chinois traditionnel), 陳浥 (CHEN Yi en chinois traditionnel), 院長 (doyen de la faculté en chinois traditionnel), 委員會 (commission en chinois traditionnel), 答辯委員會 (commission de thèses en chinois traditionnel), 論文審核組 (jury de thèses en chinois traditionnel)

3.2. Construction des corpus

Une fois les dictionnaires créés, nous nous prenons à construire le corpus. Cette étape s'avère être difficile puisqu'il s'agit du *web scraping*, lequel nous était complètement étranger. Tout d'abord, nous nous renseignons sur le CSDN (Chinese Software Developer Network), le plus grand site consacré aux développeurs chinois. Nous y trouvons plein de procédés possibles pour recueillir des données sur Weibo, mais certains d'entre eux nous étaient pourtant peu accessibles.

Des codes en Python à Octopus[1], de Houyi[2] à Jiguang[3], nous avons essayé plusieurs outils de *web scraping*, mais nous n'avons pas réussi à recueillir de données propres et structurées jusqu'à ce que l'on rencontre GooSeeker. Ce dernier est une plateforme à multiples fonctionnalités, dont l'une spéciale dédiée à recueillir les commentaires sur *Weibo*. Les données recueillies sont bien structurées. Nous n'avons donc pas besoin d'effectuer le détourage.

(Interface graphique de GooSeeker)

Nous avons construit six corpus à partir des pages *Weibo* de ces six comptes importants ayant trait à cet événement :

[1] 八抓鱼, outil de web scraping chinois, disponible sur : https://www.bazhuayu.com/

[2] 后羿, outil de web scraping chinois, disponible sur : http://www.houyicaiji.com/

[3] 极光, outil de web scraping chinois, disponible sur : https://www.jiguang.cn/izone

1. 翟天临 (compte personnel de ZHAI Tianlin)
2. 北京大学 (compte officiel de l'Université de Pékin)
3. 北京电影学院 (compte officiel de l'Université de cinéma de Pékin)
4. 人民日报 (compte officiel du *Quotidien du peuple*, presse écrite chinoise)
5. 新浪娱乐 (compte officiel du Divertissement de *SINA*)
6. 娱乐圈神评 (compte officiel d'un self-média connu)

Les nombres de commentaires recueillis de chaque compte ne sont pas équilibrés en raison de la qualité de connexion et de la profondeur autorisée du site. Voici ci-dessous un tableau qui résume les données recueillies.

Compte *Weibo*	Nombre de commentaires recueillis
翟天临	4825
北京大学	288
北京电影学院	5637
人民日报	492
新浪娱乐	339
娱乐圈神评	2148

4. Résultats

Nous avons écrit le code ci-dessous et l'avons exécuté pour tous les six corpus.

```
# Analyse effectuée sur la page Weibo de l'individu, ZHAI Tianlin
import pandas as pd
df = pd.read_excel('翟天临.xlsx')
df = df.fillna("")

dict_a = ["翟天临", "天临", "老翟", "翟", "ztl", "翟同学",
          "您", "你", "他",
          "翟天臨", "天臨", "翟同學",
          "妳",]
# Le "dict_a" contient des variations possibles des appellations pour l'individu que nous prédisons.

dict_b = ["北大", "北京大学", "北京大学光华管理学院", "光华", "光华管理学院", "北电", "北京电影学院", "电影学院",
          "大学", "学术圈", "学校",
          "导师", "老师", "博导", "陈浥",
          "院长", "委员会", "答辩委员会", "论文审核组",
          "北京大學", "北京大學光華管理學院", "光華", "光華管理學院", "北電", "北京電影學院", "電影學院",
          "大學", "學術圈", "學校",
          "導師", "老師", "博導", "陳浥",
          "院長", "委員會", "答辯委員會", "論文審核組"]
# Le "dict_b" contient des variations possibles des appellations pour l'environnement que nous prédisons.

count_a = 0
count_b = 0
for i in range(len(df)):
    for j in range(len(dict_a)):
        if dict_a[j] in df.loc[i,'评论内容']:
            count_a += 1
    for j in range(len(dict_b)):
        if dict_b[j] in df.loc[i,'评论内容']:
            count_b += 1

print(count_a)
print(count_b)
```

```
4046
297
```

Et nous avons obtenu les résultats comme suit :

Compte *Weibo*	Individu	Environnement
ZHAI Tianlin	4046	297
Université de Pékin	89	113
Université de cinéma de Pékin	2854	2071
Quotidien du peuple	131	149
Divertissement de *SINA*	183	125
Self-média connu	847	46
Total	8150	2801

Les résultats en vert prouvent que notre hypothèse est vraie alors que ceux en rouge, fausse. Nous n'avons que deux résultats qui soutiennent notre hypothèse et les deux chiffres de ces deux pairs sont très proches, ce qui ne revêt pas trop d'importance.

On peut en conséquence affirmer que l'opinion publique critique plus l'individu que l'environnement.

5. Discussion

Notre méthode ayant fonctionné et l'hypothèse prouvée fausse, notre projet laisse pourtant encore à désirer.

5.1. Limites

Premièrement, notre méthode n'est pas parfaitement fiable. S'il on s'en tient à ce seul exemple, un utilisateur de *Weibo* a écrit sur la page de l'individu : « (Tu) as déshonoré l'Université de Pékin ! ». Dans cette phrase, le sujet « tu » a été omis. En l'occurrence, l'entité nommée détectée appartient au dictionnaire représentant l'environnement alors que l'on sait bien qu'il critique l'individu. On a donc une erreur comme résultat.

Deuxièmement, les six corpus ne sont pas équilibrés à cause de conditions de recueil quelque peu favorables. De plus, on n'a pas pu commencer le recueil à temps. On ne sait pas donc si certains commentaires, de grande quantité peut-être, ont été supprimés car ils avaient menacé la face de l'environnement académique tel que l'Université de Pékin, université illustre de rang mondial.

Troisièmement, dans les corpus classés suivant le compte de Weibo, nous constatons que le compte oriente, d'une façon ou une autre, l'objet des critiques dans les

commentaires. Généralement, le public aurait tendance à critiquer l'objet en lien avec le compte. Par exemple, dans les posts de ZHAI, on critique plus lui-même par rapport aux établissements. En revanche, dans la page de l'Université de Pékin, le public a visé l'établissement ainsi que le monde académique.

5.2. Perspectives

Somme toute, notre sujet basé sur l'affaire de plagiat de ZHAI Tianlin concerne non seulement une fraude académique, mais aussi la question de l'égalité éducatrice, la qualité de l'enseignement supérieur et la corruption du monde académique. Ce genre de projet mérite bien d'être développé par autrui prenant en compte des limites précitées. Pour avoir des résultats plus précis, il conviendrait d'utiliser un outil *web scraping* plus puissant pour recueillir les plus possibles des données et d'en proposer une méthode avec des critères entraînant pas de mauvaises interprétations des données.

Nous souhaitons que ce scandale puisse devenir un tournant pour la société chinoise faisant réfléchir le public à notre éducation supérieure et aux problèmes sociaux concernés.

6. Références

- https://www.weibo.com/aj/v6/comment/big?ajwvr=6&id=4339602683247836&root_comment_max_id=865885525798421&__rnd=1553476859375&page=1
- https://baijiahao.baidu.com/s?id=1627083838390708378&wfr=spider&for=pc
- http://www.xinhuanet.com/comments/2019-02/13/c_1124107413.htm