COMP 330 Assignment #4

1 Description

In this assignment, you will be implementing a kNN classifier to classify text documents. The implementation will be in Python, on top of Spark.

You will be asked to perform three subtasks: (1) data preparation, (2) classification, and (3) evaluation.

2 Data

You will be dealing with the widely-used "20 newsgroups" data set. This is the data set that you used in Lab 5. A newsgroup post is like an old-school blog post, and this data set has 19,997 such posts from 20 different categories, according to where the blog post was made. The 20 categories are listed in the file `categories.txt`. The format of the text data set is exactly the same as the text format used for the documents in Lab 5; the category name can be extracted from the name of the document. For example, the document with identifier `20_newsgroups/comp.graphics/37261` is from the `comp.graphics` category. The document with the identifier `20_newsgroups/sci.med/59082` is from the `sci.med` category. The data file has one line per document—39 MB of text. It can be accessed at:

```
https://s3.amazonaws.com/chrisjermainebucket/comp330_A6/20_news_same_line.txt
```

or as direct S3 address, so you can use it in a Spark job:

```
s3://chrisjermainebucket/comp330_A5/20_news_different_lines.txt
```

3 The Tasks

There are three separate tasks that you need to complete to finish the assignment.

3.1 Task 1

First, you need to write Spark code that builds a dictionary that includes the 20,000 most frequent words in the training corpus—this was part of Lab 5. The words in your dictionary should be ordered based upon the relative frequency of the word. For example, the 0th word should be the most frequent word, and the $19,999$th word should be the least frequent word in the dictionary. In the case of ties, order by the values of the words themselves, from lowest to greatest.

Then, you need to use this dictionary to create an RDD where each document is represented as one entry in the RDD. Specifically, the key of the document is the document identifier
(like `20_newsgroups/comp.graphics/37261`) and the value is a NumPy array with 20,000 entries where the ith entry in the array is the number of times that the ith word in the dictionary appears in the document.

Once you do this, print out the arrays that you have created for documents:

`20_newsgroups/comp.graphics/37261`,
`20_newsgroups/talk.politics.mideast/75944`, and
`20_newsgroups/sci.med/58763`.

Since each array is going to be huge, with a lot of zeros, the thing that you want to print out is just the non-zero entries in the array (that is, for an array `a`, print out `a[a.nonzero ()]`.

3.2 Task 2

It is often difficult to classify documents accurately using raw count vectors. Thus, the next task is to write some more Spark code that converts each of those 19,997 count vectors to TF-IDF vectors ("term frequency/inverse document frequency vectors"), as described in the "Intro to Supervised Learning" lecture. The ith entry in a TF-IDF vector for document d is computed as:

$$TF(i,d) \times IDF(i)$$

Where $TF(i,d)$ is:

$$\frac{\text{Number of occurences of word } i \text{ in } d}{\text{Total number of words in } d}$$

Note that the "Total number of words" is not the number of distinct words. The "total number of words" in "Today is a great day today" is six. And the $IDF(i)$ is:

$$\log \frac{\text{Size of corpus (number of docs)}}{\text{Number of documents having word } i}$$

Again, once you do this, print out the arrays that you have created for documents:

`20_newsgroups/comp.graphics/37261`,
`20_newsgroups/talk.politics.mideast/75944` and
`20_newsgroups/sci.med/58763`.

Again, print out just the non-zero entries.

3.3 Text 3

Next, your task is to build a kNN classifier, embodied by the Python function `predictLabel`. This function will take as input a text string and a number k, and then output the name of one of the 20 newsgroups. This name is the newsgroup that the classifier thinks that the text string is "closest" to. It is computed using the classical kNN algorithm. This algorithm first converts the input string into a TF-IDF vector (using the dictionary and count information computed over the original corpus). It then finds the k documents in the corpus that are "closest" to the query vector (where distance is computed using the L_2 norm), and returns the newsgroup label that is most frequent in those top k. Ties go to the label with the closest corpus document.

Once you have written your function, run it on the following (each is an except from a Wikipedia article, chosen to match one of the 20 newsgroups). Note that I've included these tests on Canvas, for easier copy/paste):

```
predictLabel (10, 'Graphics are pictures and movies created using computers  usually referring to image
data created by a computer specifically with help from specialized graphical hardware and software.
It is a vast and recent area in computer science.  The phrase was coined by computer graphics researchers
Verne Hudson and William Fetter of Boeing in 1960.  It is often abbreviated as CG, though sometimes
erroneously referred to as CGI. Important topics in computer graphics include user interface design,
sprite graphics, vector graphics, 3D modeling, shaders, GPU design, implicit surface visualization
with ray tracing, and computer vision, among others.  The overall methodology depends heavily on the
underlying sciences of geometry, optics, and physics.  Computer graphics is responsible for displaying
art and image data effectively and meaningfully to the user, and processing image data received from
the physical world.  The interaction and understanding of computers and interpretation of data has
been made easier because of computer graphics.  Computer graphic development has had a significant
```

impact on many types of media and has revolutionized animation, movies, advertising, video games, and graphic design generally.')

predictLabel (10, 'A deity is a concept conceived in diverse ways in various cultures, typically as a natural or supernatural being considered divine or sacred. Monotheistic religions accept only one Deity (predominantly referred to as God), polytheistic religions accept and worship multiple deities, henotheistic religions accept one supreme deity without denying other deities considering them as equivalent aspects of the same divine principle, while several non-theistic religions deny any supreme eternal creator deity but accept a pantheon of deities which live, die and are reborn just like any other being. A male deity is a god, while a female deity is a goddess. The Oxford reference defines deity as a god or goddess (in a polytheistic religion), or anything revered as divine. C. Scott Littleton defines a deity as a being with powers greater than those of ordinary humans, but who interacts with humans, positively or negatively, in ways that carry humans to new levels of consciousness beyond the grounded preoccupations of ordinary life.')

predictLabel (10, 'Egypt, officially the Arab Republic of Egypt, is a transcontinental country spanning the northeast corner of Africa and southwest corner of Asia by a land bridge formed by the Sinai Peninsula. Egypt is a Mediterranean country bordered by the Gaza Strip and Israel to the northeast, the Gulf of Aqaba to the east, the Red Sea to the east and south, Sudan to the south, and Libya to the west. Across the Gulf of Aqaba lies Jordan, and across from the Sinai Peninsula lies Saudi Arabia, although Jordan and Saudi Arabia do not share a land border with Egypt. It is the worlds only contiguous Eurafrasian nation. Egypt has among the longest histories of any modern country, emerging as one of the worlds first nation states in the tenth millennium BC. Considered a cradle of civilisation, Ancient Egypt experienced some of the earliest developments of writing, agriculture, urbanisation, organised religion and central government. Iconic monuments such as the Giza Necropolis and its Great Sphinx, as well the ruins of Memphis, Thebes, Karnak, and the Valley of the Kings, reflect this legacy and remain a significant focus of archaeological study and popular interest worldwide. Egypts rich cultural heritage is an integral part of its national identity, which has endured, and at times assimilated, various foreign influences, including Greek, Persian, Roman, Arab, Ottoman, and European. One of the earliest centers of Christianity, Egypt was Islamised in the seventh century and remains a predominantly Muslim country, albeit with a significant Christian minority.')

predictLabel (10, 'The term atheism originated from the Greek atheos, meaning without god(s), used as a pejorative term applied to those thought to reject the gods worshiped by the larger society. With the spread of freethought, skeptical inquiry, and subsequent increase in criticism of religion, application of the term narrowed in scope. The first individuals to identify themselves using the word atheist lived in the 18th century during the Age of Enlightenment. The French Revolution, noted for its unprecedented atheism, witnessed the first major political movement in history to advocate for the supremacy of human reason. Arguments for atheism range from the philosophical to social and historical approaches. Rationales for not believing in deities include arguments that there is a lack of empirical evidence; the problem of evil; the argument from inconsistent revelations; the rejection of concepts that cannot be falsified; and the argument from nonbelief. Although some atheists have adopted secular philosophies (eg. humanism and skepticism), there is no one ideology or set of behaviors to which all atheists adhere.')

predictLabel (10, ' President Dwight D. Eisenhower established NASA in 1958 with a distinctly civilian (rather than military) orientation encouraging peaceful applications in space science. The National Aeronautics and Space Act was passed on July 29, 1958, disestablishing NASAs predecessor, the National Advisory Committee for Aeronautics (NACA). The new agency became operational on October 1, 1958. Since that time, most US space exploration efforts have been led by NASA, including the Apollo moon-landing missions, the Skylab space station, and later the Space Shuttle. Currently, NASA is supporting the International Space Station and is overseeing the development of the Orion Multi-Purpose Crew Vehicle, the Space Launch System and Commercial Crew vehicles. The agency is also responsible for the Launch Services Program (LSP) which provides oversight of launch operations and countdown management for unmanned NASA launches.')

predictLabel (10, ' The transistor is the fundamental building block of modern electronic devices, and is ubiquitous in modern electronic systems. First conceived by Julius Lilienfeld in 1926 and practically implemented in 1947 by American physicists John Bardeen, Walter Brattain, and William Shockley, the transistor revolutionized the field of electronics, and paved the way for smaller and cheaper radios, calculators, and computers, among other things. The transistor is on the list of IEEE milestones in electronics, and Bardeen, Brattain, and Shockley shared the 1956 Nobel Prize in Physics for their achievement. ')

predictLabel (10, ' The Colt Single Action Army which is also known as the Single Action Army, SAA, Model P, Peacemaker, M1873, and Colt .45 is a single-action revolver with a revolving cylinder holding six metallic cartridges. It was designed for the U.S. government service revolver trials of 1872 by Colts Patent Firearms Manufacturing Company todays Colts Manufacturing Company and was adopted as

```
the standard military service revolver until 1892.  The Colt SAA has been offered in over 30 different
calibers and various barrel lengths.  Its overall appearance has remained consistent since 1873.  Colt
has discontinued its production twice, but brought it back due to popular demand.  The revolver was
popular with ranchers, lawmen, and outlaws alike, but as of the early 21st century, models are mostly
bought by collectors and re-enactors.  Its design has influenced the production of numerous other models
from other companies.  ')
```

```
predictLabel (10, ' Howe was recruited by the Red Wings and made his NHL debut in 1946.  He led the
league in scoring each year from 1950 to 1954, then again in 1957 and 1963.  He ranked among the top
ten in league scoring for 21 consecutive years and set a league record for points in a season (95)
in 1953.  He won the Stanley Cup with the Red Wings four times, won six Hart Trophies as the leagues
most valuable player, and won six Art Ross Trophies as the leading scorer.  Howe retired in 1971 and
was inducted into the Hockey Hall of Fame the next year.  However, he came back two years later to
join his sons Mark and Marty on the Houston Aeros of the WHA. Although in his mid-40s, he scored over
100 points twice in six years.  He made a brief return to the NHL in 197980, playing one season with
the Hartford Whalers, then retired at the age of 52.  His involvement with the WHA was central to their
brief pre-NHL merger success and forced the NHL to expand their recruitment to European talent and
to expand to new markets.  ')
```

4 Turnin

Please create a document (.txt or .pdf) with your results from all three tasks, copied and pasted from the screen. Then zip up all of your code (as a `.py` file) and your document (use .gz or .zip only, please!), or else attach each piece of code as well as your document to your submission individually.

5 Grading

Each task is worth one third of the overall grade.