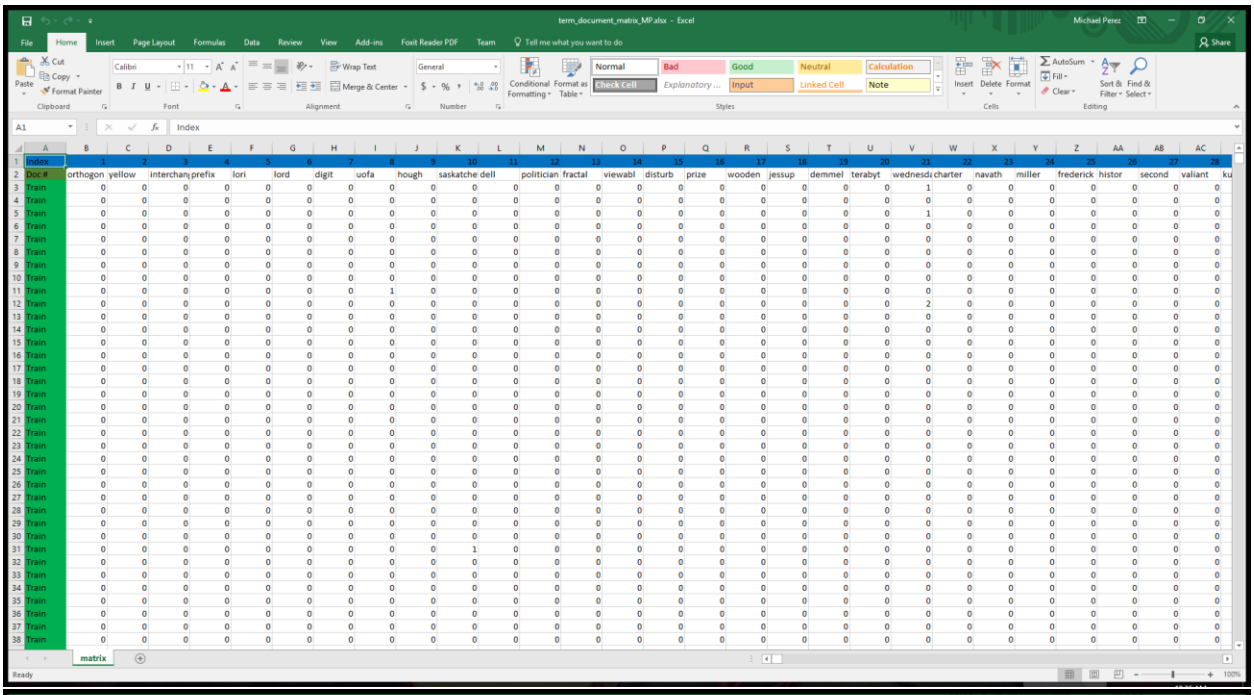
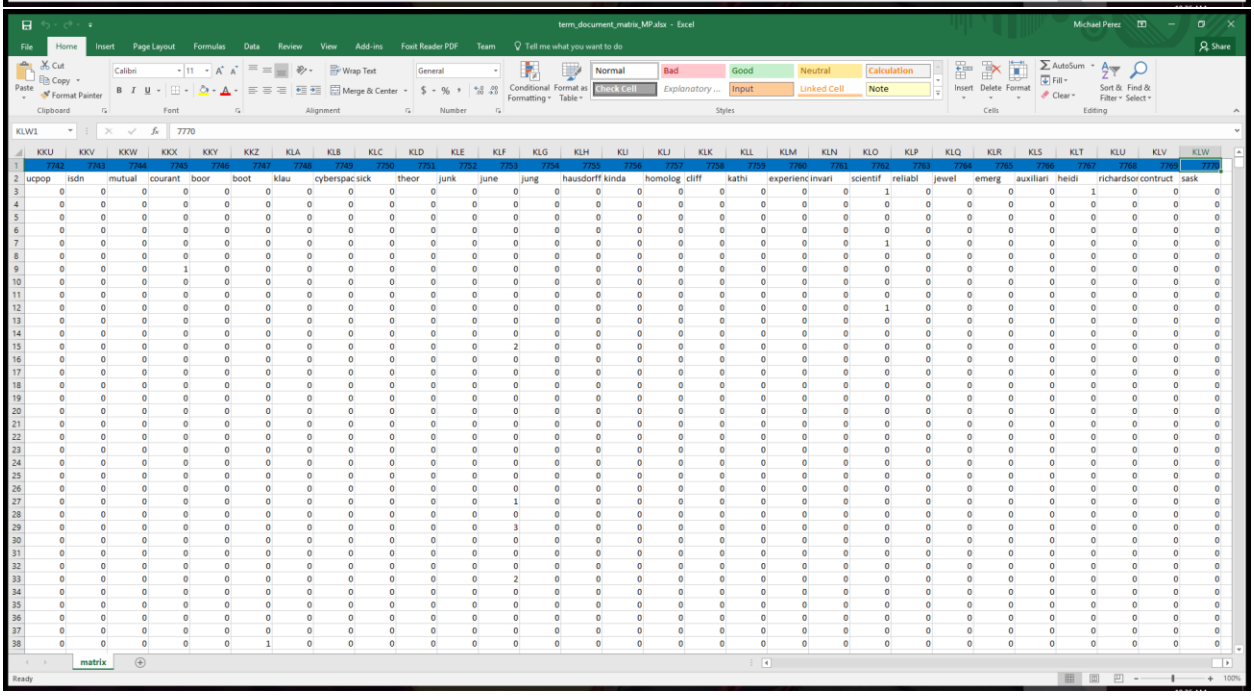
Michael Perez
Homework 2

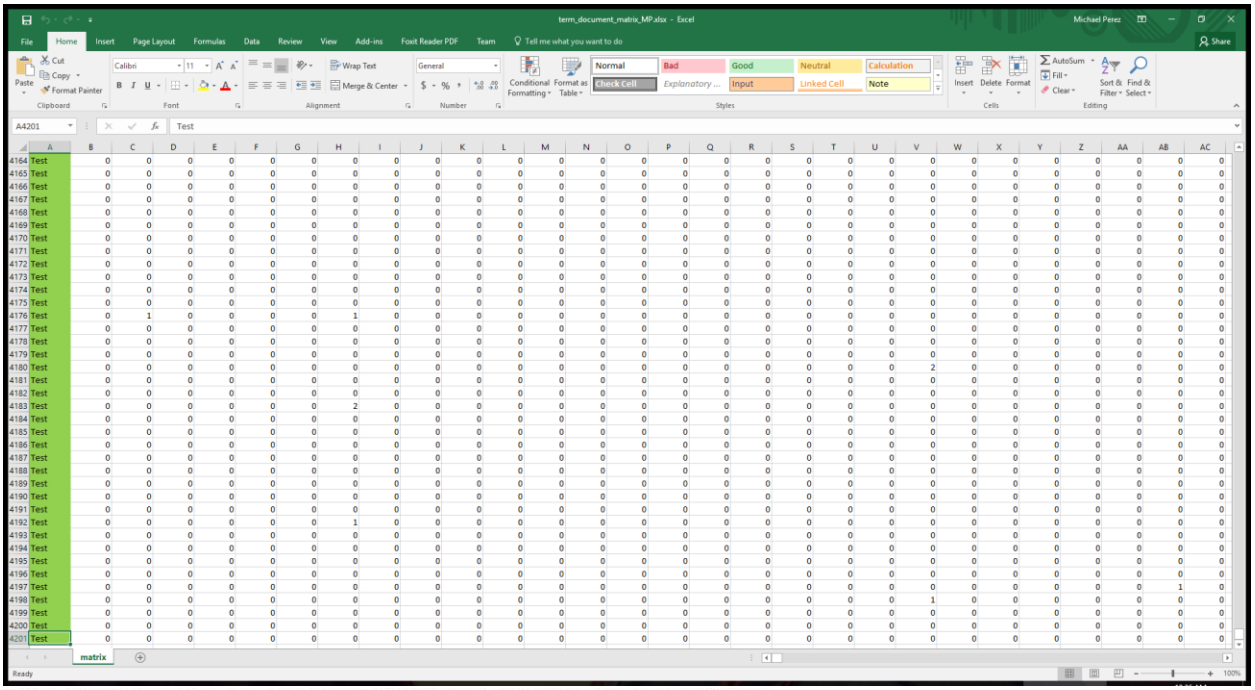
Natural Language Processing (Spring 2017)
Homework # 2
DUE: 2/26/2017 11:59pm EST Total: 10 points

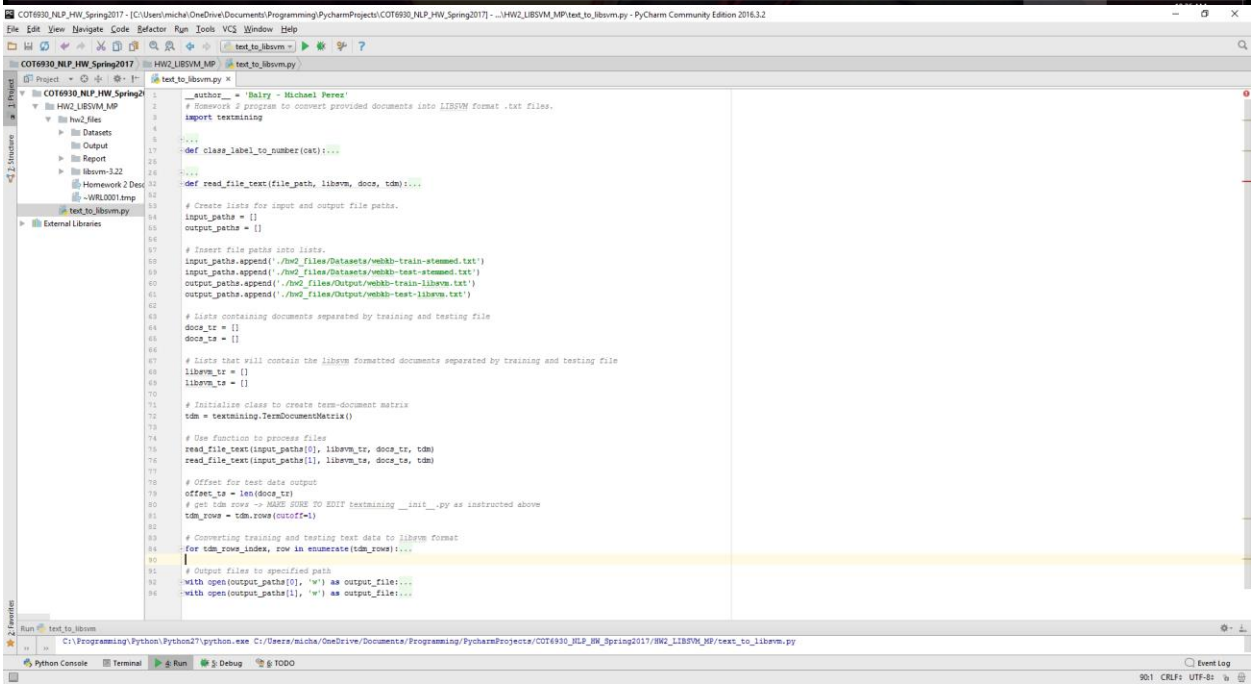
1) Term-Document Matrix

This turned out to be a large matrix when outputted into .csv and it is largely useless because of its sparsity. See screenshots below for a sample. Matrix will be attached, but it is 86.2 MB.
The code for making the conversion will be attached (made using python). Limited screenshots available below as well.

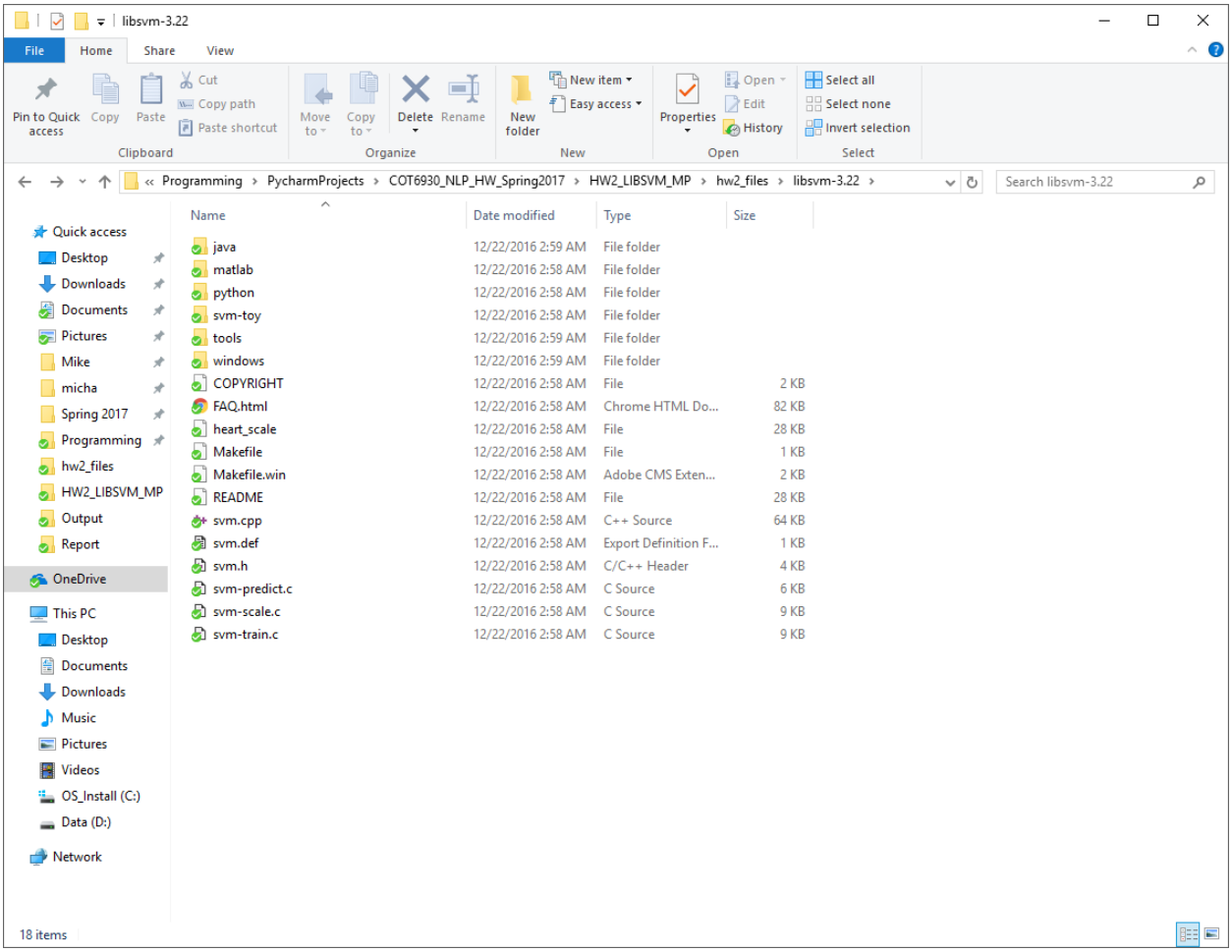





term_document_matrix_MP.xlsx - Excel
Michael Perez
matrix


COT6930_NLP_HW_Spring2017 - [C:\Users\micha\OneDrive\Documents\Programming\PycharmProjects\COT6930_NLP_HW_Spring2017] - ...\HW2_LIBSVM_MP\text_to_libsvm.py - PyCharm Community Edition 2016.3.2
__author__ = 'Balry - Michael Perez'
# Homework 2 program to convert provided documents into LIBSVM format .txt files.
import textmining
def class_label_to_number(cat):...
def read_file_text(file_path, libsvm, docs, tdm):...
# Create lists for input and output file paths.
input_paths = []
output_paths = []
# Insert file paths into lists.
input_paths.append('./hw2_files/Datasets/webkb-train-stemmed.txt')
input_paths.append('./hw2_files/Datasets/webkb-test-stemmed.txt')
output_paths.append('./hw2_files/Output/webkb-train-libsvm.txt')
output_paths.append('./hw2_files/Output/webkb-test-libsvm.txt')
# Lists containing documents separated by training and testing file
docs_tr = []
docs_ts = []
# Lists that will contain the libsvm formatted documents separated by training and testing file
libsvm_tr = []
libsvm_ts = []
# Initialize class to create term-document matrix
tdm = textmining.TermDocumentMatrix()
# Use function to process files
read_file_text(input_paths[0], libsvm_tr, docs_tr, tdm)
read_file_text(input_paths[1], libsvm_ts, docs_ts, tdm)
# Offset for test data output
offset_ts = len(docs_tr)
# get tdm rows -> MAKE SURE TO EDIT textmining __init__.py as instructed above
tdm_rows = tdm.rows(cutoff=1)
# Converting training and testing text data to libsvm format
for tdm_rows_index, row in enumerate(tdm_rows):...
Output files to specified path
with open(output_paths[0], 'w') as output_file:...
with open(output_paths[1], 'w') as output_file:...
C:\Programming\Python\Python27\python.exe C:/Users/micha/OneDrive/Documents/Programming/PycharmProjects/COT6930_NLP_HW_Spring2017/HW2_LIBSVM_MP/text_to_libsvm.py

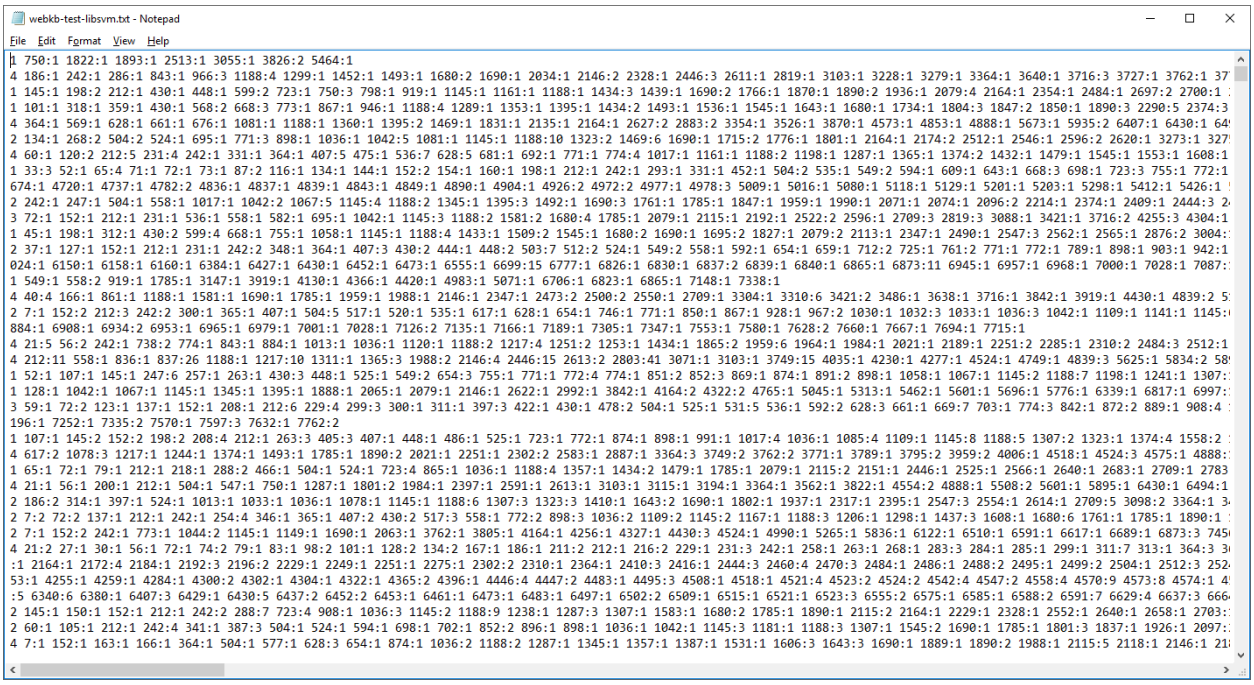
2) LIBSVM Download



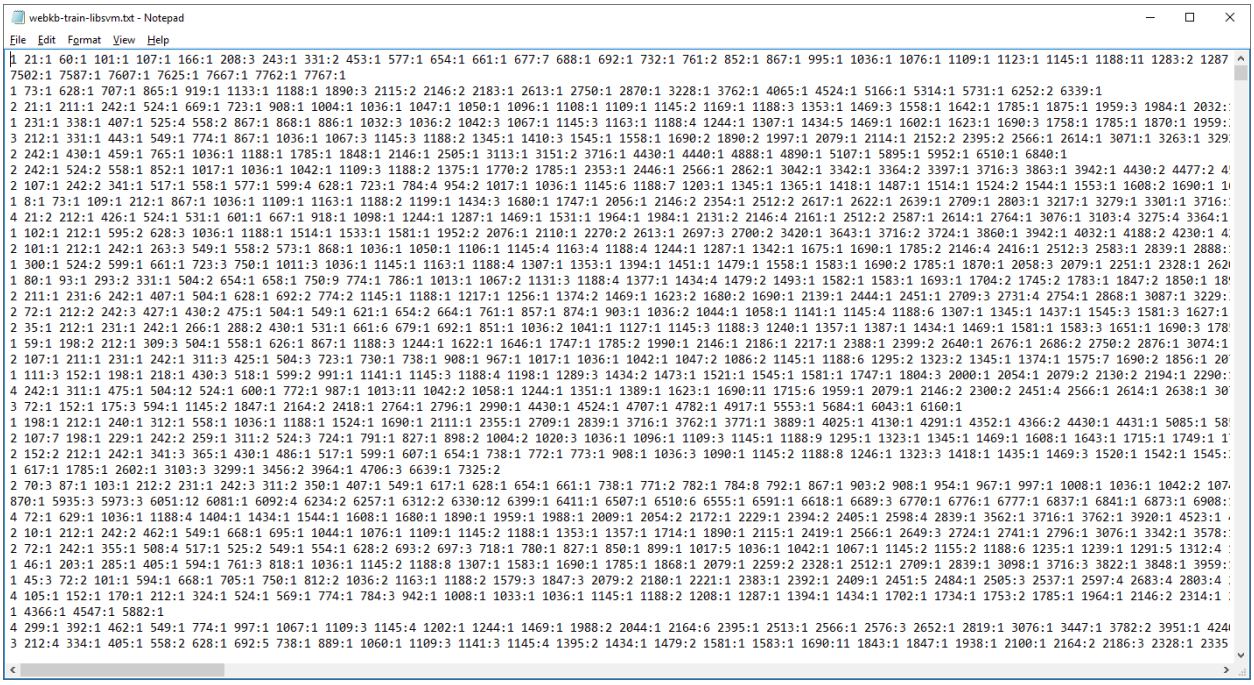
3) LIBSVM Formatted Text

Attached files under the following names “webkb-test-libsvm.txt” and “webkb-train-libsvm.txt”. Screenshots can be found below as well:

webkb-test-libsvm.txt



webkb-train-libsvm.txt



4) LIBSVM Training and Testing Data Demo

svm-train.exe -t 0 webkb-train-libsvm.txt

svm-predict.exe webkb-test-libsvm.txt webkb-train-libsvm.txt.model webkb_output_file_MP