Report

Emanuele Alessi 1486470

ANSWER 1

In this homework i have implemented word2vec model with skip-gram model. The architecture of the neural network is represented in 3 layers: the first layer contains as many neurons as the batch size; the second layer contains the embeddings matrix connected to the hidden neurons (the hidden neurons are as many as the dimension of the embeddign size), and the last layer contains as many neurons as is the size of the dictionary. In order to generate a good train set, i have implemented sentence splitting and sentence shuffling techniques .The training phase is based on the number of epochs given in input, in which the neural network is trained on the entire dataset.
The chosen parameters are the following:

- Window size = 2
- Iteration number per epoch = length of the dataset / batch size
- Embedding size = 128
- Batch size = 32
- Negative samples = 64
- Vocabulary size = 30000

ANSWER 2

After the training phase, i have obtained with 3 epochs a model with an accuracy score of 45% and a loss score of 3,3%. I found with Tensorboard the similar words required in the homework. The similarities are the following:

	top 1	**top 2**	**top 3**	**top 4**	**top 5**
german	austrian	swedish	dutch	danish	russian
most	historically	highly	nonetheless	generally	extremely
general	appointed	representative	special	lieutenant	terms
food	foods	nutrition	meat	dairy	milk
cat	cats	dogs	dog	pet	rabbits
eat	eaten	ate	prefer	feed	grow
teach	teaching	learn	taught	teaches	attend

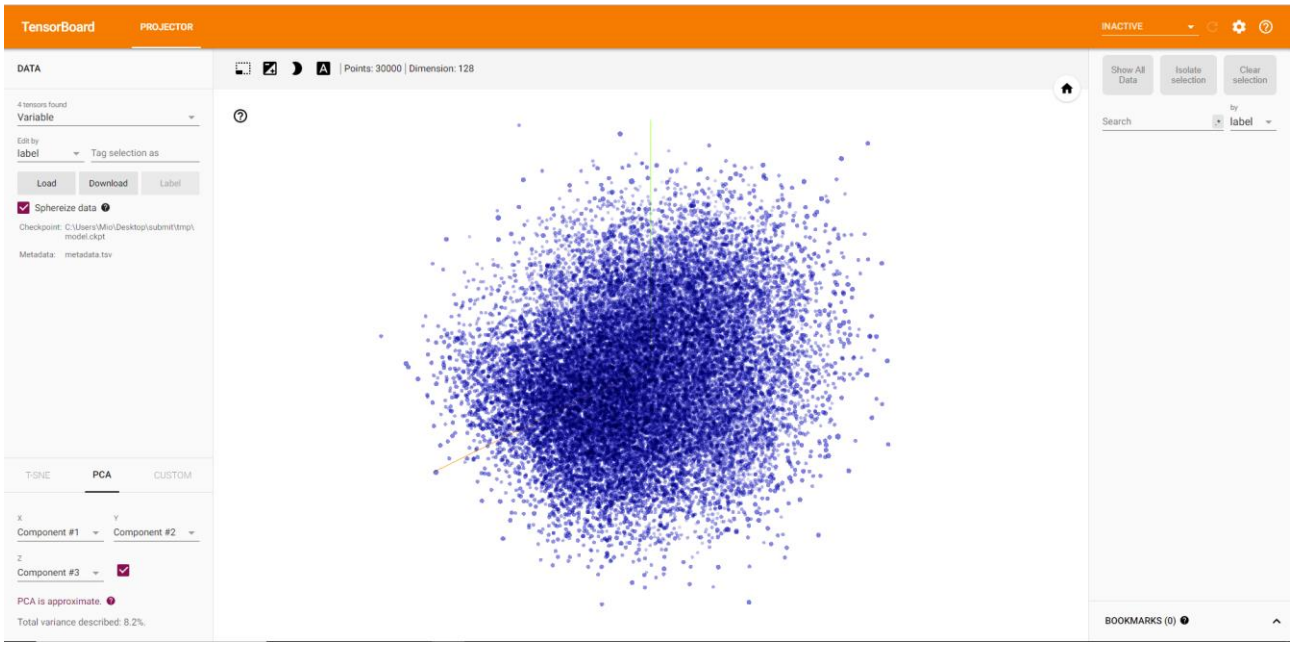
ANSWER 3

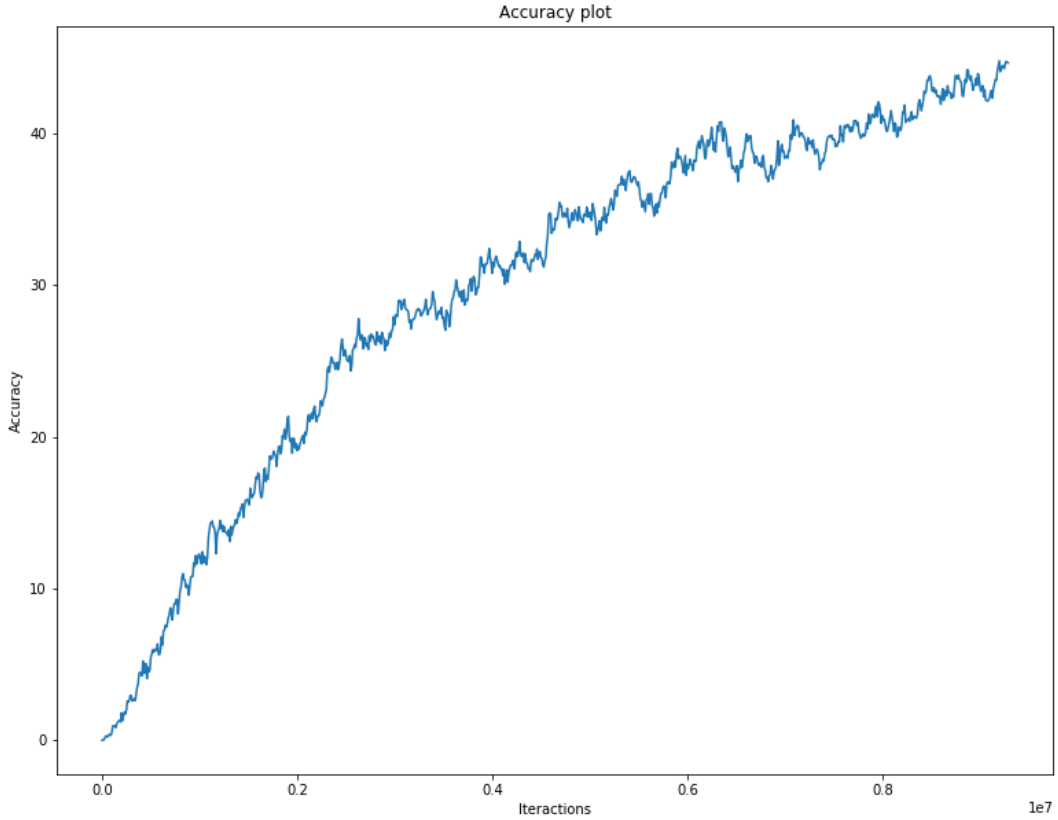
I used my embeddings vectors to calculate the mean of the word embeddings. First of all, for each document in f (where f = TRAIN or DEV or TEST) i preprocess the text inside it in the same way as it is preprocessed before training the neural network, then i calculate the sum of each preprocessed word in the document and I divide it by the number of preprocessed words, finally i add another column to the vector to save the domain of the document. The resulting vector is saved inside the dataset that will be used for the domain classification.

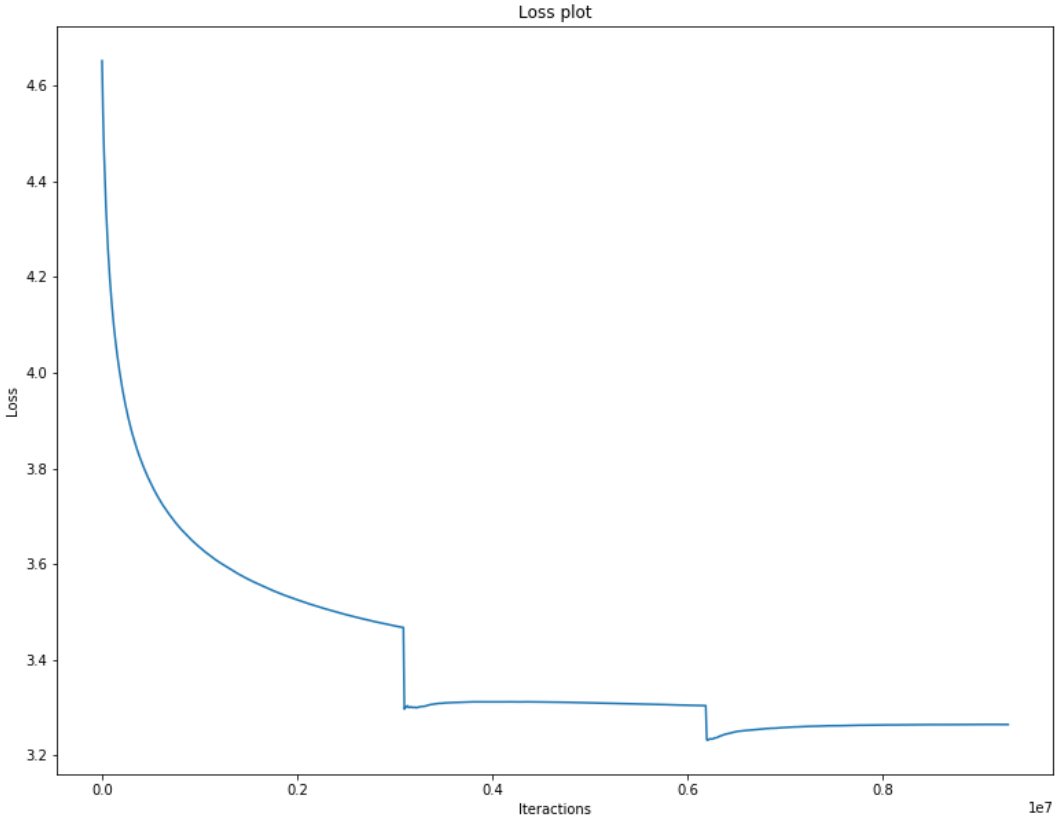
Images concerning the classification report and the confusion matrix have been inserted on page *4* and *5*.

In the next pages I've inserted the images of vector space, accuracy plot, loss plot, classification report and confusion matrix plot.

NOTICE: in *notebook.ipynb* you can see an accurate documentation about the procedures with which I have calculated the domain classification.

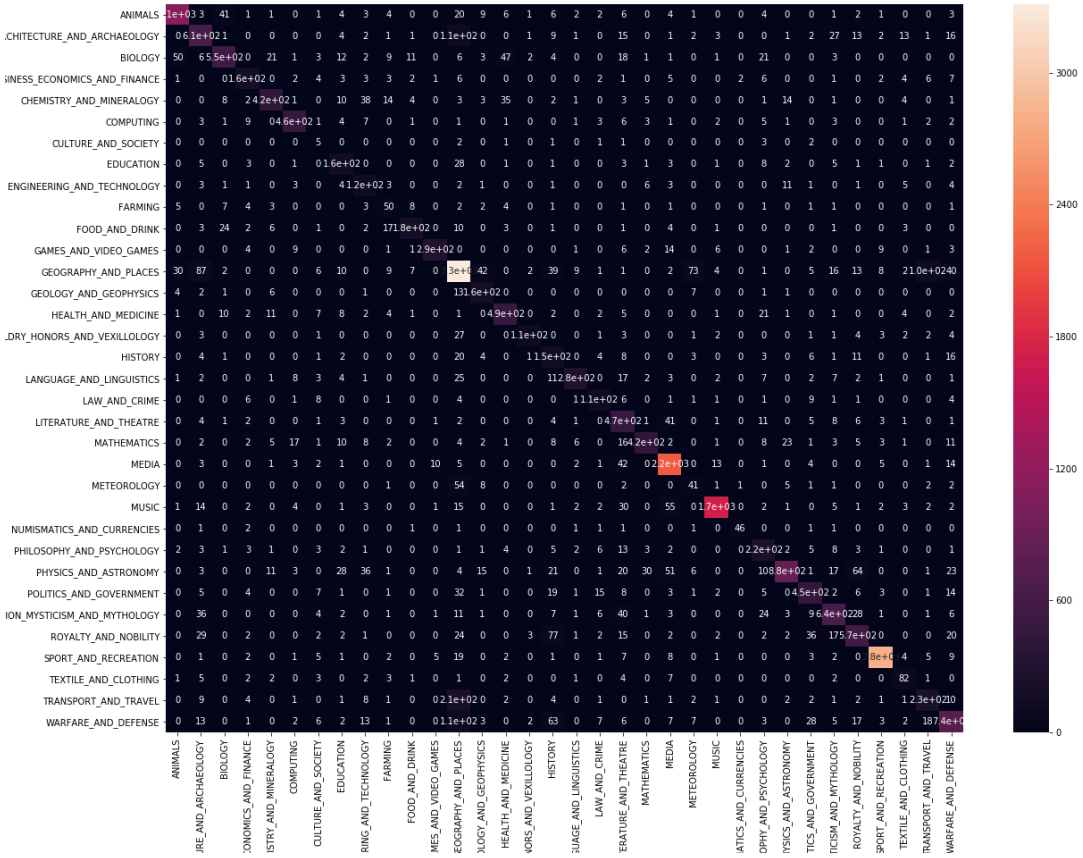

TensorBoard
PROJECTOR
INACTIVE
DATA
4 tensors found
Variable
Edit by
label
Tag selection as
Load
Download
Label
Sphereize data
Checkpoint: C:\Users\Mio\Desktop\submit\tmp\model.ckpt
Metadata: metadata.tsv
T-SNE
PCA
CUSTOM
X
Component #1
Y
Component #2
Z
Component #3
PCA is approximate.
Total variance described: 8.2%.
Points: 30000 | Dimension: 128
Show All Data
Isolate selection
Clear selection
Search
by
label
BOOKMARKS (0)


Accuracy plot
Accuracy
Iteractions
1e7
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8


Loss plot
Loss
Iteractions
1e7
3.2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
0.0
0.2
0.4
0.6
0.8

	precision	recall	f1-score	support
ANIMALS	0.92	0.90	0.91	1241
ART_ARCHITECTURE_AND_ARCHAEOLOGY	0.71	0.73	0.72	841
BIOLOGY	0.85	0.71	0.78	776
BUSINESS_ECONOMICS_AND_FINANCE	0.72	0.72	0.72	217
CHEMISTRY_AND_MINERALOGY	0.86	0.74	0.79	569
COMPUTING	0.89	0.89	0.89	515
CULTURE_AND_SOCIETY	0.07	0.31	0.11	16
EDUCATION	0.57	0.70	0.62	222
ENGINEERING_AND_TECHNOLOGY	0.46	0.70	0.56	167
FARMING	0.39	0.53	0.45	95
FOOD_AND_DRINK	0.82	0.69	0.75	258
GAMES_AND_VIDEO_GAMES	0.94	0.83	0.88	354
GEOGRAPHY_AND_PLACES	0.81	0.87	0.84	3827
GEOLOGY_AND_GEOPHYSICS	0.62	0.81	0.70	192
HEALTH_AND_MEDICINE	0.82	0.85	0.83	577
HERALDRY_HONORS_AND_VEXILLOLOGY	0.90	0.67	0.77	166
HISTORY	0.34	0.63	0.44	232
LANGUAGE_AND_LINGUISTICS	0.89	0.73	0.80	376
LAW_AND_CRIME	0.65	0.70	0.67	155
LITERATURE_AND_THEATRE	0.61	0.83	0.70	571
MATHEMATICS	0.88	0.75	0.81	564
MEDIA	0.91	0.95	0.93	2273
METEOROLOGY	0.28	0.34	0.31	119
MUSIC	0.97	0.92	0.95	1861
NUMISMATICS_AND_CURRENCIES	0.94	0.81	0.87	57
PHILOSOPHY_AND_PSYCHOLOGY	0.60	0.75	0.67	294
PHYSICS_AND_ASTRONOMY	0.93	0.72	0.81	1223
POLITICS_AND_GOVERNMENT	0.78	0.77	0.78	580
RELIGION_MYSTICISM_AND_MYTHOLOGY	0.82	0.77	0.80	823
ROYALTY_AND_NOBILITY	0.76	0.71	0.73	811
SPORT_AND_RECREATION	0.98	0.97	0.98	2899
TEXTILE_AND_CLOTHING	0.62	0.69	0.65	119
TRANSPORT_AND_TRAVEL	0.60	0.47	0.53	495
WARFARE_AND_DEFENSE	0.77	0.70	0.73	1061
avg / total	0.84	0.82	0.83	24546

Accuracy: 0.82%


ANIMALS
ARCHITECTURE_AND_ARCHAEOLOGY
BIOLOGY
BUSINESS_ECONOMICS_AND_FINANCE
CHEMISTRY_AND_MINERALOGY
COMPUTING
CULTURE_AND_SOCIETY
EDUCATION
ENGINEERING_AND_TECHNOLOGY
FARMING
FOOD_AND_DRINK
GAMES_AND_VIDEO_GAMES
GEOGRAPHY_AND_PLACES
GEOLOGY_AND_GEOPHYSICS
HEALTH_AND_MEDICINE
HERALDRY_HONORS_AND_VEXILLOLOGY
HISTORY
LANGUAGE_AND_LINGUISTICS
LAW_AND_CRIME
LITERATURE_AND_THEATRE
MATHEMATICS
MEDIA
METEOROLOGY
MUSIC
NUMISMATICS_AND_CURRENCIES
PHILOSOPHY_AND_PSYCHOLOGY
PHYSICS_AND_ASTRONOMY
POLITICS_AND_GOVERNMENT
RELIGION_MYSTICISM_AND_MYTHOLOGY
ROYALTY_AND_NOBILITY
SPORT_AND_RECREATION
TEXTILE_AND_CLOTHING
TRANSPORT_AND_TRAVEL
WARFARE_AND_DEFENSE
3000
2400
1800
1200
600
0