

Objectifs

- Appliquer les notions de Vrai/Faux Positif, Vrai/Faux Négatif
- Calculer et comprendre Rappel, Précision et F-Mesure

1 Parseur Automatique : tester des règles

- Etablissez une liste de règles à appliquer pour faire la tokenisation du français (moins de 10 règles au total). Une fois votre liste établie, ne la modifiez plus.
- Comptez (à la main) le nombre de mots présents dans le texte ci-dessous

Les membres permanents du Conseil de sécurité ne sont pas parvenus mercredi 28/08 à s'accorder sur une résolution britannique justifiant une action armée en Syrie, Londres assurant qu'elle n'aurait pas lieu avant que les résultats de l'enquête de l'O.N.U. soient connus. La ligne de fracture entre ces cinq pays – Chine et Russie d'un côté, France, Royaume-Uni et Etats-Unis de l'autre – reflète fidèlement les positions de chacun sur le conflit qui a fait plus de 100 000 morts et poussé des millions de Syriens à la fuite depuis mars 2011. "Rendez-vous compte", insiste le Premier Ministre, "que ce conflit dure depuis des années".

<http://www.lemonde.fr/proche-orient/article/2013/08/28/syrie-londres-presentera-un-projet-de>

- Listez les règles que vous avez appliquées pour faire votre découpage manuel : gestion des " l' ", " s' ", " n' ", gestion des mots composés, etc.
- Maintenant, scindez le texte selon les règles que vous avez listées à la question a). En les suivant à la lettre ! Combien de mots obtiendriez-vous avec un tel système ?
- Si l'on doit évaluer la performance de vos règles, comment définir les Faux Positifs et les Faux Négatifs pour cette tâche ?

2 Mesures de performance

2.1 Les mesures à connaître

- $Rappel = \frac{\text{Nombre d'éléments correctement classés (Vrais Positifs)}}{\text{Nombre d'éléments à classer (Vrais Positifs + Faux Négatifs)}}$
- $Precision = \frac{\text{Nombre d'éléments correctement classés (Vrais Positifs)}}{\text{Nombre d'éléments classés (Vrais Positifs + Faux Positifs)}}$
- Pour agréger (fusionner) les deux mesures en une seule:
 - $F_1 Mesure = \frac{2 * Precision * Rappel}{Precision + Rappel}$
- Le paramètre β permet de donner plus d'importance au rappel ($\beta < 1$) ou à la précision ($\beta > 1$)
 - $F_\beta Mesure = \frac{(1+\beta) * Precision * Rappel}{\beta * Precision + Rappel}$

2.2 De la matrice de confusion à l'évaluation proprement dite

NB: vous pouvez réaliser ces exercices à la main ou en PYTHON en exploitant le NOTEBOOK en ligne.

Classe réelle :		A	B	C	D	E	Total
Classe prédite	A	1		2			3
	B	1	9	4	80		94
	C	20	1	28	5		54
	D		18		1		19
	E	1				59	60
Total		23	28	34	86	59	230

Table 1: Matrice de confusion 1

- Calculez le rappel, la précision et la F-mesure avec les résultats obtenus dans l'exercice précédent.
- Les données du tableau 1 présentent la matrice de confusion d'un classifieur pour un jeu de données comprenant des instances des classes A, B, C, D et E. Analysez la matrice : quelle est la classe la mieux reconnue ? La moins bien reconnue ?
- Remplissez le tableau 2

2.3 Performance moyenne et moyenne des performances

- Calculez pour la matrice du tableau 3 le pourcentage d'instances correctement classées, et la moyenne des rappels.
- Pour la matrice ci-dessus, comment expliquez-vous qu'en moyenne, les rappels soient bons, alors que le pourcentage d'instances correctement classées est très mauvais ?

Rappel				
A	B	C	D	E
Précision				
A	B	C	D	E
F-mesure				
A	B	C	D	E

Table 2: Tableau récapitulatif des scores issus des données du tableau 1

Classe réelle :		A	B	C	D	E	Total
Classe prédite	A	5			584	1	589
	B		4		84		88
	C	1		8	12	2	23
	D				42		42
	E				168	8	176
	Total	6	4	8	890	11	919

Table 3: Matrice de confusion 2