Assignment Description

PartA：question

1. **a:** the function of reading json file is the **data\_reprocessing.py,** and it is imported into the PartA.py main running file. You can run it independently.

**b:** the reading json file function contains extracting caption in assignment. json

**c:** please see the python file **PartA.py** and **data\_reprocessing.py.**

**d:** The result of term frequency and the inverse document frequency please see in the **tf\_result.xlsx** and **idf\_result.xlsx,** and empty cells mean the value is zero. If you want to get the result of tf\_idf, please run the tf\_idf.py python file and output file is tfidf\_result.xlsx.

**Attention: empty cells mean the value is zero, the results are filtered by stop words.**

**e: please try three k cluster number parameter and run the code**

**because of the random pick centroids, the Silhouette Coefficient result and wordcloud may be different.**

* + 1. cluster = 10：

the average Silhouette Coefficient of sample: 0.01708

* + 1. cluster = 20：

the average Silhouette Coefficient of sample: 0.01755

* + 1. cluster = 40

the average Silhouette Coefficient of sample:0.02062

* + 1. wordcloud\_image: **the images are place in ‘images\_cluster40’ folder.**

**Attention:**

**in each cluster, we only \use the top 30 features.**

**The python is captions\_cluster.py and it is import into the PartA.py. You can run it independently.**



PartB：question

1. **please see the PartB.py to know the specific steps and codes.**
   1. please see the PartB.py
   2. Please the python file
   3. Result:

----------------------------------------NavieBays\_Classification----------------------------------------

NB\_Accuracy: 0.774

NB\_confusion\_matrix:

[[166 0 0 1 0 1 0 0 1 1 1 3 0 6 3 123 4 8

0 1]

[ 1 252 15 12 9 18 1 2 1 5 2 41 4 0 6 15 4 1

0 0]

[ 0 14 258 45 3 9 0 2 1 3 2 25 1 0 6 23 2 0

0 0]

[ 0 5 11 305 17 1 3 6 1 0 2 19 13 0 5 3 1 0

0 0]

[ 0 3 8 23 298 0 3 8 1 3 1 16 8 0 2 8 3 0

0 0]

[ 1 21 17 13 2 298 1 0 1 1 0 23 0 1 4 10 2 0

0 0]

[ 0 1 3 31 12 1 271 19 4 4 6 5 12 6 3 9 3 0

0 0]

[ 0 1 0 3 0 0 4 364 3 2 2 4 1 1 3 3 4 0

1 0]

[ 0 0 0 1 0 0 2 10 371 0 0 4 0 0 0 8 2 0

0 0]

[ 0 0 0 0 1 0 0 4 0 357 22 0 0 0 2 9 1 1

0 0]

[ 0 0 0 0 0 0 0 1 0 4 387 1 0 0 1 5 0 0

0 0]

[ 0 2 1 0 0 1 1 3 0 0 0 383 1 0 0 3 1 0

0 0]

[ 0 4 2 17 5 0 2 8 7 1 2 78 235 3 11 15 2 1

0 0]

[ 2 3 0 1 1 3 1 0 2 3 4 11 5 292 6 52 6 4

0 0]

[ 0 2 0 1 0 3 0 2 1 0 1 6 1 2 351 19 4 0

1 0]

[ 2 0 0 0 0 0 0 0 1 0 0 0 0 1 2 392 0 0

0 0]

[ 0 0 0 1 0 0 2 0 1 1 0 10 0 0 1 6 341 1

0 0]

[ 0 1 0 0 0 0 0 0 0 1 0 2 0 0 0 24 3 344

1 0]

[ 2 0 0 0 0 0 0 1 0 0 1 11 0 1 7 35 118 5

129 0]

[ 33 2 0 0 0 0 0 0 0 1 1 3 0 4 4 131 29 5

3 35]]

NB\_classification\_report:

precision recall f1-score support

alt.atheism 0.80 0.52 0.63 319

comp.graphics 0.81 0.65 0.72 389

comp.os.ms-windows.misc 0.82 0.65 0.73 394

comp.sys.ibm.pc.hardware 0.67 0.78 0.72 392

comp.sys.mac.hardware 0.86 0.77 0.81 385

comp.windows.x 0.89 0.75 0.82 395

misc.forsale 0.93 0.69 0.80 390

rec.autos 0.85 0.92 0.88 396

rec.motorcycles 0.94 0.93 0.93 398

rec.sport.baseball 0.92 0.90 0.91 397

rec.sport.hockey 0.89 0.97 0.93 399

sci.crypt 0.59 0.97 0.74 396

sci.electronics 0.84 0.60 0.70 393

sci.med 0.92 0.74 0.82 396

sci.space 0.84 0.89 0.87 394

soc.religion.christian 0.44 0.98 0.61 398

talk.politics.guns 0.64 0.94 0.76 364

talk.politics.mideast 0.93 0.91 0.92 376

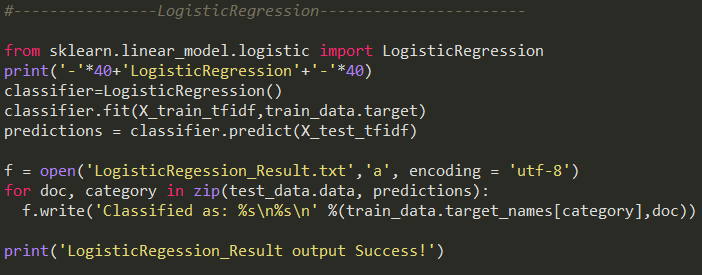
talk.politics.misc 0.96 0.42 0.58 310

talk.religion.misc 0.97 0.14 0.24 251

avg / total 0.82 0.77 0.77 7532

* 1. **Please see LogisticRegession\_Result output in the LogisticRegession\_Result.txt file.**

**Corresponding codes:**



Also provide the classification evaluation.

1. Present and compared the classification results obtained by two different classification methods.

----------------------------------------NavieBays\_Classification----------------------------------------

NB\_Accuracy: 0.774

NB\_classification\_report:

precision recall f1-score support

alt.atheism 0.80 0.52 0.63 319

comp.graphics 0.81 0.65 0.72 389

comp.os.ms-windows.misc 0.82 0.65 0.73 394

comp.sys.ibm.pc.hardware 0.67 0.78 0.72 392

comp.sys.mac.hardware 0.86 0.77 0.81 385

comp.windows.x 0.89 0.75 0.82 395

misc.forsale 0.93 0.69 0.80 390

rec.autos 0.85 0.92 0.88 396

rec.motorcycles 0.94 0.93 0.93 398

rec.sport.baseball 0.92 0.90 0.91 397

rec.sport.hockey 0.89 0.97 0.93 399

sci.crypt 0.59 0.97 0.74 396

sci.electronics 0.84 0.60 0.70 393

sci.med 0.92 0.74 0.82 396

sci.space 0.84 0.89 0.87 394

soc.religion.christian 0.44 0.98 0.61 398

talk.politics.guns 0.64 0.94 0.76 364

talk.politics.mideast 0.93 0.91 0.92 376

talk.politics.misc 0.96 0.42 0.58 310

talk.religion.misc 0.97 0.14 0.24 251

avg / total 0.82 0.77 0.77 7532

----------------------------------------LogisticRegression----------------------------------------

LR\_Accuracy: 0.828

LR\_classification\_report:

precision recall f1-score support

alt.atheism 0.80 0.74 0.77 319

comp.graphics 0.69 0.78 0.74 389

comp.os.ms-windows.misc 0.76 0.75 0.75 394

comp.sys.ibm.pc.hardware 0.73 0.72 0.72 392

comp.sys.mac.hardware 0.81 0.83 0.82 385

comp.windows.x 0.83 0.74 0.78 395

misc.forsale 0.76 0.90 0.83 390

rec.autos 0.91 0.89 0.90 396

rec.motorcycles 0.94 0.95 0.94 398

rec.sport.baseball 0.87 0.93 0.90 397

rec.sport.hockey 0.94 0.96 0.95 399

sci.crypt 0.93 0.89 0.91 396

sci.electronics 0.76 0.78 0.77 393

sci.med 0.89 0.84 0.86 396

sci.space 0.89 0.92 0.91 394

soc.religion.christian 0.79 0.93 0.85 398

talk.politics.guns 0.71 0.90 0.80 364

talk.politics.mideast 0.96 0.89 0.92 376

talk.politics.misc 0.79 0.58 0.67 310

talk.religion.misc 0.83 0.45 0.59 251

avg / total 0.83 0.83 0.83 7532

The accuracy of Naïve Bays is 0.774 and that of logistic Regression is 0.828. The performance of Naive Bays is worse than the Logistic Regression in precision, recall and f1-score. But the precision (0.82) get closed to that of Logistic Regression (0.83). In a word, the logistic regression perform better.

1. The KK Classifier is the worst among three classifications. The accuracy only has the 65.9% and other indices such as precision, recall and f1-score all perform worst.

----------------------------------------KNN\_Classifier----------------------------------------

KNN\_Accuracy: 0.659

KNN\_classification\_report:

precision recall f1-score support

alt.atheism 0.43 0.76 0.55 319

comp.graphics 0.50 0.61 0.55 389

comp.os.ms-windows.misc 0.56 0.57 0.57 394

comp.sys.ibm.pc.hardware 0.53 0.58 0.56 392

comp.sys.mac.hardware 0.59 0.56 0.57 385

comp.windows.x 0.69 0.60 0.64 395

misc.forsale 0.58 0.45 0.51 390

rec.autos 0.75 0.69 0.72 396

rec.motorcycles 0.84 0.81 0.82 398

rec.sport.baseball 0.77 0.72 0.74 397

rec.sport.hockey 0.85 0.84 0.84 399

sci.crypt 0.76 0.84 0.80 396

sci.electronics 0.70 0.50 0.58 393

sci.med 0.82 0.49 0.62 396

sci.space 0.79 0.76 0.78 394

soc.religion.christian 0.75 0.76 0.76 398

talk.politics.guns 0.70 0.73 0.72 364

talk.politics.mideast 0.62 0.76 0.69 376

talk.politics.misc 0.55 0.61 0.58 310

talk.religion.misc 0.56 0.49 0.52 251

avg / total 0.67 0.66 0.66 7532