CS4361/5361 Machine Learning

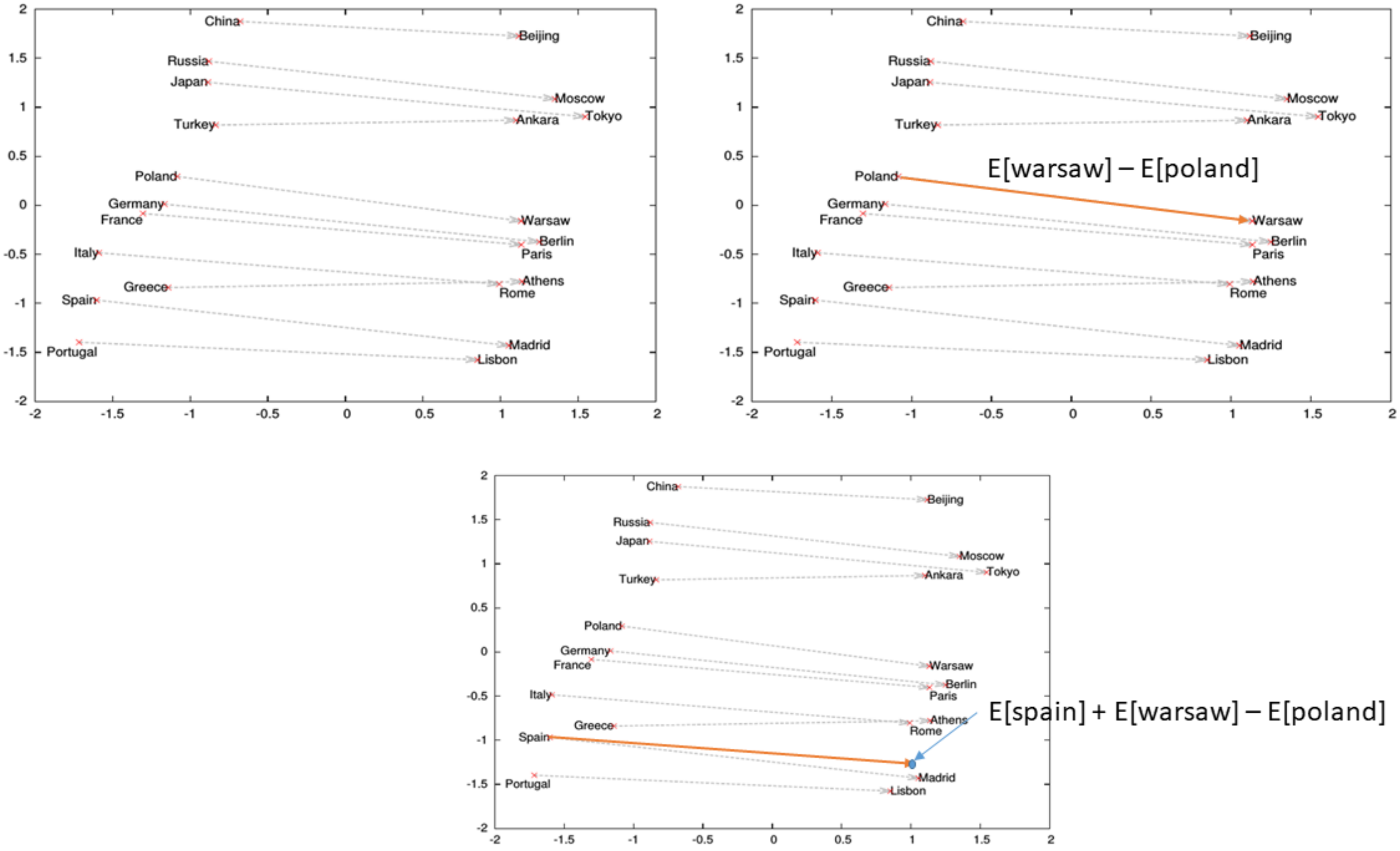
Fall 2020

Exam 2, Part 2

Part 2 has 2 questions and 2 deadlines. Submit your answer to one of the questions (your choice) by 4:20 and your answer to the other by the end of the day. Submit in separate programs using the following names: *lastname_firstname_wordanalogies.py* and *lastname_firstname_cluster_embeddings.py*. If you submit your answers to both questions by 4:30 today you will receive 20% extra credit.

1. Consider the following analogy question:
   Poland is to Warsaw as Spain is to:
   a) Beijing
   b) Moscow
   c) Paris
   d) Madrid

   We could write a program to answer this type of questions using word embeddings and simple geometry as follows:



The original relationship is represented by the vector E['warsaw'] - E['poland']
If we add that vector to the embedding of the first member of the goal analogy, E['spain'], we get to a point in embedding space that should be close to the result of the analogy. Thus the solution to the question is:
Of the four options (Beijing, Moscow, Paris, Madrid), whose embedding is closest to the point
E['spain'] + E['warsaw'] - E['poland']
Visually, we can observe that the answer is Madrid.

Your task is to implement a program to solve this type of problem. Starter code is provided.

2. Cluster the word embeddings in the dictionary into 10 clusters using the sklearn implementation of k-means. For each cluster, output the word whose embedding is closest to the cluster center.