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.

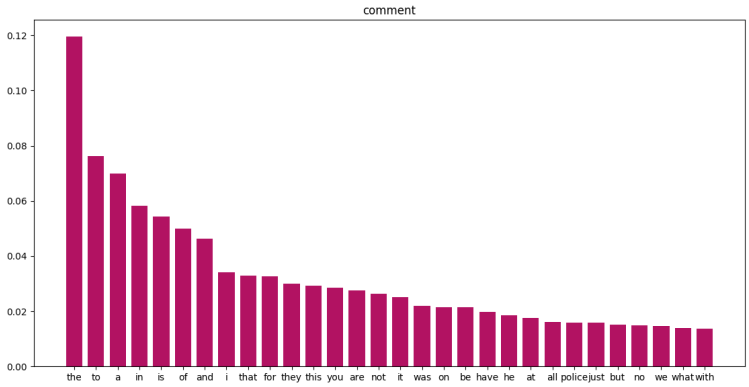

Figure 1: Distibution des mots de réponses catégorie ”comment”

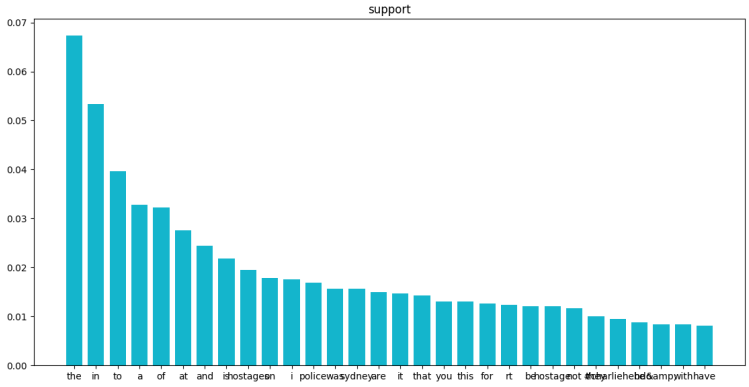

Figure 2: Distibution des mots de réponses catégorie ”support”

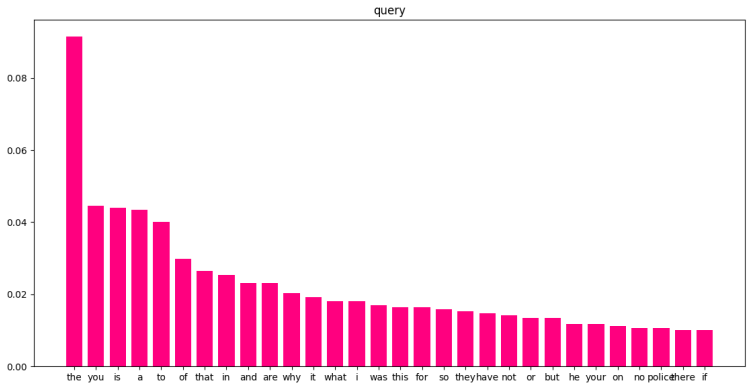

Figure 3: Distibution des mots de réponses catégorie ”query”

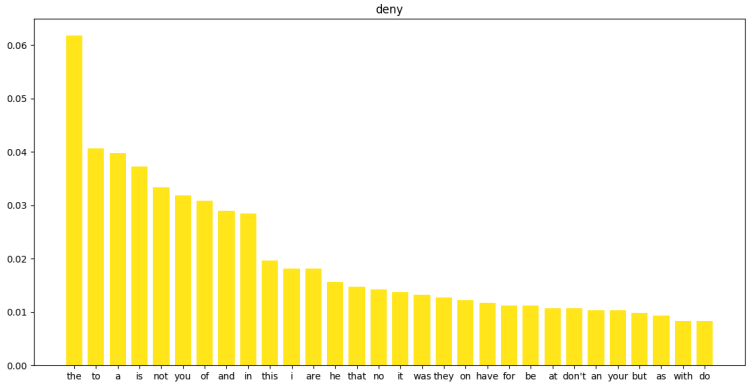

Figure 4: Distibution des mots de réponses catégorie ”deny”