Université Paul Sabatier
EIMAB3H1 - Analyse et exploitation de données

Enseignant : José G. Moreno

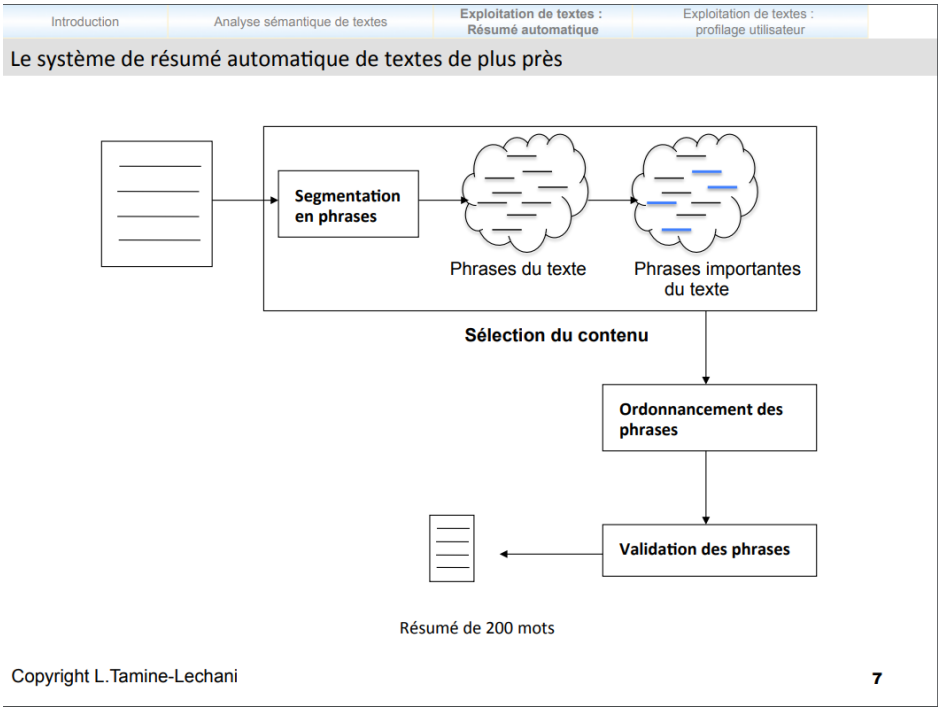
03/10/2019

TP 3. Résumé automatique

Les résumés automatiques sont devenus un moyen commun pour explorer de grandes collections de documents.

Les défis de la génération automatique de résumés résident dans l'extraction de points importants du document. Par exemple, un bon système de résumé automatique du printemps arabe devrai capturer des événements comme "le président de l'Egypte, Hosni Moubarak, a démissionné le 11 février 2011. ", "Mouammar Kadhafi a été tué le 20 octobre 2011.", etc.

Dans ce TP nous utiliserons une collection de critiques d'utilisateurs sur 5 produits afin de construire un résumé automatique pour chaque produit de type extractif comme vu dans le cours.



1. Téléchargez le fichier Critiques.zip et décompressez-le. Ce fichier contient sur le dossier « projects/test-summarization » trois dossiers : « reference », « system » et « topics ». Dans « topics » vous trouverez 5 fichiers et chaque fichier contient plusieurs phrases (1 par ligne), ces seront notres documents à résumer (ou les documents d'entrée). Dans « reference » vous trouverez 5 fichiers avec le résumé de chaque topique fait manuellement par un humain. Finalement, dans « system » vous devez créer 5 fichiers avec le résumé de votre système. Un résumé naïf est déjà dans le dossier « system », il consiste à utiliser les 5 premières lignes de chaque document d'entré (dossier « topics »). Notez que

vous pouvez facilement améliorer ces résultats.

2. Utilisez le fichier jar pour évaluer le système naïf, pour le faire utilisez la commande (sur une console) :

```
java -jar rouge2-1.2.1.jar
```

Vous aurez comme sortie :

```
========Results Summary=======

ROUGE-L+StopWordRemoval PRICE AMAZON      Average_R:0.50000
      Average_P:0.11538 Average_F:0.18750 Num Reference Summaries:1


ROUGE-1+StopWordRemoval PRICE AMAZON      Average_R:0.46154
      Average_P:0.09836 Average_F:0.16216 Num Reference Summaries:1


ROUGE-2+StopWordRemoval PRICE AMAZON      Average_R:0.09091
      Average_P:0.01786 Average_F:0.02985 Num Reference Summaries:1


ROUGE-SU4+StopWordRemoval     PRICE AMAZON      Average_R:0.22222
      Average_P:0.03922 Average_F:0.06667 Num Reference Summaries:1


ROUGE-L+StopWordRemoval SOUND IPOD  Average_R:0.28571 Average_P:0.05128
      Average_F:0.08696 Num Reference Summaries:1


ROUGE-1+StopWordRemoval SOUND IPOD  Average_R:0.28571 Average_P:0.04167
      Average_F:0.07273 Num Reference Summaries:1


ROUGE-2+StopWordRemoval SOUND IPOD  Average_R:0.20000 Average_P:0.02326
      Average_F:0.04167 Num Reference Summaries:1


ROUGE-SU4+StopWordRemoval     SOUND IPOD  Average_R:0.18750
      Average_P:0.01579 Average_F:0.02913 Num Reference Summaries:1


ROUGE-L+StopWordRemoval BATTERY-LIFE      NETBOOK     Average_R:0.25000
      Average_P:0.09091 Average_F:0.13333 Num Reference Summaries:1


ROUGE-1+StopWordRemoval BATTERY-LIFE      NETBOOK     Average_R:0.33333
      Average_P:0.05660 Average_F:0.09677 Num Reference Summaries:1


ROUGE-2+StopWordRemoval BATTERY-LIFE      NETBOOK     Average_R:0.00000
      Average_P:0.00000 Average_F:0.00000 Num Reference Summaries:1


ROUGE-SU4+StopWordRemoval     BATTERY-LIFE      NETBOOK
```

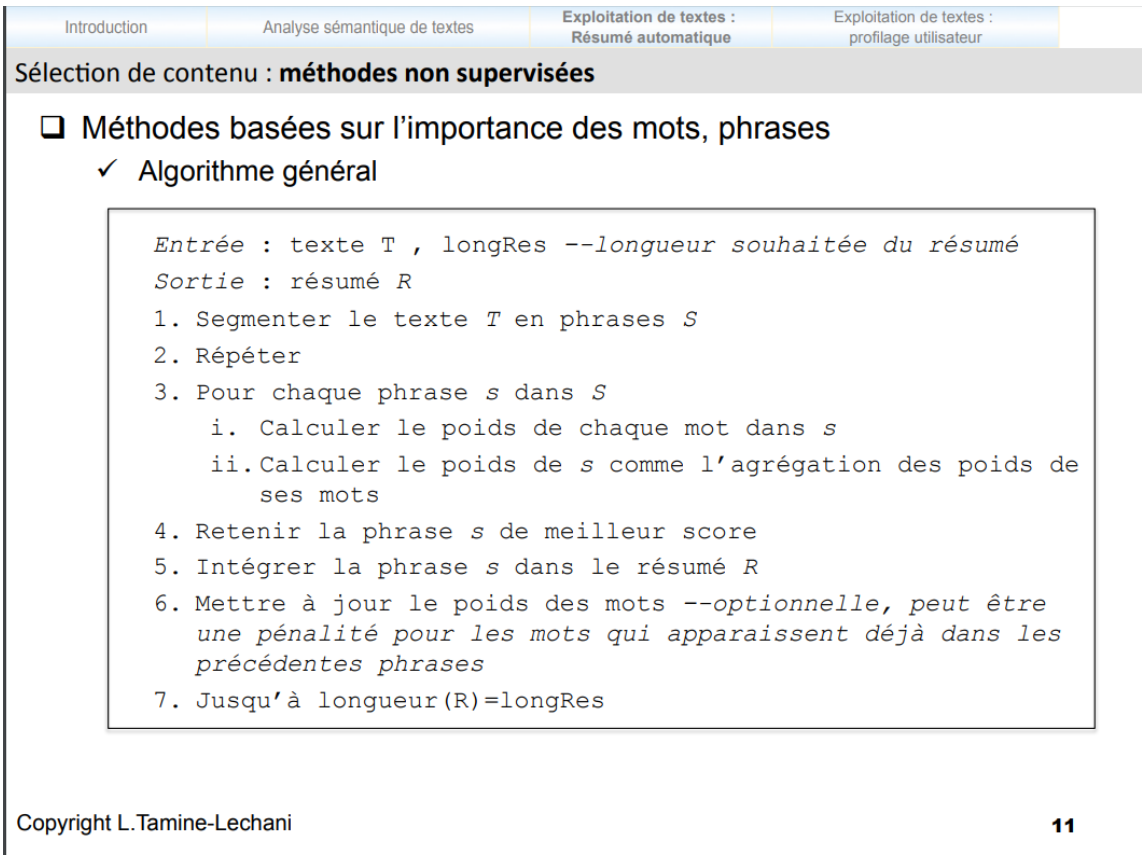
```
      Average_R:0.16000 Average_P:0.01860 Average_F:0.03333 Num
Reference Summaries:1


ROUGE-L+StopWordRemoval ROOM  HOLIDAY     Average_R:0.37500
      Average_P:0.06383 Average_F:0.10909 Num Reference Summaries:1


ROUGE-1+StopWordRemoval ROOM  HOLIDAY     Average_R:0.37500
      Average_P:0.04918 Average_F:0.08696 Num Reference Summaries:1


ROUGE-2+StopWordRemoval ROOM  HOLIDAY     Average_R:0.00000
      Average_P:0.00000 Average_F:0.00000 Num Reference Summaries:1


ROUGE-SU4+StopWordRemoval     ROOM  HOLIDAY     Average_R:0.13043
      Average_P:0.01176 Average_F:0.02158 Num Reference Summaries:1


ROUGE-L+StopWordRemoval SPEED WINDOWS7.TXT.SYS.HEAD5  Average_R:0.20000
      Average_P:0.08333 Average_F:0.11765 Num Reference Summaries:1


ROUGE-1+StopWordRemoval SPEED WINDOWS7.TXT.SYS.HEAD5  Average_R:0.18750
      Average_P:0.06122 Average_F:0.09231 Num Reference Summaries:1


ROUGE-2+StopWordRemoval SPEED WINDOWS7.TXT.SYS.HEAD5  Average_R:0.07692
      Average_P:0.02273 Average_F:0.03509 Num Reference Summaries:1


ROUGE-SU4+StopWordRemoval     SPEED WINDOWS7.TXT.SYS.HEAD5
      Average_R:0.08000 Average_P:0.02041 Average_F:0.03252 Num
Reference Summaries:1


======Results Summary End======
```

Notez que les valeurs de ROUGE-L, ROUGE-1, ROUGE-2 et ROUGE-SU4 sont affichées en termes de Rappel, Précision et F-mesure. Les résultats sont calculés pour les 5 topiques que nous avons : « battery life netbook », « price amazon kindle », « room holiday inn london », « sound ipod nano 8g » et « speed windows 7 ».

3. Inspectez les documents sur le dossier « topics ». Calculez de statistiques basiques de chaque fichier, comme par exemple le nombre de lignes, de mots, mots plus fréquents, etc.

4. Pour chaque document dans le dossier topiques et à l'aide de python, stockez toutes le lignes dans un tableau de tel façon que chaque ligne corresponde à une dimension du tableau.

5. Calculez un résumé automatique en utilisant l'algorithme du cours (diapo 11). Alternativement, vous pouvez calculer la similarité de chaque ligne avec toutes les autres lignes du même tableau. Pour ça, vous pouvez utiliser wordnet, ppmi, la formule donné en cours basés sur le tf-idf ou les word embeddings.



6. Sauvegardez votre résumé de chaque document d'entrée dans le dossier « system » en écrasant les fichiers existants.

7. Calculez à nouveau les valeurs de ROUGE en utilisant le fichier jar comme dans le point 2. Avez-vous une amélioration par rapport au système naïf ? Si non, vérifiez les étapes 4 et 5 jusqu'à trouver un une amélioration dans les résultats.

8. Utilisez une autre méthode (ou d'autres paramètres de votre système) pour améliorer encore vos résultats.