

## TP (3-4-)5: Conception et implémentation d'un système de questions réponses en langue naturelle sur des données structurées

### Continuation du TP3-4.

**Tache 3. Évaluation.** Pour chacune des questions, il faut évaluer les réponses fournies par votre système, en les comparant aux réponses fournies par le document goldstandard *questions.xml*. Vous devez évaluer la précision, le rappel et la F-mesure pour chaque question:

$$Recall(q) = \frac{\text{number of correct system answers}}{\text{number of gold standard answers}}$$

$$Precision(q) = \frac{\text{number of correct system answers}}{\text{number of system answers}}$$

$$F\text{-measure}(q) = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Il faut ensuite calculer également la précision et le rappel globale en prenant la moyenne de toutes les valeurs de précision et de rappel de chaque question, ainsi que la mesure F globale.

Analysez les erreurs. Sur quel type de question votre système se trompe ? Quelles améliorations vous pouvez envisager ?