M2 UE Appli spécialisée pour le TALNE

(Séance 2 : Prédiction de Note + Explication de la prise de décision)

Avignon, le 11 décembre 2013 Enseignant : Marc El-Bèze marc.elbeze@univ-avignon.fr

Cette nouvelle séance de TP a un double objectif

- trouver les moyens de prédire la note mise sur un film par un membre du réseau social VodKaster
- expliquer la raison de son évaluation en faisant apparaître les éléments qui ont le plus contribué à la décision du système de choisir une note négative ou positive.

On ne tiendra compte pour cela que des micro-critiques (MC) ayant reçu une note négative ou une note positive (champ 5)
Une note comprise entre 0.5 et 2 (inclus) sera considérée comme négative
Une note supérieure ou égale à 4 sera considérée comme positive
Les MC ayant une note égale à 0 (refus de noter) sont donc écartées

Le sous corpus obtenu à partir des 6999 MC du fichier AppVk13 sera ainsi réduit après filtrage. On fera de même avec le corpus DeVk13 qui est mis à votre disposition pour la mise au point des paramètres. Ce corpus contient des MC plus récentes.

Dans un premier temps, on complétera l'ensemble des questions posées lors de la première séance par les 2 questions suivantes :
1/ Vérifier l'homogénéité des 2 corpus App et Dev
2/ Extraire de AppVk13 les expressions (1 à n mots) qui contribuent le plus à attribuer note négative et celles qui contribuent le plus à donner une note positive. Pour éviter que la liste ne soit trop longue on se limitera à un pouvoir* discriminant = 1 et une fréquence supérieure à 7.

Enfin c'est cela qui est au centre de cette session, pour chacune des MC du corpus DeVk13 filtré, on essaiera de deviner si la note mise était positive ou négative. De plus, on affichera pour chaque note prédite LA raison majeure de la prise de décision par le système.

Les résultats seront présentés sous forme de micro et macro mesures en calculant la_précision et le rappel et le f-score moyen

***** Le pouvoir discriminant *G(t)* d'un terme *t* est sa capacité à déterminer plus ou moins l'une des classes *c* dans lesquelles il a été employé. Il peut être estimé comme la somme sur *c* des carrés de la probabilité de *c* sachant *t*.

$$Rappel_i = \frac{(nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i)}{(nombre\ de\ documents\ appartenant\ à\ la\ classe\ i)} \qquad Rappel = \frac{\sum_{i=1}^{n} Rappel_i}{n}$$

$$Précision_i = \frac{(nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i)}{(nombre\ de\ documents\ attribués\ à\ la\ classe\ i)} \qquad Précision = \frac{\sum_{i=1}^{n} Précision_i}{n}$$

$$F_{score}(\beta) = \frac{(\beta^2+1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel}$$