2018-6-15

Lab Report 2

Saltzman,Thomas Renner; Fei Wu; Widner,John Drury

1. Introduction

In the second two week lab assignment 2, 4 program questions were designed to require us to get familiar with python skills and basic machine learning methods

1. Objectives

To learn more from the practice of classification of Regression and Clustering, to familiarize Scikit-Learn libraries for classifiers and use Scikit-Learn libraries to evaluate dataset

1. Approaches/Methods

**In question 1**, this program uses Linear Discriminant Analysis and Logistic Regression on the Iris dataset from Sklearn and graphs the results using matplotlib. LDA more often got better results than LR. Note this run LR did worse than normal. Applied packages: import numpy as np; import matplotlib.pyplot as plt; from sklearn.model\_selection import train\_test\_split; from sklearn.discriminant\_analysis; from sklearn.linear\_model import LogisticRegression; from sklearn.datasets import load\_iris.

**In question 2**, this is a comparison of the SVC using an RDF Kernel and SVC using a Linear Kernel. Applied packages: import numpy as np; import matplotlib.pyplot as plt; from mpl\_toolkits.mplot3d import Axes3D; from sklearn import datasets, svm; from sklearn.cluster import KMeans; from sklearn.decomposition import PCA;

from sklearn.model\_selection import train\_test\_split, cross\_val\_score;

from sklearn.metrics import accuracy\_score, average\_precision\_score; f1\_score; from sklearn.naive\_bayes import GaussianNB as Guass;

from sklearn.neighbors import KNeighborsClassifier; import pandas

**In question 3**, we used packages from NLP (Natural language processing) NLTK (Natural Language Toolkit) to aid us in processing text. Applied packages: import nltk;

from nltk.stem import WordNetLemmatizer;

from nltk.tokenize import word\_tokenize,sent\_tokenize;

from nltk.util import ngrams;

import re, collections.

**In question 3,** With Knn whether we want a larger K or a smaller K depends on the dataset, the type of data and the sample size of the data.

import numpy as np

import matplotlib.pyplot as plt

from sklearn.datasets import load\_digits

from sklearn.neighbors import KNeighborsClassifier

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score

1. Workflow

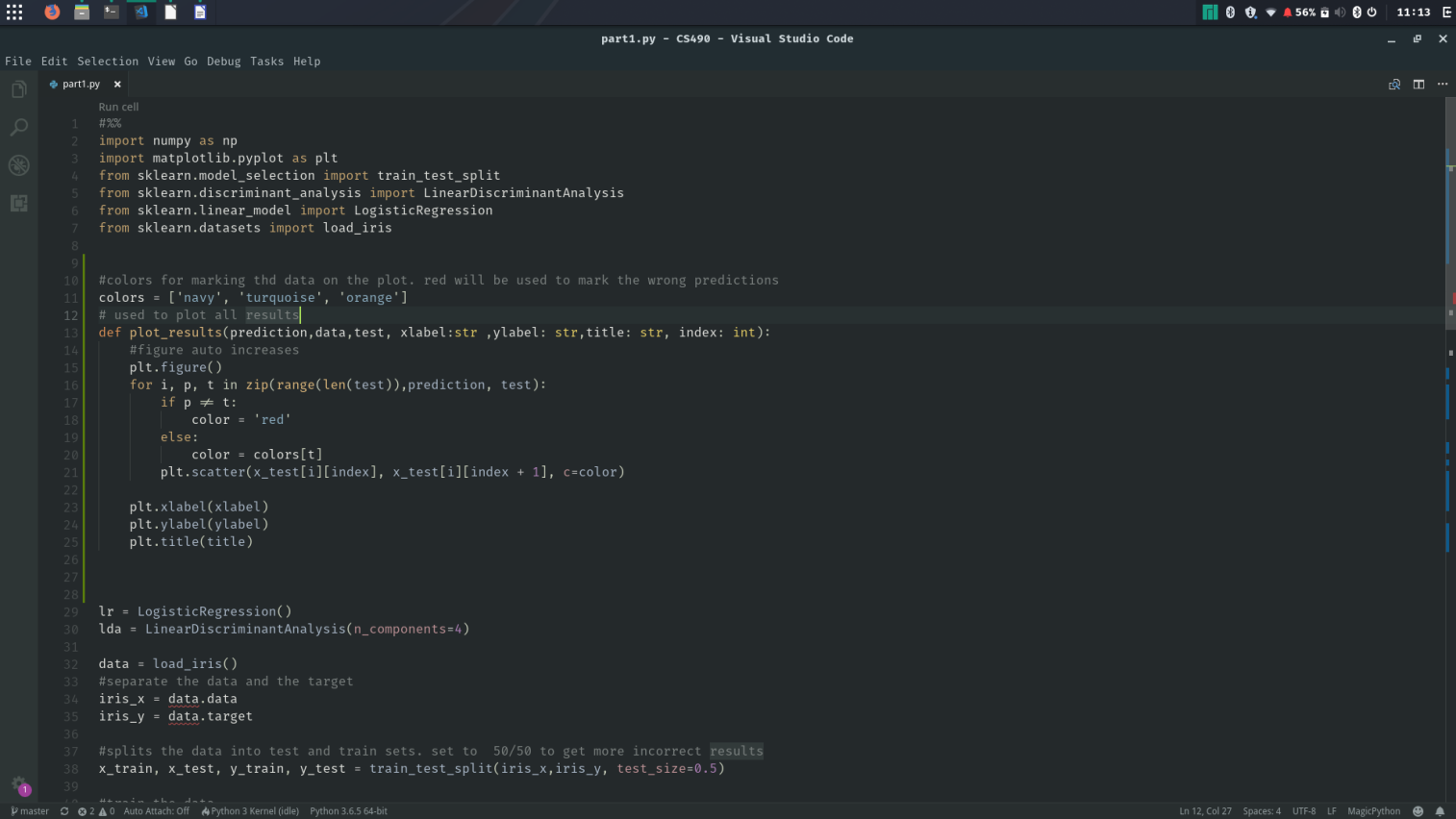
Until now the four python classes covered few topics with respect to the lab 2 assignment(fundamental parts of ML techinique). After each class, we reviewed the lecture and started with the related question in the lab. For each question, we worked independently and try to solve the problem, we created our own ideas, coded, discussed and finally picked a best answer, and everyone was participating in this lab with hundred percentage effort.

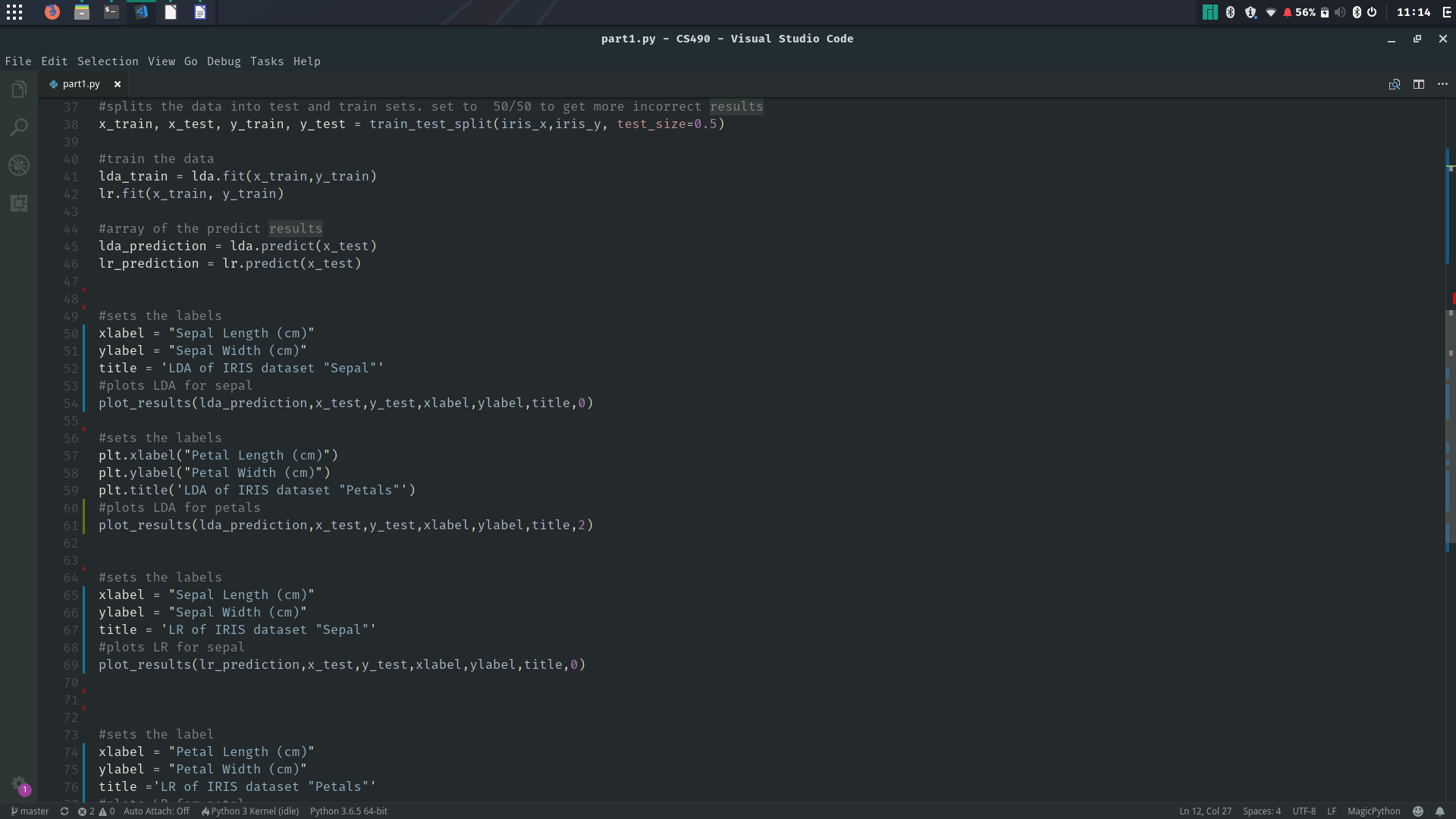
1. Datasets (if applicable)

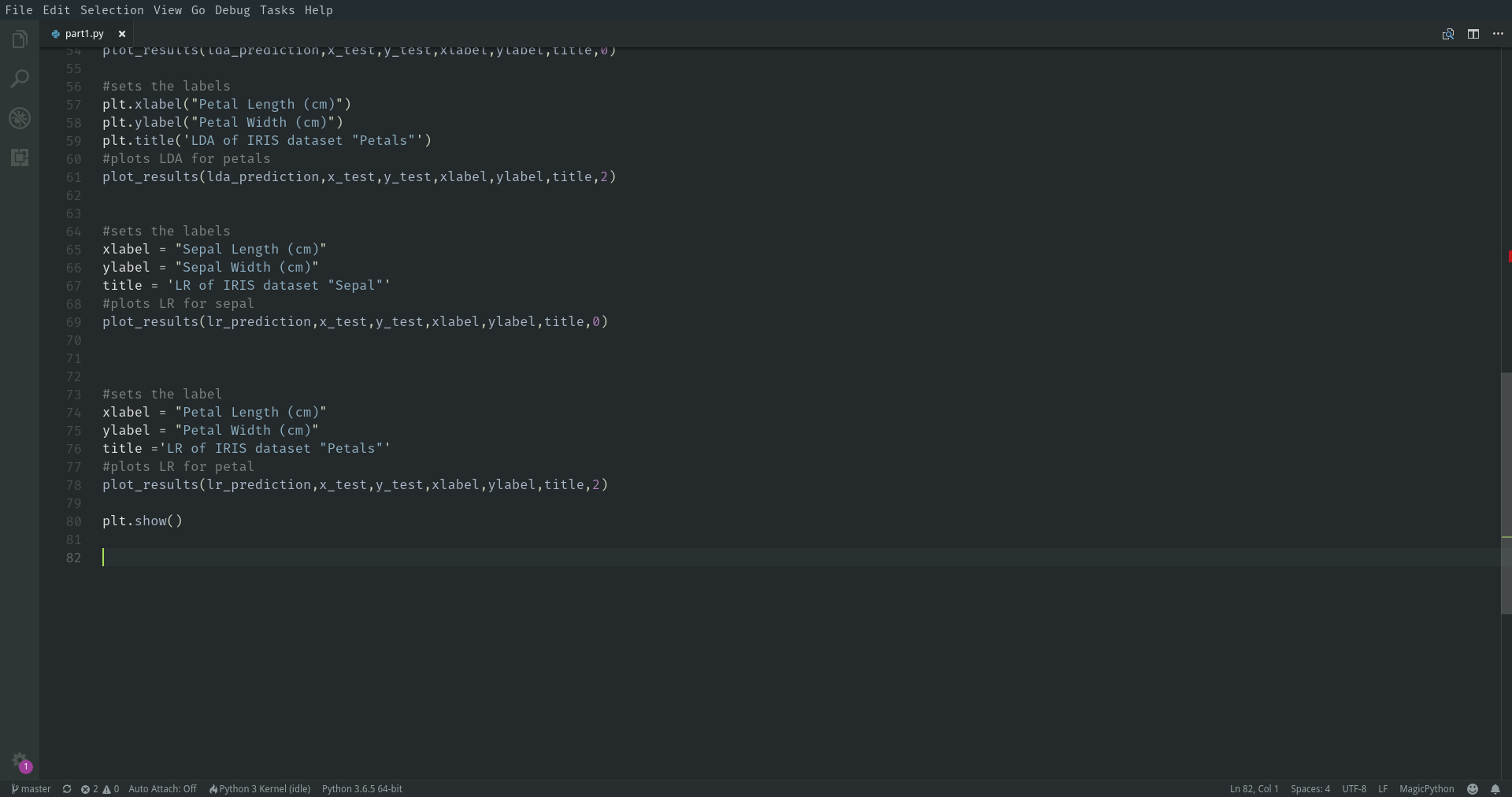
Part 1 and part 2: sklearn.datasets. iris. Part3: SimpleDoc.txt(see github)

1. Parameters/Results

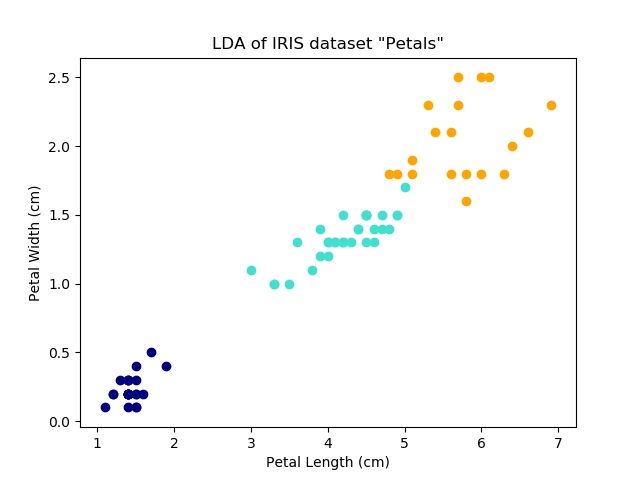
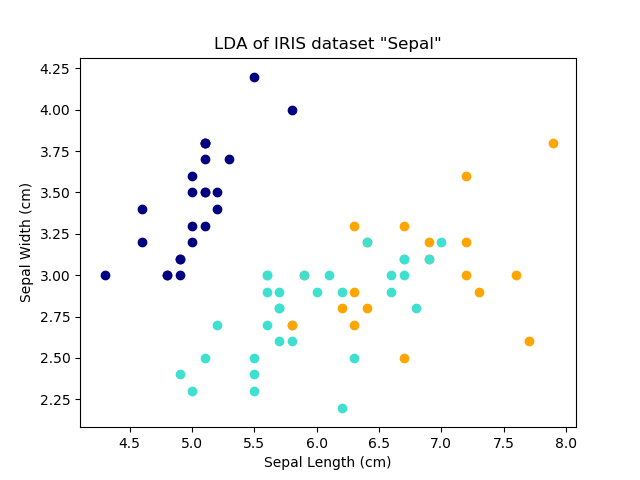
# Part 1-code:



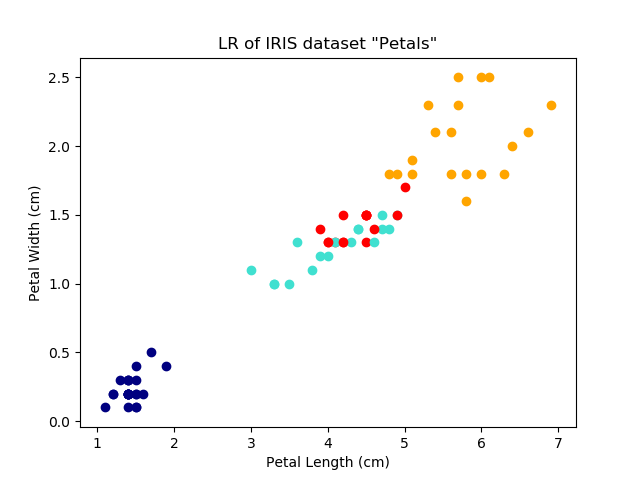
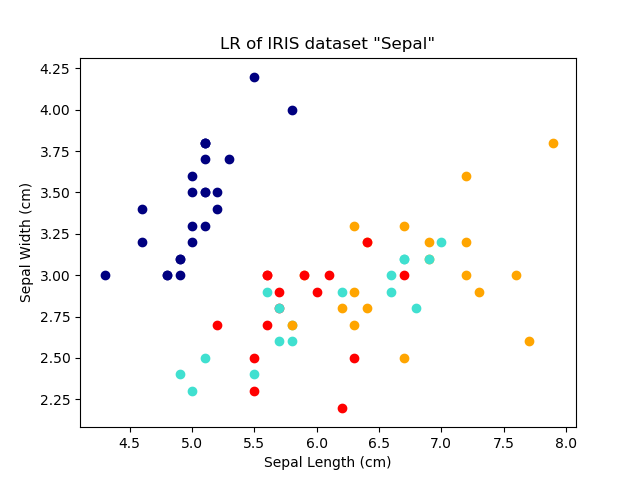




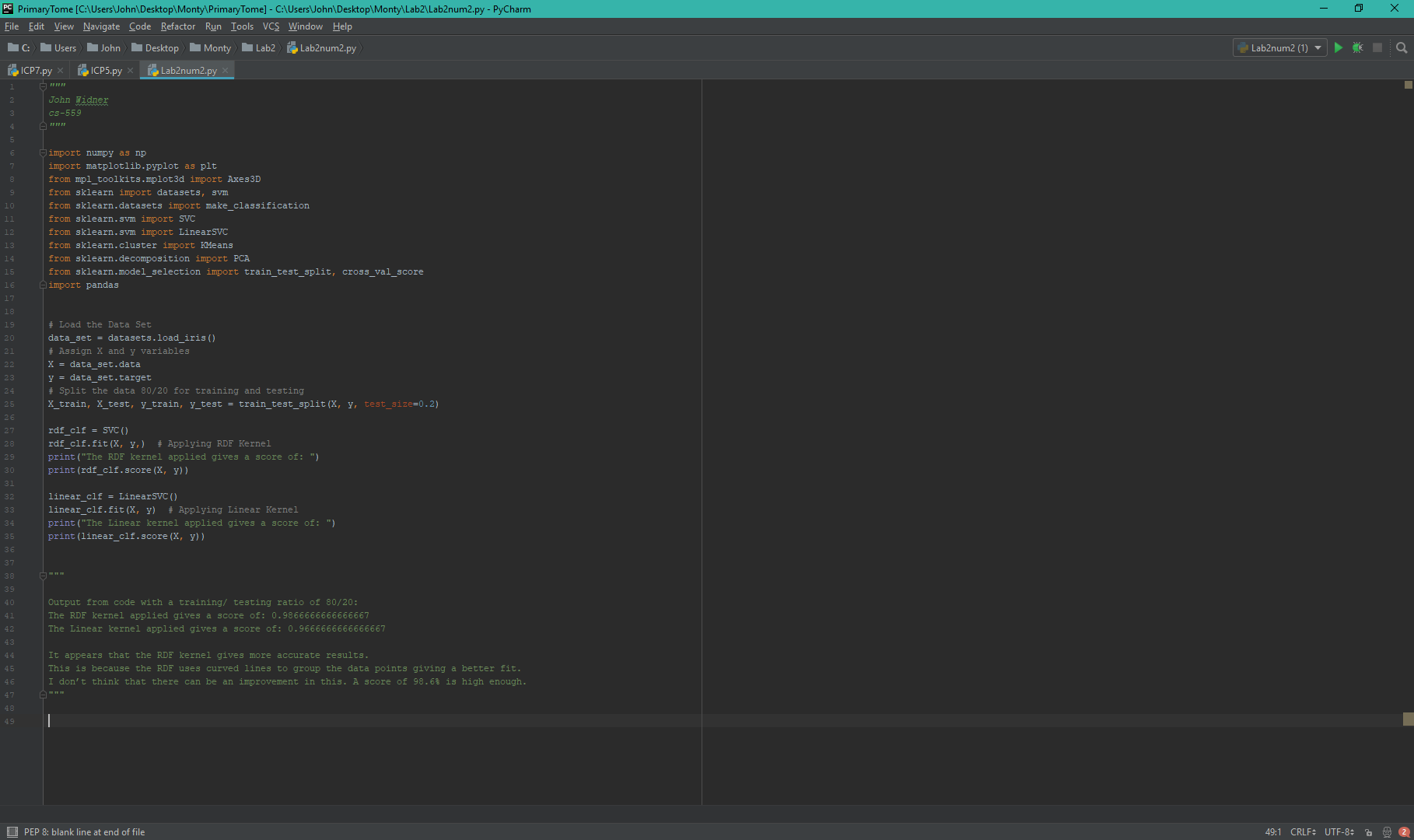
**Part 1-result:**

**Linear Discriminant Analysis. Wrong predictions are in red** 

**Logistic Regression results. Wrong predictions are in red:**

**Part 2-code:**



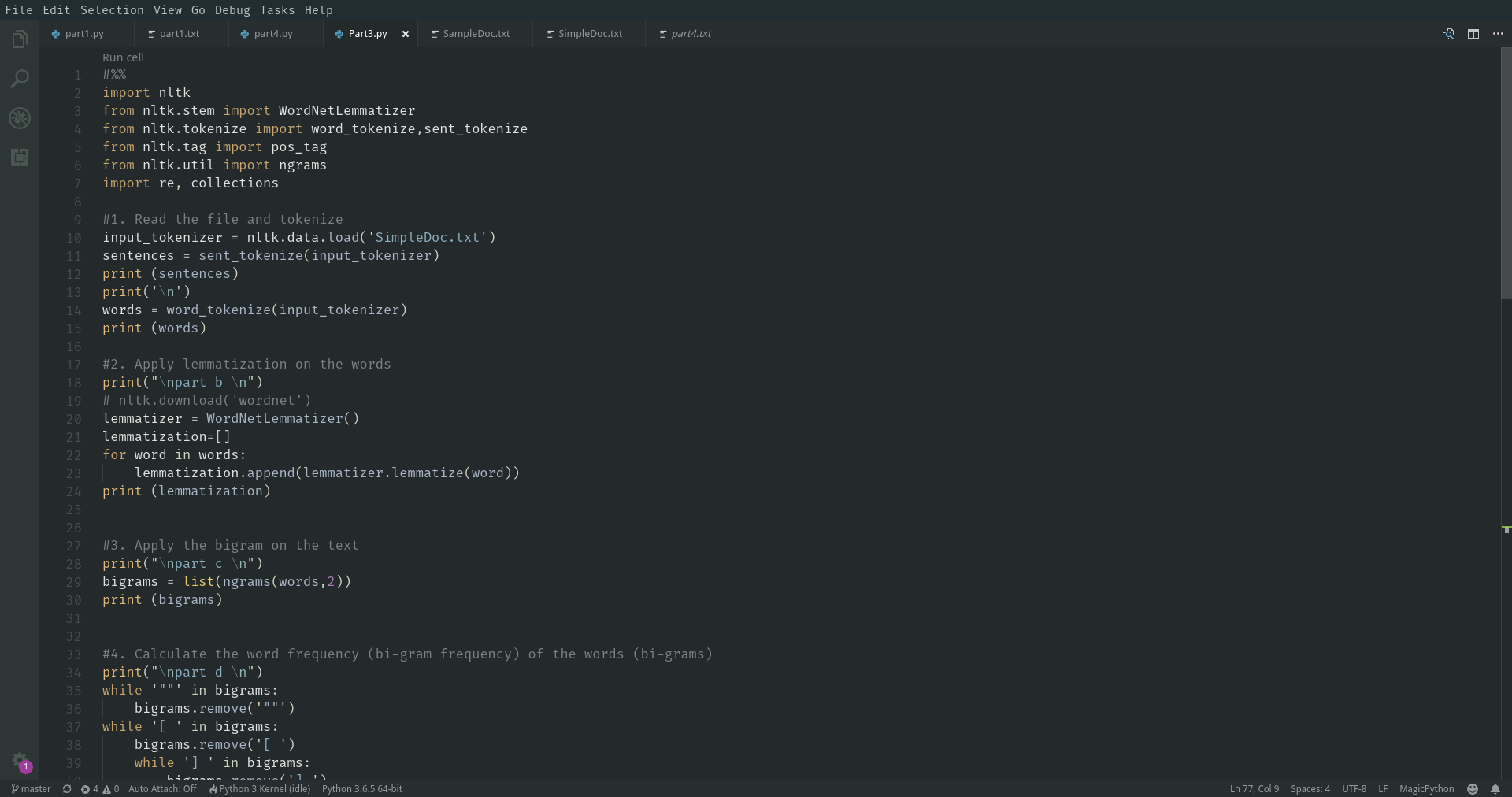
**Part 2-result:**

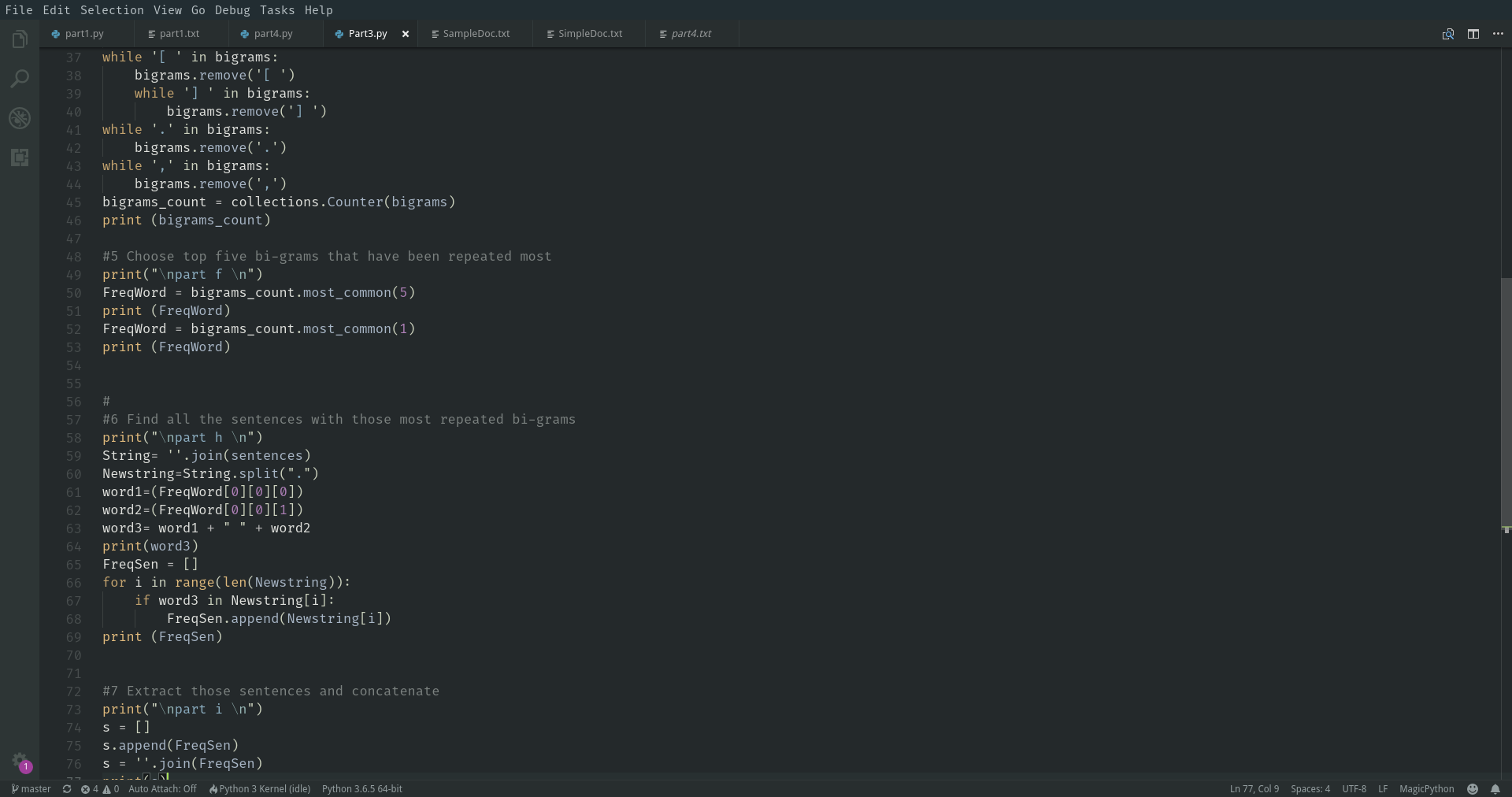
Output from code with a training/ testing ratio of 80/20:

The RDF kernel applied gives a score of: 0.9866666666666667

The Linear kernel applied gives a score of: 0.9666666666666667

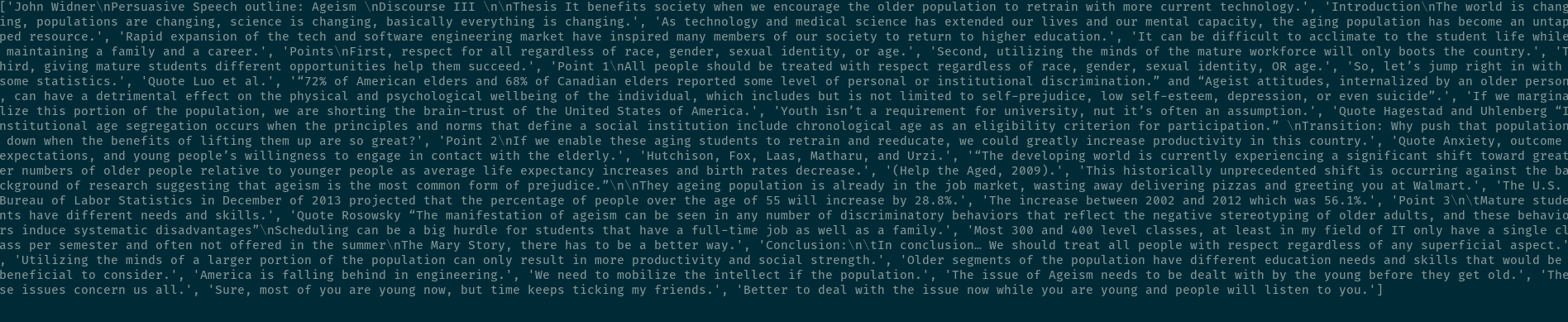
**Part 3-code:**



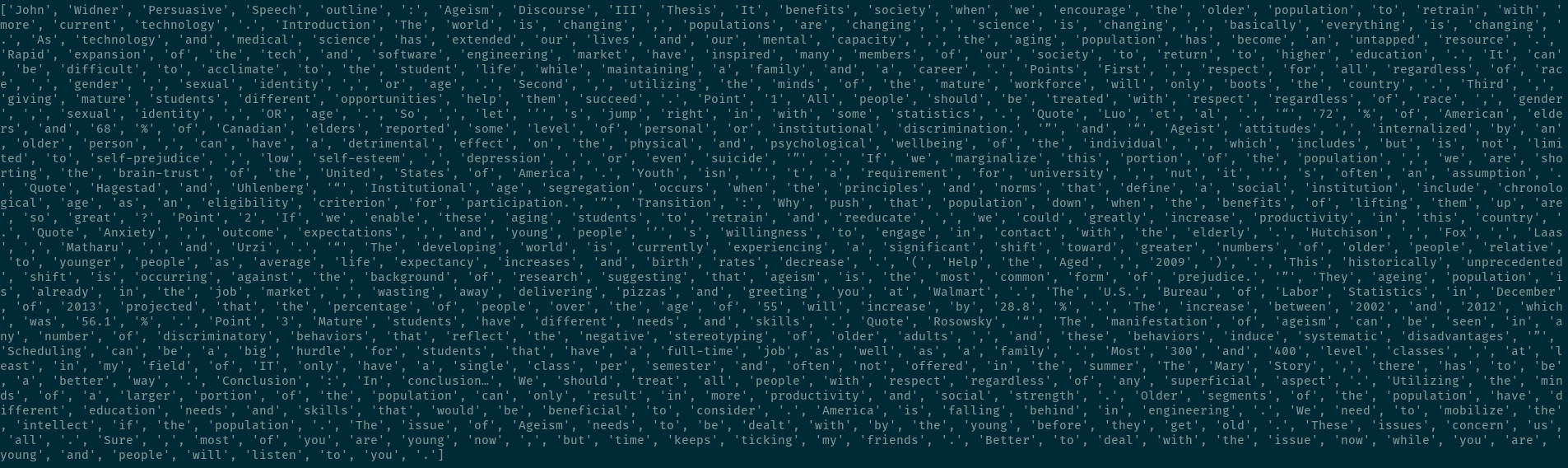


**Part 3-results:**

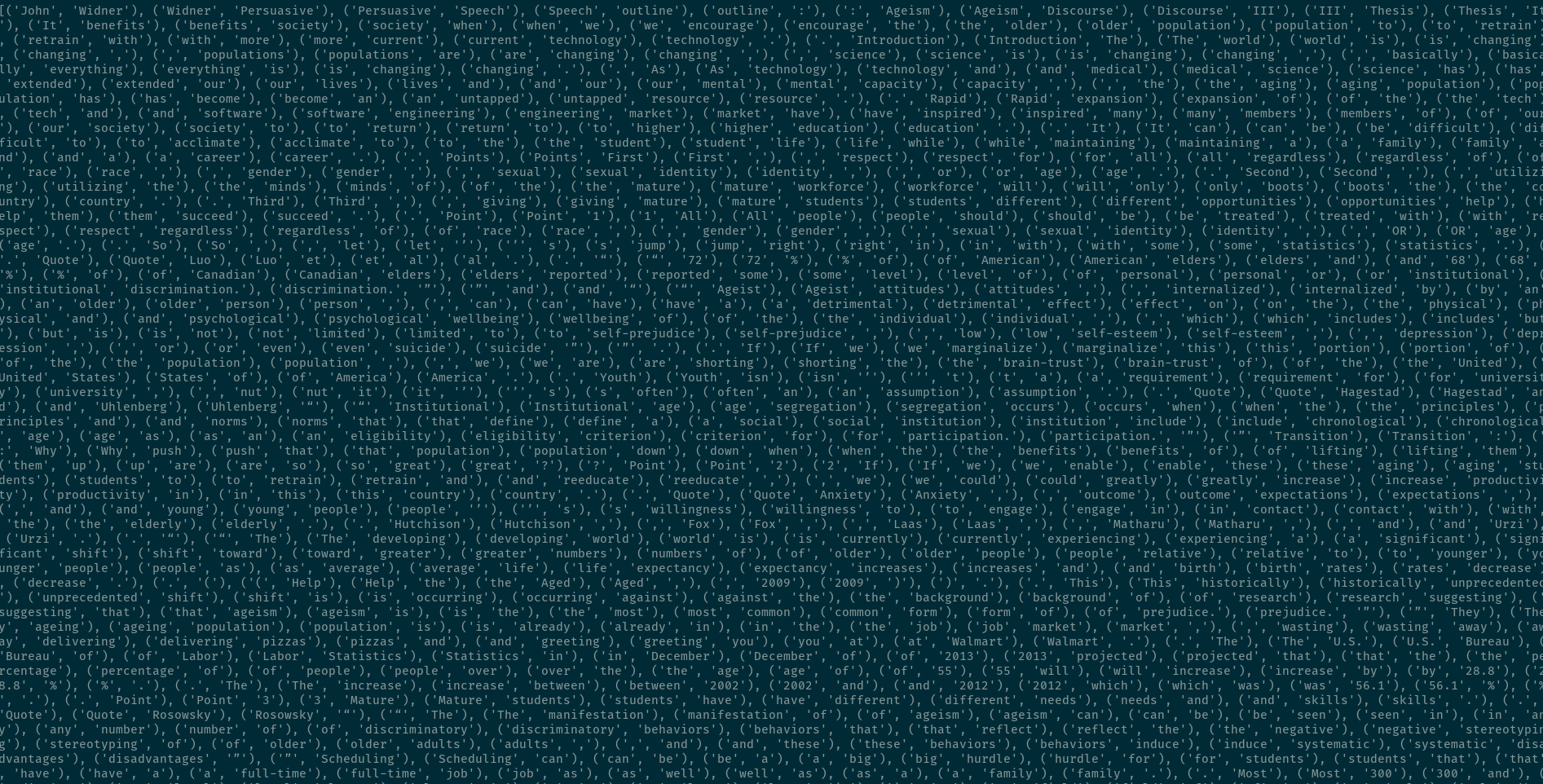
Tokenize the document



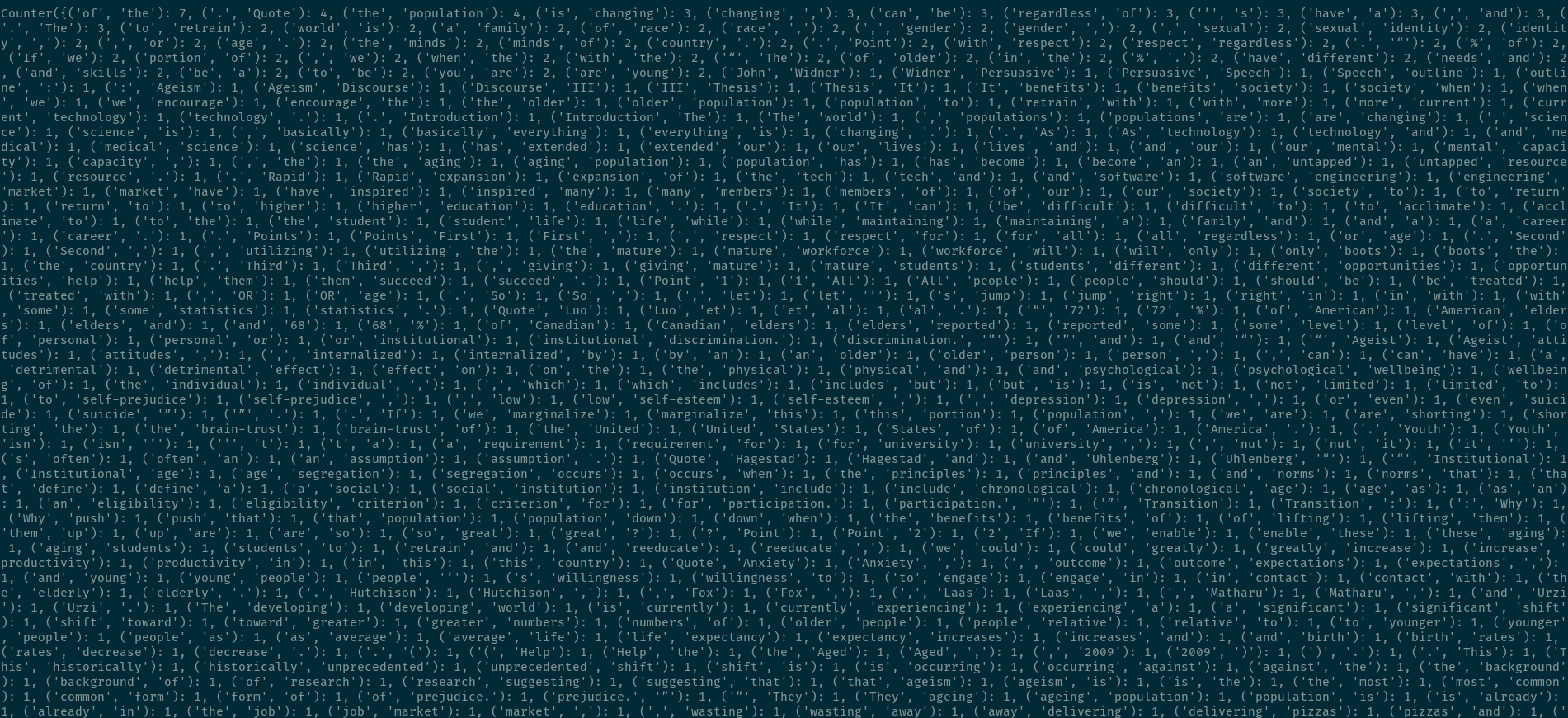
B. Apply lemmatization to the words



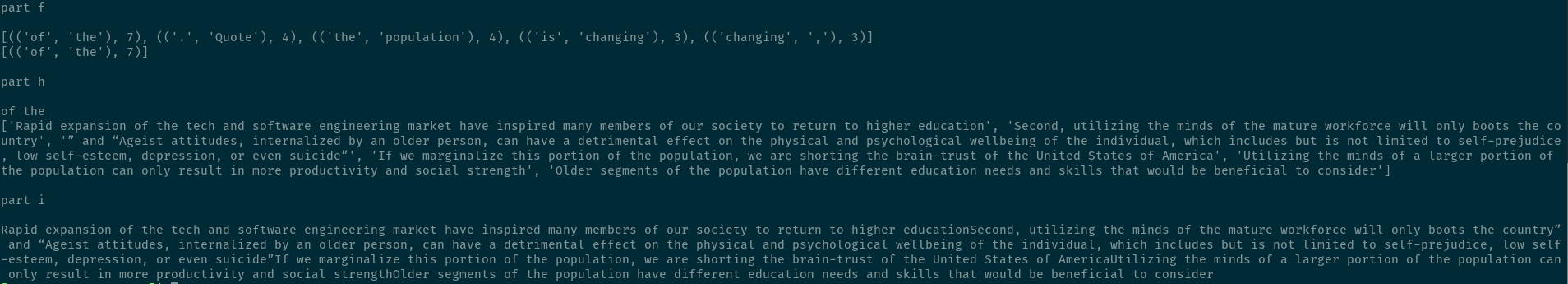
C. Apply bigram on the text



D. Calculate the frequency of the bigrams

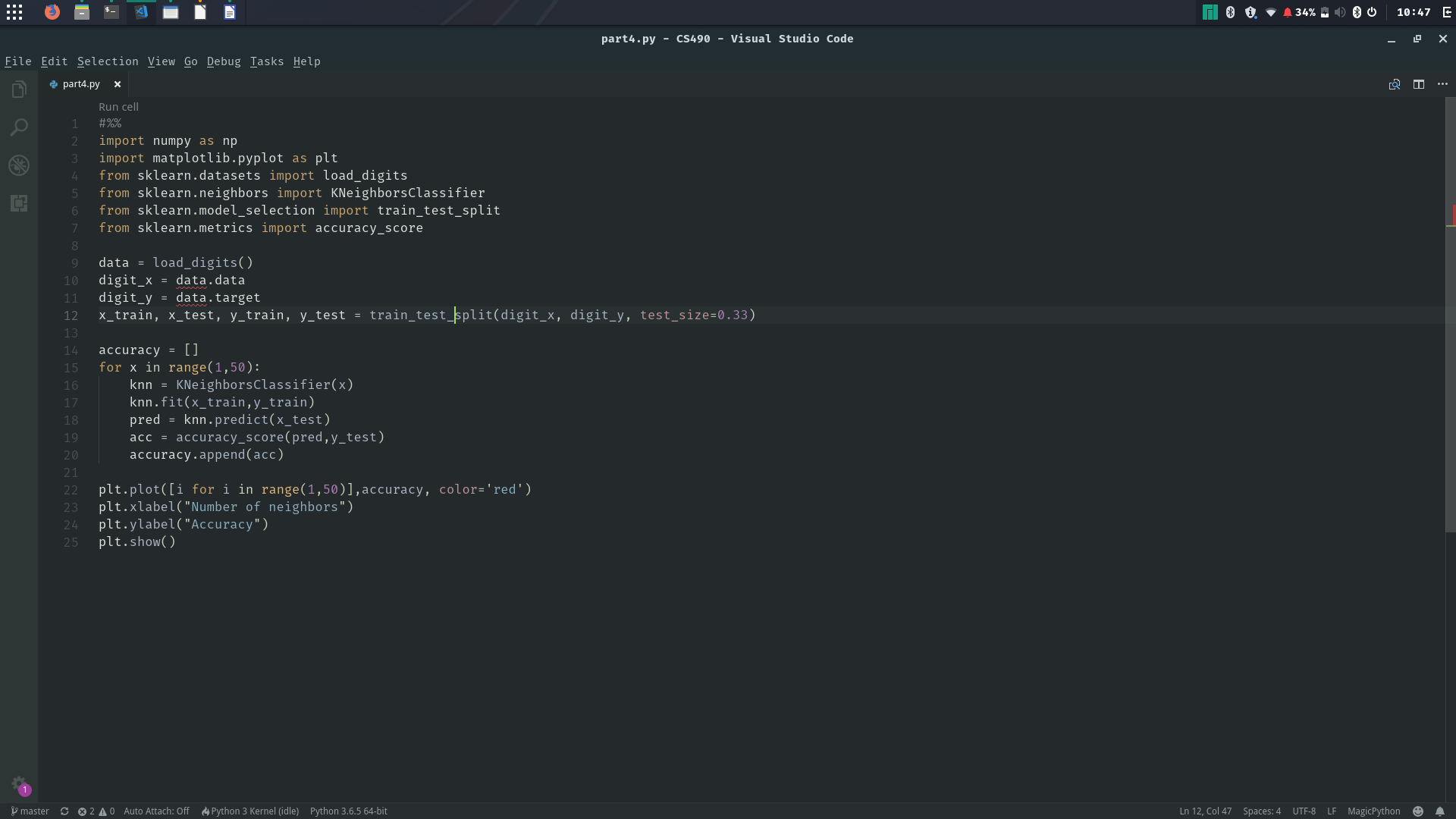


F-J. Finds the top 5 bigrams and finds the sentences with the most frequent bigrams and concatenate them.

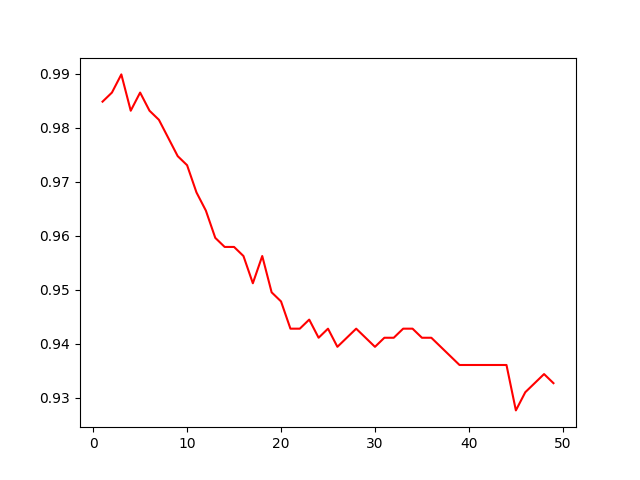


**Part 4-code:**

How does the size of K affect the accuracy of Knn



**Part 4-result:**



1. Evaluation & Discussion

**Part 1:** The differences between the results of Linear Discriminant Analysis and Logistic Regression could be caused by a number of items. One reason could be that Logistic Regression does not do quite as well as Linear Discriminant Analysis when the sample size is small and I chose a less than optimal split in the training and test sets. Another reason could be that the petals of the iris are well separated and Logistic Regression is not as stable as Linear Discriminant Analysis when this is the case. This could also explain the wide range of accuracy scores we got when using Logistic Regression.

**Part 2:** It appears that the RDF kernel gives more accurate results. This is because the RDF uses curved lines to group the data points giving a better fit. I don’t think that there can be an improvement in this. A score of 98.6% is high enough.

**Part 3:** It seems thatsummarization (processing) of a text file needs to start from tokenization although it was not required from the question, the separated words or sentences then can be continually processed by lemmatization or Bigram(or Trigram depends on requirements). Symbols or special characters need to be removed sometimes to reduce random errors.

**Part 4:** With Knn whether you want a larger K or a smaller K depends on the dataset, the type of data and the sample size of the data. In the cases I tried, when K was small it led to overfitting as outliers tended to have more influence than they should and, when K was large it led to underfitting as outliers were suppressed more than they should have been. With the digit dataset from SKlearn the accuracy of Knn peaked when K was around 6 and then steadily decreased as K increased.

1. Conclusion

All the requirements in lab 2 were met, and more program skills are needed which can be obtained by more practices.