Homework 6:
Supervised Learning: K nearest neighbors

Marina Sedinkina,Benjamin Roth
Symbolische Programmiersprache

Due: Thursday December 6, 2018, 16:00

In this exercise you will:

- implement k nearest neighbours classifier

Exercise 1: K nearest neighbours [6 points]

Train k nearest neighbours classifier using training set of newsgroups data and classify test documents (test set) into one of the 20 newsgroups.

Download and unpack `20news-bydate.tar.gz` - 20 Newsgroups sorted by date from `http://qwone.com/~jason/20Newsgroups/` into the `data/` folder of your project. **Do not push downloaded data in Git!!!** The dataset contains train and test folders consisting of several newsgroups folders and their documents. Take a look at the data and the file `hw06_knn/classification.py`. In this exercise you will have to complete some methods to make the classification work.

This homework will be graded using unit tests by running:
`python3 -m unittest -v hw06_knn/test_knn.py`
Implement methods:

1. `calculate_similarities(self,vecTestDoc,vectorsOfTrainDocs)`: calculate similarities between test document and other train documents; do not forget to label them ([(similarity, label),...])
2. `order_nearest_to_farthest(self,distances)`: order the pairs of similarity and label from most similar to less similar
3. `labels_k_closest(self,sorted_distances)`: find k closest labels
4. `choose_one(self, labels)`. This method should return unique neighbor (label) from the given k nearest neighbors (labels). If there is a unique winner, return it, otherwise, reduce the number of k and search again.

5. `classify(self,test_file)`. This method should classify the given test document. Use the methods you have implemented before.

6. `get_accuracy(self,gold,predicted)`. This method should return the accuracy: proportion of correctly classified test documents over the whole test set of documents