CS 256: Homework2

SJSU ID: 012447851
Name: Mayuri Wadkar

Instructions to run the program from command line:

```
$ python mayuri_knn _kmeans.py
```

My python implementation requires following external libraries:
1. BeatifulSoup

System Design:

In order to determine relationship between keyword sequence and relevant job amongst all the fetched jobs, I am building a percept for corresponding keyword sequence.

This percept sequence represents a job document and is a list of averages of term frequencies over all documents. This percept has been normalized using tf*idf. And, every document in the corpus has been normalized using tf*idf.

tf = count(word, document) / len(document)

idf = log(len(collection) / count(document_containing_term, collection)

To find similarity between keyword and job documents fetched, I am using this normalized percept and tf*idf normalized TFDocuments for jobs fetched.

In order to find K nearest neighbors, I used cosine distance metric as mentioned below,

$$\text{Cosine } d(p,q) = 1 - \cos(p,q) = 1 - \frac{p_1q_1 + \cdots + p_nq_n}{\sqrt{p_1^2 + \cdots p_n^2}\sqrt{q_1^2 + \cdots q_n^2}}$$

Lookup table Implementation:

- Lookup table gets occupied with 15 jobs for popular key words when the program starts.
- Whenever a keyword which is not present in lookup table has been searched by user, it fetches at least 15 jobs and adds those to lookup table for corresponding keyword.
- Whenever user searches keywords which are already present in lookup table, jobs are fetched from table.
- Environment class stores a timestamp (date) of lookup table build. Every day, lookup table will be rebuilt only once.