CMSC 409: Artificial Intelligence
Project No. 4
Due: Nov. 12, 2019, noon

Student certification:
Team member 1:
Print Name: ___Aidan Kierans________________ Date: __11/14/2019_______
I have contributed by doing the following:____Everything.________________________________
Signed: _______Aidan Kierans______________ (you can sign/scan or use e-signature)

1. I have provided the feature vector in the file feature_vector.csv. I'm not able to create a large enough table in this report to show the whole thing, so I recommend opening the file using Microsoft Excel or another CSV reader.

2.
    a. As with the feature vector, I have provided the file term_document_matrix.csv, which contains the TDM and can be opened in your CSV reader of choice.
    b.
        A. Tokenize sentences: Performed to break documents down into manageable parts that can be matched against quickly by queries and rearranged as needed. If the sentences are formatted at all before this step, such as by being split into paragraphs or italicized, that information is lost.

            Words that are hyphenated are also broken down into two separate tokens, which can be necessary for queries to be matched against the word on either side of the hyphen, at the expense of whatever meaning the author of the original sentence intended to communicate when they hyphenated those words in the first place.

        B. Remove punctuation and special characters: Performed to simplify tokens so that they can be used and compared more generally and efficiently. Once punctuation/special characters such as hyphens, commas, and periods left over from tokenization are removed, the information that they communicated (e.g. which subjects are connected or separate) is lost.

        C. Remove numbers: Performed because there are infinitely many numbers and there's no good/easy way to split the range of negative infinity to positive infinity into a finite number of discrete values that are useful for a wide variety of contexts. Splitting the number line into many discrete sections would increase the dimensionality of the feature vector and TDM too much to be useful, and splitting it into only a few would result so much lost meaning that even if the discretization was a useful description in one context, it probably wouldn't be in another. Without numbers, we can only determine that a document is talking about miles per hour, rather than 60 miles per hour, for example, but that's probably all we need to know.

D. Convert upper-case to lower-case: This step is performed to reduce dimensionality and increase performance by removing information about which letters are/aren’t capitalized. This means that the first words of sentences are now indistinguishable from the last words (when viewed out of order) and acronyms/names are indistinguishable from normal words (without analyzing their frequency or usage, that is).

E. Remove stop words: Stop words are very common and generally communicate superficial and/or grammatical context, such as ownership, tense, and comparison of the non-stop words. Removing them makes it easier to identify topics by increasing the relative frequency of the more important (subject-specific) words, in addition to making the search more efficient by reducing the dimensionality of the feature space. For example, if I search for “woman with hat”, I probably want to see documents containing “woman” and “hat”, and I probably don’t care whether the woman is “with” the hat or “has” it or “had” it or “got” it, so there’s no point spending the space and computation time to record which of those options is true for a given document.

F. Perform stemming: As with stop words, the grammatical specifics of suffixes and plurality are important in context but fairly useless to a person or algorithm trying to group information by topic. Reducing words to their stems means that patterns that indicate tense and other grammatical features are removed from each token, so the words “finale”, “finalize”, “finalization”, and “finalizations”, etc. all get represented as “final”. This simplification reduces the dimensionality of the feature space by a lot, at the cost of that grammatical information.

G. Combine stemmed words: For any two stemmed words or phrases that describe the same topic, pick one and change the second to a copy of the first. This is done to reduce the number of dimensions of the feature space. It removes a lot of nuance from the documents, since words often have specific connotations, but to the extent that words are only combined when their meaning is redundant (as used in the document corpus), that loss of information is unimportant.

H. Extract most frequent words: Words that appear very infrequently aren’t useful for distinguishing one group of documents from another, because even if they correlate with certain topics/types of documents, there might not be enough data to figure that out. Because of this, only words that are somewhat frequent are worth keeping, and any that aren’t can be removed.

Note that removing these words makes the remaining words appear more significant, which can make clusters easier to distinguish, but can imply that a cluster of documents are more similar than they really are. This aspect can be reduced by removing the words that have a low frequency in a given document relative their frequency in the other documents rather than the words that have a low frequency overall, so as to specifically remove the words with low discriminatory power.

3. I chose to implement Forming Clusters As Needed (FCAN).
   a. I identified four distinct clusters in the data.

   b. The dimensionality of the TDM is driven by the number of documents and the number of terms. In general, the dimensionality can be reduced by simplifying, combining, and removing terms; by removing punctuation, special characters, and numbers, converting upper-case letters to lower-case, removing stop words, performing stemming, and removing terms that are too infrequent to be useful, the number of terms can be reduced with an acceptably small loss of information. In my case, the dimensionality was smaller than it could have been, due to my application of the methods described above, but I still could have made it smaller by combining more stemmed words and/or removing more infrequent words. I chose not to reduce the dimensionality of the TDM any further because I felt that the resulting loss of data wouldn't be worth the minor increase in processing speed. Regardless, the dimensionality of the TDM is unaffected by the order of the data, but the results of the FCAN algorithm are.

   c. I have pasted my results below. Each cluster has a handful of sentences, and all of the sentences in a given cluster are clearly related to each other. However, I was unable to cluster many sentences that I felt should have been clustered with each other, such as the sentences about the IoT and artificial intelligence in general; instead, they were all treated as individual clusters. I tried adjusting the learning rate of the clusters in addition to the activation threshold, but that caused seemingly unrelated sentences to be clustered together, so I eventually settled on a learning rate of 1.1 and a threshold of 3.5. I think the reason the sentences about artificial intelligence and the IoT were so difficult to cluster was that they had a lot of unique words that were hard to combine with other words, so they were "noisy" compared to the sentences about cars and houses. Overall, my clusters contained 17 of the 46 sentences, so I was able to cluster 37% of the data.

```
Cluster 0
ï»¿The autonomous sedan will be able to travel on any type of road at speeds of up to 60 miles per hour.
This gets the car from 0 to 60 miles per hour (that is, to 97 kilometers per hour) in 3.2 seconds.
The autonomous sedan will do a lap or 2 at around 250 kilometers per hour (149 miles per hour).
The car went round the 3 mile off road lap in 11 minutes and 50 seconds, which is an average of around 15 miles per hour.


Cluster 1
On the road test, we were able to achieve a range of 220 kilometers (around 138 miles) on a charge.
The car will go 443 kilometers (275 miles) on a charge, up 3.7 percent from before.


Cluster 2
Entire interior of home is freshly painted, large living room and bedroom.
Single family home with 5 bedroom and 2.5 bath, conveniently located near all major routes.
```

Very cute and classy house with open living area and kitchen, kitchen is updated with great appliances.

Cluster 3
Three parking spaces in back, pets are possible with approval from the owner.
Decoration is the furnishing or adorning of a space with fashionable or beautiful things.
In a bedroom, a closet is most commonly used for clothes and other small personal items that one may have.
The mirror area should definitely have at least two sources of light at least 1 feet apart to eliminate any shadows on the face.

Cluster 4
The 31.5 kilometers of roads that are off limits to the public, will be used for testing the autonomously driven car.
They have completed over 300,000 autonomous driving miles (500,000 kilometers) accident-free.
The car had to autonomously deal with a number of situations on the road.
Over 839 miles of driving, we averaged 29 miles per gallon, for the 3317 pound sedan.

Unable to cluster:
On future of machine learning, Ray Kurzweil has predicted that we are only 28 years away from the Singularity or when self-improving artificial super-intelligence will far exceed human intelligence.
Newly remodeled home for rent, 4 bedrooms with 1 bath, living room, large eat in kitchen with a full sized utility room.
Musk said that the way to escape human obsolescence may be by having some sort of merger of biological intelligence and machine intelligence.
While artificial intelligence could possibly lead to intelligence in machines with machine learning, intelligence will not necessarily lead to sentience.
IoT devices are a part of the larger concept of home automation, which can include lighting, heating and air conditioning, media and security systems.[28][29] Long term benefits could include energy savings by automatically ensuring lights and electronics are turned off.
All appliances are included, as well as security system, tenant is responsible for electric and water gas, pets negotiable based on animal.
Four bedroom 3 bath row house, home comes with 2 washers and 2 dryers and finished basement.
Artificial intelligence is combining two paradigms, that everything that we know about our reality comes by way of our senses, and that the knowledge comes from our experiences via five senses.
One of the bedroom is a large suit with a king size bed, the other one is a very nice size bedroom with a queen size.
A two- or four-door design built on a normal chassis, but with a shorter roof and interior space, club sedans were most often available in high-level U.S. models from the mid-1920s to the mid-1950s
House has been updated and renovated with an updated kitchen and new flooring.
John McCarthy, inventor of the programming language LISP, coined the term â€œartificial intelligenceâ€? in 1955.
General Artificial Intelligence involves self-aware computer programs that can engage in common-sense reasoning and learning, attain knowledge in multiple domains, feel, express and understand emotions.
Ray Kurzweil, author of the 1999 book The Age of Spiritual Machines described that intelligence would spread throughout the cosmos.

Recent artificial intelligence work has been fundamental, with techniques like deep learning laying the groundwork for computers that can automatically through machine learning increase their understanding of the world around them.Two bedroom 1 bathroom townhouse, central heat and air, water trash sewage included, living room, eat in kitchen.

Vehicles using alternative fuels such as ethanol flexible-fuel vehicles and natural gas vehicles are also gaining popularity in some countries and Electric cars, which were invented early in the history of the car, began to become commercially available in 2008.

While Hollywood movies and science fiction novels depict AI as human-like robots that take over the world, the current evolution of AI technologies isnâ€™t that scary â€“ or quite that smart.

Interior design is the art and science of understanding people's behavior to create functional spaces within a building.

An individualâ€™s bedroom is a reflection of their personality, as well as social class and socioeconomic status, and is unique to each person.

Interior designers must be highly skilled in order to create interior environments that are functional, safe, and adhere to building codes, regulations and ADA requirements.

For example, machines that calculate basic functions or recognize text through methods such as optimal character recognition are no longer said to have artificial intelligence, since this function is now taken for granted as an inherent computer function.

In larger bedrooms, a small desk and chair or an upholstered chair and a chest of drawers may also be used. In Western countries, some large bedrooms, called master bedrooms, may also contain a bathroom.

Modern parking lots use a variety of technologies to help motorists find unoccupied parking spaces using parking guidance and information system, retrieve their vehicles, and improve their experience.

In the Internet of things, if things are able to take actions on their own initiative, this human-centric mediation role is eliminated.

One of the primary functions of the family involves providing a framework for the production and reproduction of persons biologically and socially.

Passenger vehicles are a major pollution contributor, producing significant amounts of nitrogen oxides, carbon monoxide, and other pollution.

Self-driving cars have been the subject of controversy, as their machines tend to be designed for the lowest possible risk and the least casualties.

The IoT can realize the seamless integration of various manufacturing devices equipped with sensing, identification, processing, communication, actuation, and networking capabilities.

The IoT's major significant trend in recent years is the explosive growth of devices connected and controlled by the Internet/