CS381/780 Data Analytic Review Quiz 8

Instruction: For multiple choice questions, clearly circle one of the choice; for all other questions, write your answer right below the questions. All questions carry the same weights.

Name:

Question 1: Which of the following are true regarding clustering?

```
    1. Sum of square of the distance from the data point to its centroid increases
when K increases
    2. Elbow graph is just a heuristic for determining the value of K, i.e. there
is no absolute answer for K.
    3. Clustering is considered as supervised learning because K has to be an
external input (thus supervised by the user)
```

A. Only 1

[✓ B. Only 2]

C. Only 3

D. 1 and 2 are true

E. 2 and 3 are true

Answer:

Question #2

1. Pick a value for k (the number of clusters to create)
2. Initialize k ‘centroids’ (starting points) in your data.
3. Create your clusters. Assign each point to the nearest centroid.
4. Make your clusters better. Move each centroid to the center of its cluster
5. Repeat steps 3-4 unti your centroids converge

Question 2: Write down the pseudo code for K-means clustering algorithm

Answer:

Question 3: Say we have collected all emails sent by the (1) all presidents of the US, (2) all professors from the computer science (3) all the law department professors from Queens College, fill in the blank in the following statements with one of words among: "higher", "lower", "bigger", "smaller", or "similar".

Common Word = lower IDF, higher Term Frequency
Unusual Word = higher IDF, lower Term Frequency
Stop words = useless common words

A. The IDF of the word "database" from a corpus of all emails sent by the computer science professors should be lower than that from a corpus of all emails sent by the law professors.

B. The term frequency of the word "allegation" from emails sent by law professors should be higher than the term frequency of the same word from all emails collected from the whole corpus.

C. Average IDF of emails sent from President Trump's should have a lower IDF than that average from all the Presidents of the US.

D. The set of stops word collected from President Trump emails should be similar than that of the average from the whole corpus.

E. The set of stops word collected from President Trump emails should be **bigger** than that of the average from all the Presidents of the US.

Question 4: Write down the results from the statements.

```
  1. Results from running tozenizer of the sentence:  Good morning Americans, have
a great day!
```

Answer: **[‘Good’, ‘morning’, ‘Americans’, ‘have’, ‘a’, ‘great’, ‘day’,’!’]**

```
  2. Results from running stemming from the words:   jumping, jumped, jump
```

Answer: **jump, jump, jump**

```
  3. Results after removing stopwords of the sentence:  Cristiano Ronaldo was born
on February 5, 1985.
```

Answer: **Cristiano Ronaldo born Feburary 5, 1985**