

Soutenance TER

Appariement Questions/Questions

Félix Jamet

30 mai 2018

Outline

- 1 Présentation du sujet
- 2 Approches explorées

Appariement de questions/questions

- Communautés de questions / réponses (AskUbuntu, StackExchange)
- Beaucoup de données (questions dupliquées)
- Intéressant de pouvoir évaluer la similarité de questions

- *Workshop*
- Évaluation de systèmes d'analyse de sémantique computationnelle
- Organisé en tâches

Tâche 3

- Similarité question / réponse
- Sous-tâche 3B : similarité question / question
- Données extraites du forum Qatar Living

Organisation des données

- N questions originales
- $N \times 10$ questions reliées
- Dans un fichier XML

- Fautes de frappe ou langage abrégé
- Parties non pertinentes à la sémantique (ex: remerciements)

Document

- Texte à analyser (en l'occurrence question)

Corpus

- Ensemble de documents

Sac de mots

- Approche d'analyse de document en tant que multi-ensemble de tokens

- *Term Frequency - Inverse Document Frequency*
- Donne une idée de l'importance d'un terme dans un document et dans un corpus

$$TF(\text{terme}, \text{document}) = \frac{\text{occurrences}(\text{terme}, \text{document})}{\text{taille}(\text{document})}$$

$$IDF(\text{terme}, \text{corpus}) = \log \left(\frac{\text{taille}(\text{corpus})}{\|\{\text{doc}/\text{doc} \in \text{corpus} \wedge \text{terme} \in \text{doc}\}\|} \right)$$

$$TF-IDF(\text{terme}, \text{document}, \text{corpus}) =$$

$$\begin{cases} TF(\text{terme}, \text{document}) \times IDF(\text{terme}, \text{corpus}) & \text{si } \text{terme} \in \text{corpus} \\ \max(\{IDF(el, \text{corpus})/el \in \text{corpus}\}) & \text{sinon} \end{cases}$$

Méthode de référence

- Somme des valeurs TF-IDF des tokens du sac de mots
- Corpus : les questions originales et reliées mises bout à bout
- Document : les deux questions concaténées

similaritéRéf érence(Q_1, Q_2) =

$$\sum_{\text{terme} \in Q_1 \cap Q_2} TF\text{-}IDF(\text{terme}, Q_1 \cup Q_2, \text{corpus})$$

Méthode de référence - Scores

Édition	Méthode	Score MAP
2016	UH-PRHLT-contrastive2	77.33
2016	UH-PRHLT-primary	76.70
2016	UH-PRHLT-contrastive1	76.56
2016	<i>IR baseline</i>	74.75
2016	Référence	71.48
2017	KeLP-contrastive1	49.00
2017	SimBow-contrastive2	47.87
2017	SimBow-primary	47.22
2017	<i>IR baseline</i>	41.85
2017	Référence	44.21

Table: SemEval - Meilleurs scores MAP