## Manipulations : mesure de l'accord inter-annotateur

1. Annoter, dans le tableur dont le numéro vous a été attribué, les extraits en remplissant la colonne annotateur 1 (M(ono-négatif), B(i-négatif), A(mbigu), N(/A)).

2. En formant un groupe de 2 personnes, procédez à une "adjudication" pour les extraits que vous avez chacun annotés. Créer un nouveau tableur qui contient les 50 extraits annotés, avec une colonne pour chaque annotateur (ne regroupez pas vos annotations dans la même colonne) et ajoutez une colonne ("gold") pour le résultat de vos discussions.

3. Ajoutez dans le tableur une colonne pour l'accord entre la colonne Gold et la colonne d'annotation : faire en sorte que la case contienne 1 en cas d'accord, 0 sinon [1]. En déduire une mesure d'accord global (moyen) pour chaque annotateur. Exemple (factice) :

| *Extrait* | *Ann.1* | *Ann.2* | *Gold* | *Accord* | |
|---|---|---|---|---|---|
| Personne n'a rien dit | M | | M | 1 | |
| Aucun nouveau n'écoute rien | B | | A | 0 | |
| Rien ni personne ne me feront reculer | N | | N | 1 | 2/3 (66 %) |
| Il écouta. Rien, aucun bruit. | | N | N | 1 | |
| Max ne dit jamais rien | | B | A | 0 | |
| Personne ne va nulle part. | | M | B | 0 | 1/3 (33 %) |
| | | | | | 3/6 (50 %) |

4. Ajouter une feuille dans le tableur pour créer, pour chaque annotateur, une *matrice de contingence* : dans cette table on compte le nombre de situations différentes qu'on peut rencontrer. Par exemple les cas où l'annotateur choisit M mais la valeur "gold" est A (case en rouge) ; ou les cas où l'annotateur et la référence sont d'accord (cases de la diagonale).

| | | *Gold* | | | | |
|---|---|---|---|---|---|---|
| | | M | B | A | N | |
| *ann. 1* | M | 10 | 0 | 1 | 0 | 11 |
| | B | 1 | 3 | 1 | 0 | 5 |
| | A | 2 | 0 | 3 | 1 | 3 |
| | N | 1 | 0 | 0 | 2 | 3 |
| | | 14 | 3 | 5 | 3 | (25) |

Si on fait la somme des valeurs sur la diagonale on obtient un nombre qu'on peut diviser par le nombre total d'extraits (25 ici) pour obtenir une exactitude globale : $\frac{10+3+3+2}{25} = 68\%$.

5. En prenant comme référence l'annotation "Gold", on peut calculer, pour chaque annotateur, et pour chaque catégorie, une *précision* et un *rappel*. Soit $\gamma$ une catégorie. Soit $h_\gamma$ (*hits*) le nombre de cas où l'annotateur a correctement attribué la catégorie $\gamma$ à un extrait, soit $m_\gamma$ le nombre de cas où l'annotateur a attribué une autre catégorie à un extrait catégorisé $\gamma$ dans la référence (*miss*), soit $f_\gamma$ le nombre de cas où l'annotateur a attribué la catégorie $\gamma$ alors que l'extrait est d'une autre catégorie (*false positive*). La **précision** pour la catégorie $\gamma$ s'obtient en calculant le rapport $\frac{h_\gamma}{h_\gamma+f_\gamma}$, le **rappel** vaut quant à lui $\frac{h_\gamma}{h_\gamma+m_\gamma}$.

Pour chaque annotateur, et pour chaque catégorie, calculer la précision et le rappel.

1. On peut le faire à la main, ou en utilisant une formule du genre `=SI(B2=C2;1;0)`.

6. La mesure d'accord inter-annotateur de Cohen (« Kappa ($\kappa$) de Cohen ») aborde la question en évaluant l'accord obtenu par rapport à un accord théorique qu'on obtiendrait si les juges répondaient complètement au hasard. En utilisant le tableau déjà pré-rempli `calcul_KappaFleiss.xls`, calculer, pour chaque annotateur, la valeur de $\kappa$ obtenue (ici on utilise non pas la version de Cohen, limitée à deux juges/annotateurs, mais la version de Fleiss, qui fonctionne avec un nombre quelconque d'annotateurs).

Pour obtenir un bonus, me renvoyer l'ensemble des résultats **dans un seul tableur pour le binôme**, en n'oubliant de m'indiquer vos noms.