Projet TALN

Analyse des données

Répartition des données

figure 1 et *figure 2*:

On peut voir que l'ensemble d'entraînement et l'ensemble de test on a peu près la même distribution au niveau des "comment" (69% à 67%). Par contre il y a proportionnellement deux fois plus d'élément dans pour les "deny" du "train" que pour le "test".

Les sujets sont répartis de manière inéquitable entre le "test" et le "train". En effet il y a 7 sujets pour le "train" et 1 pour le "test" (germanwings-crash). On ne devra alors pas se baser sur le vocabulaire spécifique à chaque pour dissocier les catégories. On devra plutôt utiliser les relations que les mots entretiennent entre eux.

Train's Subjects:

La répartion des différentes catégories varient en fonction des sujets:

subject	comment	deny	query	support
charliehebdo	72%	6%	5%	17%
ebola-essein	66%	19%	3%	12%
ferguson	69%	9%	10%	13%
ottawashooting	66%	10%	9%	15%
prince-toronto	68%	7%	12%	13%
putinmissing	60%	9%	9%	21%
sydneysiege	68%	8%	9%	14%

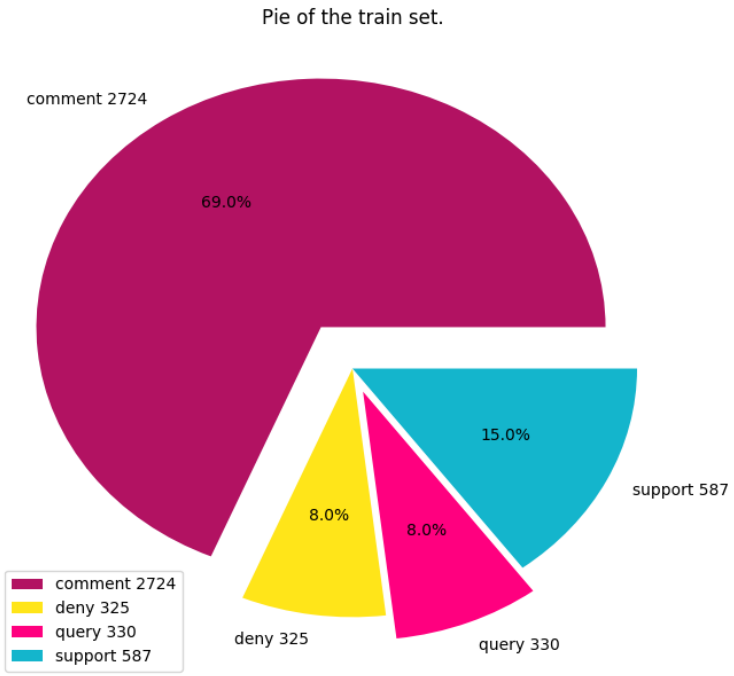

Figure 1: Répartition des données d'entraînement

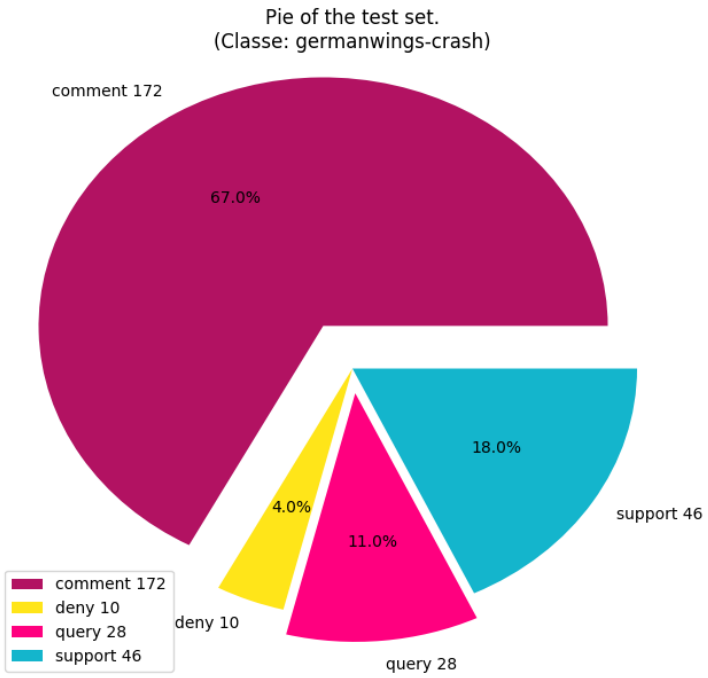

Figure 2: Répartition des données de test

Projet TALN

Analyse des données

Mesures

Pour voir si la tâche était difficile nous avons fait un algorithme Naive Bayes sur la fréquence des mots par catégories pour distinguer les différentes réponses de tweets. Nous avons fait varier le numérateur (Num) appliqué pour les les mots inconnus (UNK) sur des puissance négative de 10.

Num UNK	Accuracy
0.1	20.31%
0.01	28.52%
0.001	35.94%
0.0001	40.23%
1e-05	44.53%
1e-06	47.27%
1e-07	47.66%
1e-08	51.17%
1e-09	54.3%
1e-10	55.47%
1e-11	55.47%
inf.	conv = 55.47%

Plus on diminue Num UNK plus on a une grande exactitude. Est-ce que ce la est bien ? On élimine certes les mots qui sont sans rapport pour la catégorie (pas sur)…

Num UNK: 0.01

F-score

comment	deny	query	support
54.14%	20.27%	39.37%	24.1%

`Accuracy: 28.52%`

Num UNK: 1e-11

F-score

comment	deny	query	support
83.09%	16.08%	43.92%	17.77%

`Accuracy: 55.47%`

Num UNK: 0.01

Truth Prediction

P T	comment	deny	query	support
comment	50	54	54	14
deny	0	3	4	3
query	5	3	15	5
support	9	24	8	5

Precision

comment	deny	query	support
88.12%	13.57%	28.52%	28.52%

Recall

comment	deny	query	support
39.07%	40.0%	63.57%	20.87%

F-score

comment	deny	query	support
54.14%	20.27%	39.37%	24.1%

Accuracy: 28.52%

Num UNK: 0.001

Truth Prediction

P T	comment	deny	query	support
comment	72	38	44	18
deny	0	3	4	3
query	9	3	13	3
support	16	20	6	4

Precision

comment	deny	query	support
84.23%	14.69%	29.4%	24.29%

Recall

comment	deny	query	support
51.86%	40.0%	56.43%	18.7%

F-score

comment	deny	query	support
64.19%	21.49%	38.66%	21.13%

Accuracy: 35.94%

Num UNK: 0.0001

Truth Prediction

P T	comment	deny	query	support
comment	86	31	40	15
deny	1	3	3	3
query	11	3	11	3
support	22	15	6	3

Precision

comment	deny	query	support
81.67%	15.77%	28.33%	22.5%

comment	deny	query	support

Recall

comment	deny	query	support
60.0%	40.0%	49.29%	16.52%

F-score

comment	deny	query	support
69.18%	22.62%	35.98%	19.05%

Accuracy: 40.23%

Num UNK: 1e-05

Truth Prediction

P T	comment	deny	query	support
comment	100	28	31	13
deny	3	2	2	3
query	11	3	10	4
support	23	15	6	2

Precision

comment	deny	query	support
82.99%	14.17%	30.41%	19.09%

Recall

comment	deny	query	support
68.14%	30.0%	45.71%	14.35%

F-score

comment	deny	query	support
74.84%	19.25%	36.52%	16.38%

Accuracy: 44.53%

Num UNK: 1e-06

Truth Prediction

P T	comment	deny	query	support
comment	107	26	30	9
deny	4	2	1	3
query	11	3	10	4
support	25	14	5	2

Precision

comment	deny	query	support
82.79%	14.44%	31.74%	21.11%

Recall

comment	deny	query	support
72.21%	30.0%	45.71%	14.35%

F-score

comment	deny	query	support
77.14%	19.5%	37.47%	17.08%

Accuracy: 47.27%

Num UNK: 1e-07

Truth Prediction

P T	comment	deny	query	support
comment	109	26	29	8
deny	6	1	0	3
query	12	3	10	3
support	29	11	4	2

Precision

comment	deny	query	support
79.87%	12.44%	33.26%	22.5%

Recall

comment	deny	query	support
73.37%	20.0%	45.71%	14.35%

F-score

comment	deny	query	support
76.48%	15.34%	38.5%	17.52%

Accuracy: 47.66%

Num UNK: 1e-08

Truth Prediction

P T	comment	deny	query	support
comment	118	23	23	8
deny	6	1	0	3
query	12	3	10	3
support	31	9	4	2

Precision

comment	deny	query	support
80.66%	12.78%	37.03%	22.5%

comment	deny	query	support

Recall

comment	deny	query	support
78.6%	20.0%	45.71%	14.35%

F-score

comment	deny	query	support
79.62%	15.59%	40.91%	17.52%

Accuracy: 51.17%

Num UNK: 1e-09

Truth Prediction

P T	comment	deny	query	support
comment	126	21	18	7
deny	6	1	0	3
query	12	3	10	3
support	33	7	4	2

Precision

comment	deny	query	support
81.19%	13.12%	41.25%	23.33%

Recall

comment	deny	query	support
83.26%	20.0%	45.71%	14.35%

F-score

comment	deny	query	support
82.21%	15.85%	43.37%	17.77%

Accuracy: 54.3% — ## Num UNK: 1e-10

Truth Prediction

P T	comment	deny	query	support
comment	129	18	18	7
deny	6	1	0	3
query	12	3	10	3
support	34	7	3	2

Precision

comment	deny	query	support
81.27%	13.45%	42.26%	23.33%

Recall

comment	deny	query	support
85.0%	20.0%	45.71%	14.35%

F-score

comment	deny	query	support
83.09%	16.08%	43.92%	17.77%

Accuracy: 55.47%

Num UNK: 1e-11

Truth Prediction

P T	comment	deny	query	support
comment	129	18	18	7

P T	comment	deny	query	support
deny	6	1	0	3
query	12	3	10	3
support	34	7	3	2

Precision

comment	deny	query	support
81.27%	13.45%	42.26%	23.33%

Recall

comment	deny	query	support
85.0%	20.0%	45.71%	14.35%

F-score

comment	deny	query	support
83.09%	16.08%	43.92%	17.77%

Accuracy: 55.47%

Projet TALN

Analyse des données

Distibution des mots par catégories

Ici nous avons représenter les distributions zipfienne de la fréquence des mots par catégories ordonné du plus au moins fréquent. Sans grande surprise “the” est le mot au premier rang toutes catégories confondues. “the” n’est pas un trait distinctif, par contre tous les mots suivants sont différents ou n’ont pas le même rang selon la catégories.

Comment

figure 1 Le vocabulaire qui ressort en majorité des comment est un vocabulaire assez usuel.

Deny

figure 2 On voit des mots de négation et un vocabulaire d’interpellation se démarquer.

Query

figure 3 Encore plus que la catégorie “deny”: On voit clairement des mots interrogatifs et un vocabulaire d’interpellation se démarquer.

Support

figure 4 Le support n’ont pas l’air d’avoir un vocabulaire clairement propre. Il y a beaucoup de mots usuels (comme pour les “comment”).
Ce qui va nous demander de trouver de nouveaux traits distinctifs.

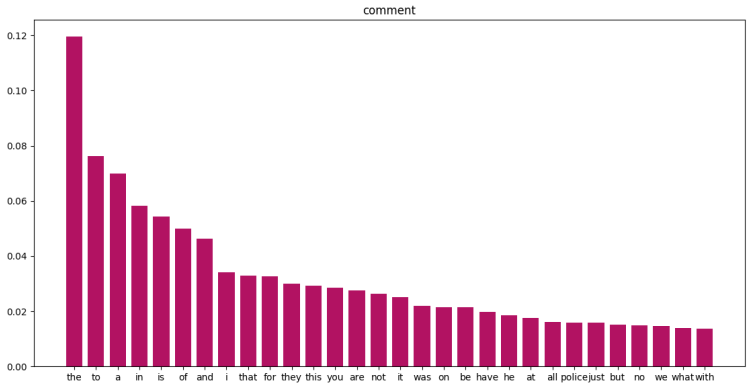

Figure 1: Distibution des mots de réponses catégorie ”comment”

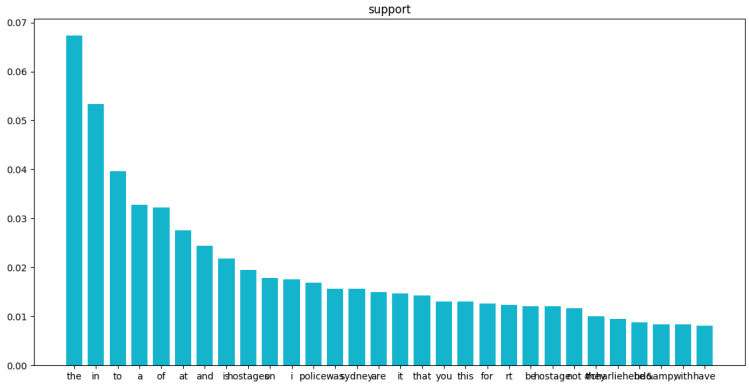

Figure 2: Distibution des mots de réponses catégorie ”support”

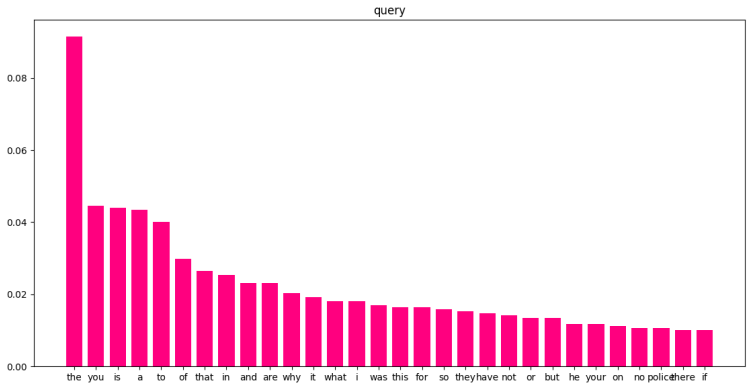

Figure 3: Distibution des mots de réponses catégorie ”query”

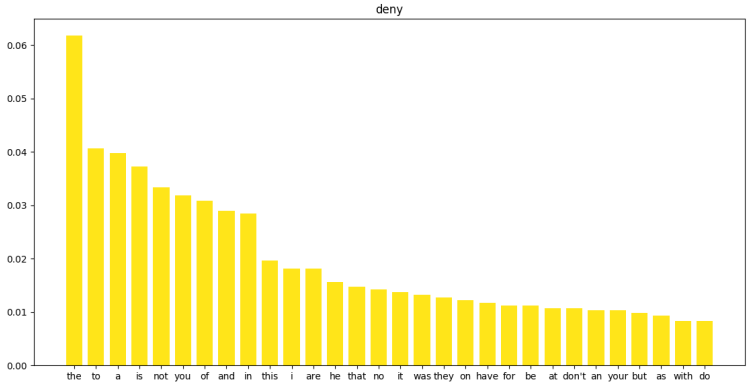

Figure 4: Distibution des mots de réponses catégorie ”deny”