ESPADON

Aymeric Luneau

Aug 02, 2022

CONTENTS

0.1 Détection des biomarqueurs dans les tweets

0.1.1 Méthode initiale

Afin de compter les occurrences des biomarqueurs dans les tweets, la méthode initiale consistait à rechercher simplement les chaînes de caractères correspondant aux noms des biomarqueurs en prenant ou non en compte la casse. Pour rappel, les biomarqueurs pris en compte sont: *ROS1, ALK, EXON, EGFR, KRAS, NTRK, BRAF, MET, RET, HER2*.

Toutefois, cette première méthode conduisait à prendre en compte de nombreux faux-amis dans le cas de certains biomarqueurs. Par exemple, suivant cette méthode, “RETweet”, “WALK”, “VincenTRK” ou “METASTASIS” étaient compté comme une occurrence de “RET”, “ALK”, “NTRK” ou “MET” respectivement.

0.1.2 Définition d’expressions régulières plus complexes

Des expressions régulières plus “complexes” ont alors été définies pour exclure les faux-amis évoqués ci-dessus et d’autres du décomptes des occurrences. Le script utilisé et affiché dans la cellule ci-dessous fonctionne de la façon suivante:

1. Pour chaque tweet, on commence par remplacer plusieurs signes de ponctuations (“/”, “.”, “,”, “-”) par des espaces.
2. Une fois le remplacement effectué, chaque tweets est découpé afin d’obtenir la liste de toutes les chaînes de caractère précédées et suivies par un espace (dans le script ci-dessous, cette liste est nommée “tt”).
3. puis on recherche la présence de chaque biormaqueur au sein de cette liste selon des expressions régulières. Par exemple, dans le cas du biomarqueur “MET”, nous avons utilisé l’expression régulière suivante (avec x est égale à “MET” et le signe “|” qui signifie “ou”):

```
"{x}[(+)]" | "[#@+]{x}[^AaSsUuEeYyHh]" | "[#@+]{x}$" | "^{x}[^AaSsUuEeYyHh]" |
"^{x}$"
```

C’est-à-dire que pour chaque terme de la liste “tt”, celui-ci sera considéré comme une occurrence du biomarqueur “MET” si, et seulement si “MET” :

- est suivi du signe +”
- **ou** est précédé par un hashtag ou un arobase et suivi par un ou plusieurs caractères sauf les lettres majuscules ou minuscules *a, s, u, e, y, h*.
- **ou** est précédé par un hashtag, un arobase ou le signe “+” et n’est suivi par aucun caractères
- **ou** n’est précédé par aucun caractère et n’est pas suivi par les lettres *a, s, u, e, y, h* majuscules ou minuscules”. Dans ce cas, l’expression “METUProg” n’est pas prise en compte à la différence de “METmuts”
- **ou** n’est précédé ni suivi de caractères.

Si certaines règles sont communes à l’ensemble des biomarqueurs, d’autres sont spécifiques. Par exemple, dans le cas du biomarqueur “RET”, les lettres *a, s, u, e, y, h* sont remplacées par *a, e, i, o, u, r, h, z, w* pour éviter que des mots comme “retweet” soient comptés. À noter également que dans le cas des biomarqueurs “ALK”, “MET” et “RET”, on ne prend en compte que les formes en majuscule. Pour les autres, l’algorithme de recherche est insensible à la casse.

Enfin, ces expressions régulières sont recherchées à l’aide de la fonction “search” du module *re* du langage Python. Cette fonction retourne “vrai” si l’expression est trouvée, “faux” dans le cas contraire. Par ailleurs, elle s’arrête à la première expression trouvée. Autrement dit, la fonction indique simplement si l’expression est présente quelque soit le nombre de

fois où elle apparaît dans le tweet. Le tableau ci-après illustre le résultat obtenu pour une dizaine de phrases avec chacune des deux méthodes (sans ou avec les expressions régulières).

La principale limite de cette seconde méthode est qu'elle conduit à définir de nombreuses exceptions pour éviter que des "vrais positifs" soit traités comme des "faux positifs". Ainsi, la règle qui consiste à ne pas compter les chaînes de caractère constituées de MET suivi d'un "a", comme *METastasis*, a pour effet également d'exclure les formes du types *METamp*. La liste des formes exclues est placée à la fin de ce document.

Illustration de l'algorythme

1. Le tweet initial :

   "MET copy number as a secondary driver of EGFR TKI resistance in EGFR-mutant NSCLC http://bit.ly/2HfGHjn #editorial #lcsm"

1. Le tweet nettoyé (les "/", ".", "," et "-" sont remplacés par des espaces) :

   "MET copy number as a secondary driver of EGFR TKI resistance in EGFR mutant NSCLC http: bit ly 2HfGHjn #editorial #lcsm"

1. On crée une liste de toutes les chaînes de caractère précédées et suivies par une espace :

   ['MET', 'copy', 'number', 'as', 'a', 'secondary', 'driver', 'of', 'EGFR', 'TKI', 'resistance', 'in', 'EGFR', 'mutant', 'NSCLC', 'http:', '', 'bit', 'ly', '2HfGHjn', '#editorial', '#lcsm']

1. La recherche du biomarqueur se fait ensuite à partir de la liste des termes entre crochets ci-dessus. Dans cet exemple, on a bien le terme "MET". On constate que le tweet fait égalment référence au biomarqueur EGFR.

```
<pandas.io.formats.style.Styler at 0x7f6bb38498b0>
```

Le tableau ci-dessous donne le nombre d'occurrences retrouvées pour chacun des marqueurs avec les deux méthodes.

```
<pandas.io.formats.style.Styler at 0x7f6bc2c8ed90>
```

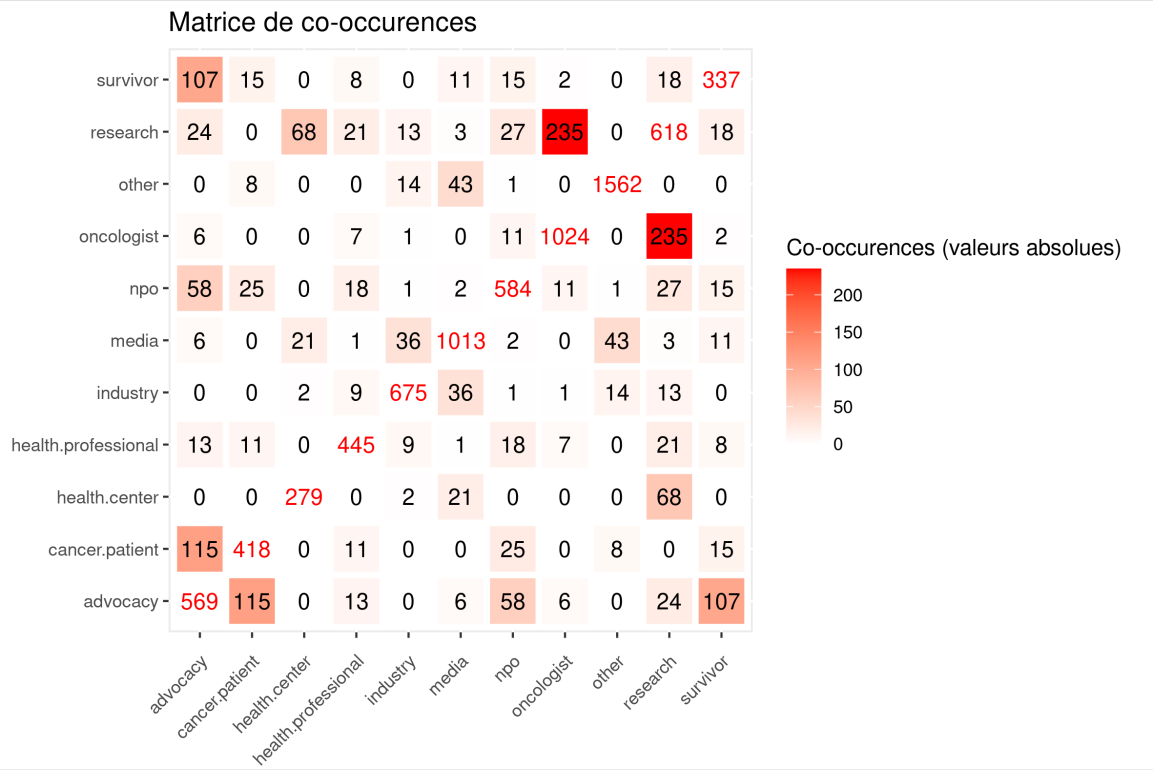
```
<pandas.io.formats.style.Styler at 0x7f6b44447af0>
```

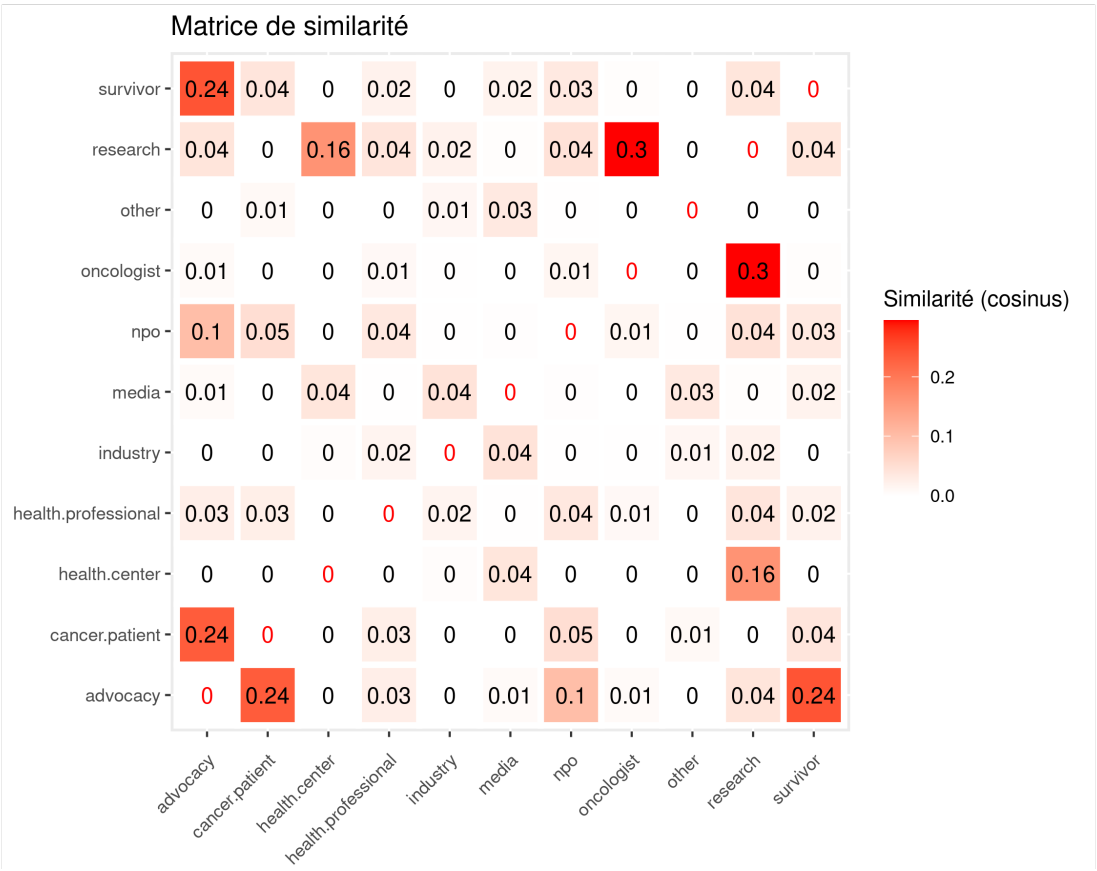
0.2 Regroupement des rôles

S'agissant des rôles et suivant les suggestions de la réunion précédente, plusieurs niveaux de regroupements ont été expérimentés. J'ai d'abord distingué les rôles caractérisant essentiellement des "personnes" (*advocacy, survivor, cancer patient, oncologist, health professional* et *research*) de ceux définissant des "collectifs" (*media, npo, industry, health center*). Toutefois, comme on le verra plus loin, il existe des collectifs codés comme "patients" ou "oncologues" et, réciproquement, des "personnes" codées comme "media" ou "npo".

Je me suis ensuite appuyer sur une analyse des "similarités" entre les rôles pour définir les rapprochements pertinent. Par "similarité", j'entends le fait que 2 rôles ou plus soient associés à un même compte. Par exemple, on a 235 comptes qui ont été annotés à la fois comme oncologues (*oncologist*) et chercheurs (*research*).

La figure X indique pour chaque couple de rôles le nombre d'individus qu'ils ont en commun (matrice de co-occurrence) et leur degré de similarité (matrice de similarité). On observe une certaine "similarité" entre les "défenseurs de cause" (*advocacy*), les "survivants" (*survivor*) et les "patients" (*cancer patient*) d'une part, les "chercheurs" (*research*) et "oncologues" (*oncologist*) d'autre part qui justifient leur rapprochement au sein d'une même catégorie. En revanche, on peut s'interroger sur l'intérêt de regrouper les "professionnels de santé" (*health professional*) avec les "oncologues" et les "chercheurs", étant donnée leur faible similarité. De même, le faible nombre de chevauchements entre les rôles collectifs plaide pour que ces derniers continuent d'être distinguer.


Matrice de co-occurences
survivor 107 15 0 8 0 11 15 2 0 18 337
research 24 0 68 21 13 3 27 235 0 618 18
other 0 8 0 0 14 43 1 0 1562 0 0
oncologist 6 0 0 7 1 0 11 1024 0 235 2
npo 58 25 0 18 1 2 584 11 1 27 15
media 6 0 21 1 36 1013 2 0 43 3 11
industry 0 0 2 9 675 36 1 1 14 13 0
health.professional 13 11 0 445 9 1 18 7 0 21 8
health.center 0 0 279 0 2 21 0 0 0 68 0
cancer.patient 115 418 0 11 0 0 25 0 8 0 15
advocacy 569 115 0 13 0 6 58 6 0 24 107
advocacy cancer.patient health.center health.professional industry media npo oncologist other research survivor
Co-occurences (valeurs absolues)
200
150
100
50
0


Matrice de similarité
survivor 0.24 0.04 0 0.02 0 0.02 0.03 0 0 0.04 0
research 0.04 0 0.16 0.04 0.02 0 0.04 0.3 0 0 0.04
other 0 0.01 0 0 0.01 0.03 0 0 0 0 0
oncologist 0.01 0 0 0.01 0 0 0.01 0 0 0.3 0
npo 0.1 0.05 0 0.04 0 0 0 0.01 0 0.04 0.03
media 0.01 0 0.04 0 0.04 0 0 0 0.03 0 0.02
industry 0 0 0 0.02 0 0.04 0 0 0.01 0.02 0
health.professional 0.03 0.03 0 0 0.02 0 0.04 0.01 0 0.04 0.02
health.center 0 0 0 0 0 0.04 0 0 0 0.16 0
cancer.patient 0.24 0 0 0.03 0 0 0.05 0 0.01 0 0.04
advocacy 0 0.24 0 0.03 0 0.01 0.1 0.01 0 0.04 0.24
advocacy cancer.patient health.center health.professional industry media npo oncologist other research survivor
Similarité (cosinus)
0.2
0.1
0.0

Le travail de recodage a alors été divisé en plusieurs étapes afin de construire de nouvelles variables permettant d'attribuer chaque compte à une catégorie unique tout en conservant les rôles définis lors de l'annotation. De cette manière, il est possible ensuite de tester différents niveaux de regroupement.

- Etape 1 : j'ai crée une première variable intutilée *User_role*. Elle comprend à la fois les rôles "purs", c'est-à-dire les comptes jouant un seul rôle, et l'ensemble des associations de rôles observés existantes. Si un compte a uniquement été classé comme *cancer patient*, il conserve cette valeur. En revanche, si un compte est à la fois classé comme *cancer patient* et *advocacy*, alors il prendra la valeur *Cancer patient & advocacy*.
- Etape 2 : les modalités de la variable *User_role* sont regroupées à l'aide d'une nouvelle variable appelée *User_role2*. Il s'agit ici d'opérer une première réduction des rôles. J'ai fait par exemple le choix de regrouper toutes les modalités contenant le terme "survivor" dans une méta-catégorie appelée *Survivor* (avec S majuscule). De même toutes les modalités contenant le terme "cancer patient" dans la méta-catégorie *Cancer patient*, sauf celles contenant aussi le terme "survivor".
- Etape 3 : La variable *User_status* réunit à son tour les modalités de la variable *User_role2*. Dans cette dernière variable, les modalités *Cancer patient* et *Survivor* sont réunies au sein de la catégorie générale *Patients*. Tandis que la catégorie *Health professionals* rassemblent les "oncologues", les "professionnels de santé" et les "chercheurs". En ce qui concerne les collectifs, j'ai regroupés les "centres de santé" et les "industries" dans une catégorie appelée "Health organisations". J'ai par ailleurs fait le choix de conserver une catégorie "média" et "NPO" dès lors que le comptes n'appartiennent pas à la classe des "industries" et des "centres de santé".

```
<pandas.io.formats.style.Styler at 0x7f3fa86a84f0>
```

0.3 Quelques graphiques

Les différents graphiques qui suivent donne un aperçu de la distribution dans le temps des rôles et des références aux différents biomarqueurs.

0.3.1 Des professionnels de plus en plus présents

L'analyse de la distribution des rôles montre la présence croissante des professionnels, en particulier des oncologues (*oncologists*) et, dans une moindre, des chercheurs. Alors que les oncologues représentent un peu moins de 5% des comptes en 2012, ils constituent environ 27% des comptes en 2020. La part des professionnel de santé (*health professional*) semble quant à elle relativement stable dans le temps. La proportion de comptes jouant les rôles de patients, c'est-à-dire les *survivors* et les *cancer patient*, ne dépasse pas les 10% sur toutes la périodes. On observe également une faible représentation des "défenseurs de causes" (*advocacy*). Rappelons toutefois que cette catégorie comprend les comptes qui ne jouent pas d'autres rôles. Par exemple, un "survivant" qui a été annoté également comme un défenseur de cause (*advocacy*) sera compté parmi les patients et non les *advocacy*.

Les deux diagrammes ci-dessous représentent respectivement la proportion de comptes et de tweets par statut et par an. Les "statuts" correspondent aux modalités de la variable *User_status*. On voit ainsi que les médias constituent moins de 10% des comptes en 2021, mais sont à l'origine de plus de 25% des tweets à la même époque.

Note: La variable *User_status* est une "réduction" de la variable *User_role2*. Ainsi, la modalité "Health professionals"

regroupe les rôles d'oncologues, de chercheurs et de professionnels de la santé (hors médecins).

0.3.2 La dynamique des biomarqueurs

Les diagrammes ci-dessous donnent la proportion de tweets contenant chacun des biomarqueurs par an et par mois respectivement, sachant qu'un tweet peut contenir plusieurs biomarqueurs. Ainsi, en 2018, le biomarqueur ALK était présent dans près 1,5% des tweets. Puis en janvier 2021, le biomarqueur EGFR était présent dans 3% des tweets.

0.3.3 Qui parle de quel marqueur ?

Enfin, la matrice ci-après indique la part occupée par les différents biomarqueurs dans les tweets qui en mentionnent au moins un en fonction du statut de leurs auteurs. Par exemple, 30% des 2548 biomarqueurs cités par les "défenseurs de cause" (*advocacy*) concernent le marqueur EGFR.

```
<pandas.io.formats.style.Styler at 0x7f67a223e4c0>
```