M1 Données et Connaissances

Recherche d'Information

Rapport de TP 6 et 7

2 novembre 2019

Table des matières

Informations

Voici le rapport qui présente notre le travail que nous avons réalisé pour les TP 6 et 7. Ce travail a été réalisé par **BARDY Benjamin** et **ORTIZ Diégo**. Vous pourrez retrouver les différents codes sources sur notre notre repository github.

TP6 : Catégorisation de documents texte

6. — Résultats avec J48 :

- Correctly Classified Instances : 116 (96.667%)
- Incorrectly Classified Instances : 4 (3.3333%)
- Kappa statistic : 0.96
- Mean absolute error : 0.0133
- Root mean squared error : 0.1033
- Relative absolute error : 4.7895 %
- Root relative squared error : 27.7204 %
- Total Number of Instances : 120

— Résultats avec Naive Bayes (multinomial) :

- Correctly Classified Instances : 101 (84.1667%)
- Incorrectly Classified Instances : 19 (15.8333%)
- Kappa statistic : 0.81
- Mean absolute error : 0.0514
- Root mean squared error : 0.2219
- Relative absolute error : 18.5074 %
- Root relative squared error : 59.5522 %
- Total Number of Instances : 120

— Résultats avec une SVM :

- Correctly Classified Instances : 78 (65%)
- Incorrectly Classified Instances : 42 (35%)

- Kappa statistic : 0.58
- Mean absolute error : 0.2365
- Root mean squared error : 0.3324
- Relative absolute error : 85.1333 %
- Root relative squared error : 89.1939 %
- Total Number of Instances : 120

7. — Résultats avec J48 + reduced error pruning (-R) :

- Correctly Classified Instances : 119 (99.1667%)
- Incorrectly Classified Instances : 1 (0.8333%)
- Kappa statistic : 0.99
- Mean absolute error : 0.0956
- Root mean squared error : 0.1331
- Relative absolute error : 34.4185 %
- Root relative squared error : 35.7173 %
- Total Number of Instances : 120

Pour Naive Bayes et SVM, aucun paramètre ne change les résultats.

8. — Pour J48 :

- Correctly Classified Instances : 23 (76.6667%)
- Incorrectly Classified Instances : 7 (23.3333%)
- Kappa statistic : 0.72
- Mean absolute error : 0.0802
- Root mean squared error : 0.2792
- Relative absolute error : 28.8571 %

- Root relative squared error : 74.9149 %
- Total Number of Instances : 30

— Pour Naïve Bayes (multinomial) :

- Correctly Classified Instances : 26 (86.6667%)
- Incorrectly Classified Instances : 4 (13.3333%)
- Kappa statistic : 0.84
- Mean absolute error : 0.0471
- Root mean squared error : 0.2087
- Relative absolute error : 16.9496 %
- Root relative squared error : 55.9892 %
- Total Number of Instances : 30

— Pour SVM :

- Correctly Classified Instances : 21 (70%)
- Incorrectly Classified Instances : 9 (30%)
- Kappa statistic : 0.64
- Mean absolute error : 0.2348
- Root mean squared error : 0.3299
- Relative absolute error : 84.5333 %
- Root relative squared error : 88.5237 %
- Total Number of Instances : 30

Même si les scores du Naïve Bayes et de la SVM sont plus élevés avec la simple évaluation du jeu de test, les valeurs enregistrées après la validation croisée sont beaucoup plus représentatives des performances réelles de l'algorithme. Ici, la simple chance a dû nous aider à avoir de meilleurs résultats.

TP7 : Résumé automatique

3. Statisques :

	Battery-life	**Amazon Kindle**	**Room holiday London**	**Sound Ipod**	**Speed Windows7**
Nombre de lignes	333	100	575	101	124
Nombre de mots	6573	2051	12658	1595	2234
Nombre de mots unique	1420	646	2046	497	675
nombre de mots moyen / ligne	19.7	20.5	22	15.8	18

On a aussi calculé les occurrences de chaque mot pour chaque documents :

— Pour le fichier **battery life netbook** :

battery (331), life (168), hours (41), netbook (35), charge (32), long (32), 10 (28), great (27), 5 (25), Battery (23), get (23), power (21), time (20), computer (20), use (18), can (17), screen (16), one (15), machine (15), charger (15)

— Pour le fichier **price amazon kindle** :

price (74), Kindle (26), books (21), book (18), prices (13), will (10), priced (9), find (9), purchase (9), many (9), high (7), cost (7), much (7), device (7), one (7), 2 (7), cover (6), reading (6), still (5), worth (5)

— Pour le fichier **room holiday london** :

room (518), clean (118), small (98), rooms (89), hotel (81), bathroom (68), bed (58), size (53), good (50), London (42), comfortable (38), floor (37), time (33), great (32), one (31), Room (28), nice (25), service (25), breakfast (25), staff (24)

— Pour le fichier **sound ipod nano** :

sound (81), quality (43), great (28), use (21), Sound (15), good (15), easy (14), Great (12), video (10), headphones (9), love (8), music (7), Easy (7), can (6), really (6),

speaker (6), iPod (6), headphone (5), excellent (5), picture (5)

— Pour le fichier **speed windows7** :

faster (76), fast (34), Vista (24), Windows (21), 7 (19), much (17), XP (16), system (10), slower (9), install (9), Win7 (9), slow (9), computer (8)

4. Calcul des résumés automatiques en utilisant les top-5 documents.

— Pour le fichier **battery life netbook** :

1. The main selling point for this netbook is its amazing battery life.
2. 0 out of 5 stars netbook with power and battery life, August 24, 2009.
3. This netbook has plenty of power for my needs and the battery life is great.
4. Battery life is one of the biggest features of a netbook so I wouldn't go cheap and get less battery.
5. Pretty standard list of specs for a netbook with the addition of bluetooth and a very long battery life.

— Pour le fichier **price amazon kindle** :

1. Amazon is not charging full price to replace a broken or damaged Kindle .
2. 99, thus offsetting the high price of the Kindle .
3. 00, so the price of Kindle effectively dropped to $349 .
4. The price for new hardbacks on the Kindle is great too .
5. 99 price and credits those who've paid more, I'll once again recommend Kindle .

— Pour le fichier **room holiday london** :

1. Room was an Executive Club floor room and was larger than average for a London Holiday Inn, well laid out and everything worked !
2. Because of the room size, the lack of continental breakfast for those who reserve outside of the Holiday Inn web site, and the distance to many attractions, we WILL NOT STAY in the hotel during our next trip to London .
3. We got in about 9am to the Holiday Inn, and the room wasn't ready yet.
4. Our room was typical holiday inn the bathroom could have done with updating

but was spotless .

5. My room was a typical London size i .

— Pour le fichier **sound ipod nano** :

1. Great Sound, Sleek Compact The only problem so far is that our Ipod speakers don't charge the new Ipod nano .

2. Received my new iPod nano and was pleased with the radio, camera, pedometer, sound quality and appearance .

3. The iPod Nano is a very sleek, cutting edge device with high, quality sound and state, of, the, art graphics .

4. New User to iPOD , , , Sound is Great

5. The nano is definitely several notches above the sound

— Pour le fichier **speed windows7** :

1. Windows 7 is overall faster than Windows XP .

2. But Windows 7 is definitely faster !

3. I admit Windows 7 performs faster than Vista as a whole .

4. Simply put, Windows 7 looks amazing, and is fast and intuitive .

5. Windows 7 is also faster in the core of the operating system .

8. Comparaisons entre les différents systèmes

ROUGE-Type	Task Name	Avg_Recall	Avg_Recall	Avg_Precisio	Avg_Precisio	Avg_F-Score	Avg_F-Score
ROUGE-L+	AMAZON	**0.66667**	0.50000	0.09195	**0.11538**	0.16162	**0.18750**
ROUGE-1	AMAZON	**0.69231**	0.46154	0.08571	**0.09836**	0.15254	**0.16216**
ROUGE-2	AMAZON	0.00000	**0.09091**	0.00000	**0.01786**	0.00000	**0.02985**
ROUGE-SU4	AMAZON	**0.24444**	0.22222	0.02316	**0.03922**	0.04231	**0.06667**
ROUGE-L+	IPOD	0.28571	0.28571	0.02500	**0.05128**	0.04598	**0.08696**
ROUGE-1	IPOD	**0.42857**	0.28571	0.03125	**0.04167**	0.05825	**0.07273**
ROUGE-2	IPOD	0.00000	**0.20000**	0.00000	**0.02326**	0.00000	**0.04167**
ROUGE-SU4	IPOD	0.18750	0.18750	0.00698	**0.01579**	0.01345	**0.02913**
ROUGE-L+	NETBOOK	**0.62500**	0.25000	0.04386	**0.09091**	0.08197	**0.13333**
ROUGE-1	NETBOOK	**0.66667**	0.33333	0.04478	**0.05660**	0.08392	**0.09677**
ROUGE-2	NETBOOK	**0.14286**	0.00000	**0.00775**	0.00000	**0.01471**	0.00000
ROUGE-SU4	NETBOOK	**0.36000**	0.16000	0.01452	**0.01860**	0.02791	**0.03333**
ROUGE-L+	HOLIDAY	0.25000	**0.37500**	0.00901	**0.06383**	0.01739	**0.10909**
ROUGE-1	HOLIDAY	0.25000	**0.37500**	0.00673	**0.04918**	0.01311	**0.08696**
ROUGE-2	HOLIDAY	0.00000	0.00000	0.00000	**0.00000**	0.00000	0.00000
ROUGE-SU4	HOLIDAY	**0.08696**	0.13043	0.00139	**0.01176**	0.00274	**0.02158**
ROUGE-L+	WINDOWS7	0.20000	0.20000	0.02913	**0.08333**	0.05085	**0.11765**
ROUGE-1	WINDOWS7	0.18750	0.18750	0.02013	**0.06122**	0.03636	**0.09231**
ROUGE-2	WINDOWS7	0.07692	0.07692	0.00694	**0.02273**	0.01274	**0.03509**
ROUGE-SU4	WINDOWS7	0.08000	0.08000	0.00576	**0.02041**	0.01074	**0.03252**
		9	4	1	19	1	18

On remarque que avec notre système par rapport au système naïf on obtient des résultats supérieurs pour le "recall moyen" mais on obtient des résultats inférieur pour la précision moyenne et pour le "F-score".

9. Comparaisons entre les différents systèmes

ROUGE-Type	Task Name	Avg_Recall	Avg_Recall	Avg_Precision	Avg_Precision	Avg_F-Score	Avg_F-Score
ROUGE-L+	AMAZON	**0.66667**	0.50000	0.06667	**0.11538**	0.12121	**0.18750**
ROUGE-1	AMAZON	**0.69231**	0.46154	0.05696	**0.09836**	0.10526	**0.16216**
ROUGE-2	AMAZON	0.09091	0.09091	0.00676	**0.01786**	0.01258	**0.02985**
ROUGE-SU4	AMAZON	**0.28889**	0.22222	0.01884	**0.03922**	0.03537	**0.06667**
ROUGE-L+	IPOD	**0.85714**	0.28571	0.05128	**0.05128**	0.09677	**0.08696**
ROUGE-1	IPOD	**0.85714**	0.28571	0.03750	**0.04167**	0.07186	0.07273
ROUGE-2	IPOD	**0.40000**	0.20000	0.01333	**0.02326**	0.02581	**0.04167**
ROUGE-SU4	IPOD	**0.56250**	0.18750	0.01286	**0.01579**	0.02514	**0.02913**
ROUGE-L+	NETBOOK	**0.62500**	0.25000	0.02646	**0.09091**	0.05076	**0.13333**
ROUGE-1	NETBOOK	**0.77778**	0.33333	0.02846	**0.05660**	0.05490	**0.09677**
ROUGE-2	NETBOOK	**0.28571**	0.00000	**0.00847**	0.00000	**0.01646**	0.00000
ROUGE-SU4	NETBOOK	**0.44000**	0.16000	0.00973	**0.01860**	0.01905	**0.03333**
ROUGE-L+	HOLIDAY	**0.50000**	0.37500	**0.01311**	0.06383	0.02556	**0.10909**
ROUGE-1	HOLIDAY	**0.50000**	0.37500	0.00937	**0.04918**	0.01839	**0.08696**
ROUGE-2	HOLIDAY	0.00000	0.00000	0.00000	**0.00000**	0.00000	0.00000
ROUGE-SU4	HOLIDAY	**0.17391**	0.13043	0.00197	**0.01176**	0.00389	**0.02158**
ROUGE-L+	WINDOWS7	0.20000	0.20000	0.02041	**0.08333**	0.03704	**0.11765**
ROUGE-1	WINDOWS7	0.18750	0.18750	0.01357	**0.06122**	0.02532	**0.09231**
ROUGE-2	WINDOWS7	0.07692	0.07692	0.00474	**0.02273**	0.00893	**0.03509**
ROUGE-SU4	WINDOWS7	0.08000	0.08000	0.00398	**0.02041**	0.00758	**0.03252**
		9	4	1	19	1	17

En modifiant la longueur du document souhaité et en retirant les stop-words on obtient de meilleurs résultats avec notre nouveau système comparé à l'ancien. En effet on améliore à la fois le recall moyen qui surpasse celui du système naïf. On améliore aussi la précision moyenne mais on arrive toujours pas à égaler celle du système naïf.

10 **Question bonus** : Nous avons appliqué le code sur moodle à notre corpus, on obtient les résultats suivants (nous exposeront les résultats seulement pour les deux premiers fichiers ; aussi nous défini le nombre max de topics à 5) :

— Pour le fichier **battery life netbook** :

Topic 0 : charge netbook adapter computer 10 time problem charging went using longer power ac hrs second plugged fine use charger started

Topic 1 : life hours long netbook great 10 hour power use screen time light laptop ve really machine running performance don nice

Topic 2 : asus eee performance pc life just problems new stars read netbook charger reviews bought 10 quickly like weeks hour doesn

Topic 3 : life cell better keyboard bluetooth best charging screen ion amazing fast great upgrade webcam computer touch currently available design li

— Pour le fichier **price amazon kindle** :

Topic 0 : price just reader cover paid lack book think worth nice k2 especially really reading purchase reviews make low device like

Topic 1 : books priced price purchase pay buy cost value hard pricey new dollars reading pretty book prices know love especially download

Topic 2 : price kindle book books amazon high version device great list tag purchase prices priced cost believe issues paying higher better

Topic 3 : prices books thought buying releases amazon new book given ability priceless addition reading user reader nice free good really high

— Pour le fichier **room holiday london** :

Topic 0 : room staff fridge plenty impressed restaurant tesco waitrose gloucester rd hours high tube really run couple happy touch wine rooms

Topic 1 : rooms tea sure basic hand noise coffee facilities connecting hair standard kettle cream away soap looks shampoo hotel making like

Topic 2 : room clean small hotel rooms bathroom bed size london good comfortable floor great breakfast time nice service check large night

Topic 3 : appeared fault finally calls room public cleared conditioning despite following areas air evening day reception minor heat working issues knew

— Pour le fichier **sound ipod nano** :

Topic 0 : time improved previous hear really sounded video compact screen sound able ipod new appearance great particularly look speakers display little

Topic 1 : great sound easy use love battery quality life headphones features pedometer right long listen nice really lots christmas need doesn

Topic 2 : sound quality great easy good use video ipod nano music headphone love picture headphones jack excellent radio camera storage awesome

Topic 3 : speaker bose added sound ipod car better recording just good great video quality feature pedometer ears compact speakers earphones able

— Pour le fichier **speed window7** :

Topic 0 : faster start boot install xp hibernation window time opening programs slower file sleep say does pretty 20 10 operating new

Topic 1 : fast windows slower vista xp faster computer run finally pretty user use photoshop bit worked great just win7 sp3 looks

Topic 2 : faster vista xp slow fast things install win7 like hardware runs time windows way upgrade bit lot work does clean

Topic 3 : faster fast vista windows stable os open performance run boots running 2009 significantly stars computer programs use december overall hp

Sources

1. Code source sur github des TP6 et 7 : https://github.com/Diegoortizz/RI_TP67