EEM 409 Makine Öğrenmesi'ne Giriş
EEE 6113 Machine Learning

Clustering Similarity Python Assignment

You are going to do three tasks in this assignment.

1) **Load data:** Load the shared *people_wiki.sframe* data and analyze.
2) **Represent the data using word counts and TF-IDF**: Use two document representations (*word counts* and *TF-IDF*) to represent the data.
3) **Compare top words according to word counts to TF-IDF:** Take a particular famous person, say 'Elton John'.
   a) What are the 3 words in his articles with highest word counts?
   b) What are the 3 words in his articles with highest TF-IDF?
   c) Which one do you think is more useful in finding important words? Why?
4) **Measuring distance: C**ompute the distance between Elton John's article and those of two other famous singers. In this assignment, you will use the *cosine distance*, which measure similarity between vectors.
   a) What is the cosine distance between the articles on 'Elton John' and 'Victoria Beckham'?
   b) What is the cosine distance between the articles on 'Elton John' and Paul McCartney'?
   c) Which one of the two is closest to Elton John?
   d) Does this result make sense to you? Why?
5) **Building nearest neighbors (kNN) models with different input features and setting the distance metric:** Build two nearest neighbors (kNN) models for retrieving articles; one model by using word counts as features and another one using TF-IDF as features. In both of these models, set the distance function to *cosine similarity*.

Use these two models to retrieve documents in order to collect the following results:

   a) What is the most similar article, other than itself, to the one on 'Elton John' using word count features?
   b) What is the most similar article, other than itself, to the one on 'Elton John' using TF-IDF features?
   c) What is the most similar article, other than itself, to the one on 'Victoria Beckham' using word count features?
   d) What is the most similar article, other than itself, to the one on 'Victoria Beckham' using TF-IDF features?