COMP 543 Assignment #4

1 Description

In this assignment, you will be implementing a kNN classifier to classify text documents. "Classification" is the task of labeling documents based upon their contents. The implementation will be in Python, on top of Spark, **USING RDDs**. You will be asked to perform three subtasks, covering data preparation and classification.

2 Data

You will be dealing with MEDLINE/PubMed data:
(`https://www.nlm.nih.gov/bsd/pmresources.html`).

MEDLINE is a bibliographic database of journal citations, abstracts, and metadata. The data set is composed of categories, identifier, and abstract triples. This data set has 26,754 such posts from 11 different categories, according to catagories assigned to the articles. The 11 categories are listed in the file `categories.txt`. The category name can be extracted from the name of the document. For example, the document with identifier `doc id=Wounds/24458063` is from the `Wounds` category. The document with the identifier `doc id=MedicalEducation/20662575` is from the `MedicalEducation` category. The data file has one line per document for a total of ˜38 MB of text. It can be accessed via Spark at: `s3://risamyersbucket/A4/pubmed.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. When you do this, please start with the code provided with THIS assignment in `A4start.py` so that we know that everyone has the same dictionary. 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 `BrainInjuries/17626737`) 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
`Wounds/23778438`,
`ParasiticDisease/2617030`, and
`RxInteractions/1966748`

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 26,754 count vectors to TF-IDF vectors "term frequency-inverse document frequency vectors"). 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:

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

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:

`PalliativeCare/16552238`,
`SquamousCellCancer/23991972` and
`HeartFailure/25940075`

Again, print out just the non-zero entries.

3.3 Task 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 11 categories. This name is the new group 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 category 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 excerpt from a pubmed abstract, chosen to match one of the 11 categories). These function call are provided in the file `predictLabelCalls.py`.

```
predictLabel (10, 'Simulation technology for health care professional skills training
and assessment.  Changes in medical practice that limit instruction time and patient availability,
the expanding options for diagnosis and management, and advances in technology are contributing
to greater use of simulation technology in medical education.  Four areas of high-technology
simulations currently being used are laparoscopic techniques, which provide surgeons with
an opportunity to enhance their motor skills without risk to patients; a cardiovascular
disease simulator, which can be used to simulate cardiac conditions; multimedia computer
systems, which includes patient-centered, case-based programs that constitute a generalist
curriculum in cardiology; and anesthesia simulators, which have controlled responses that
vary according to numerous possible scenarios.  Some benefits of simulation technology include
improvements in certain surgical technical skills, in cardiovascular examination skills,
and in acquisition and retention of knowledge compared with traditional lectures.  These
systems help to address the problem of poor skills training and proficiency and may provide
a method for physicians to become self-directed lifelong learners.')
```

predictLabel (10, 'Propofol inhibits T-helper cell type-2 differentiation by inducing apoptosis via activating gamma-aminobutyric acid receptor. Propofol has been shown to attenuate airway hyperresponsiveness in asthma patients. Our previous study showed that it may alleviate lung inflammation in a mouse model of asthma. Given the critical role of T-helper cell type-2 (Th2) differentiation in asthma pathology and the immunomodulatory role of the gamma-aminobuty acid type A (GABAFor invivo testing, chicken ovalbumin-sensitized and challenged asthmatic mice were used to determine the effect of propofol on Th2-type asthma inflammation. For invitro testing, Th2-type cytokines as well as the cell proliferation and apoptosis were measured to assess the effects of propofol on Th2 cell differentiation and determine the underlying mechanisms.We found that propofol significantly decreased inflammatory cell counts and interleukin-4 and inflammation score invivo. Propofol, but not intralipid, significantly reduced the Th2-type cytokine interleukin-5 secretion and caused Th2 cell apoptosis without obvious inhibition of proliferation invitro. A GABA receptor agonist simulated the effect of propofol, whereas pretreatment with an antagonist reversed this effect.This study demonstrates that the antiinflammatory effects of propofol on Th2-type asthma inflammation in mice are mediated by inducing apoptosis without compromising proliferation during Th2 cell differentiation via activation of the GABA receptor.Copyright 2016 Elsevier Inc. All rights reserved.')

predictLabel (10, 'Evaluation of isopathic treatment of Salmonella enteritidis in poultry. Salmonellosis is a common problem worldwide in commercially reared poultry. It is associated with human Salmonellosis. No fully satisfactory method of control is available.Nosodes to an antibiotic-resistant strain of Salmonella enterica serovar Enteritidis in D30 (30X) potency were prepared. One day old chicks (N = 180) were divided into four groups: two control and two different preparations of the nosode. Treatments were administered in drinking water for 10 days. The birds were challenged by a broth culture of the same Salmonella, by mouth, on day 17. Cloacal swabs were taken twice weekly for Salmonella enterica serovar Enteritidis.Birds receiving active treatment were less likely to grow the strain of Salmonella from cloacal swabs compared to control.Isopathy is low cost and non-toxic. It may have a role to play in the widespread problem of Salmonella in poultry. Further research should be conducted.')

predictLabel (10, 'Management of the neck after chemoradiotherapy for head and neck cancers in Asia: consensus statement from the Asian Oncology Summit 2009. The addition of a planned neck dissection after radiotherapy has traditionally been considered standard of care for patients with positive neck-nodal disease. With the acceptance of chemoradiotherapy as the new primary treatment for patients with locally advanced squamous-cell head and neck cancers, and the increasing numbers of patients who achieve a complete response, the role of planned neck dissection is now being questioned. The accuracy and availability of a physical examination or of different imaging modalities to identify true complete responses adds controversy to this issue. This consensus statement will address some of the controversies surrounding the role of neck dissection following chemoradiotherapy for squamous-cell carcinomas of the head and neck, with particular reference to patients in Asia.')

predictLabel (10, '[Quality management in the hospital with special reference to the medical departments]. Present German health legislation requires standardisation of systems and procedures not only for medical departments but for entire clinics as well. Changes in organisation and treatment possibilities should enable clinic administrations to better manage the quality of the services provided. This system of quality control and quality management is based on the introduction of ISO 9004, part 2. Medical and non-medical procedures are standardised and streamlined through flow-diagrams, the details of which are summarised in a series of quality handbooks with an emphasis on guidelines and standards. These handbooks are distributed clinic-wide, and although acceptance is not yet determined, objections are probable. Future analysis will show the effects due to quality management in the clinical setting.')

predictLabel (10, 'Suicide attempts involving power drills. A 61-year-old man was found

```
dead next to a power drill soiled with blood and bone dust.  A 5mm circular wound of the
forehead corresponded to the size of the drill bit.  Subarachnoid haemorrhage was present
over the anterior pole of the left frontal lobe with a penetrating injury extending 75mm
into the frontal lobe white matter towards, but not involving, the basal ganglia.  No major
intracranial vessels had been injured and there was no significant intraparenchymal haemorrhage.
Death was due to haemorrhage from self-inflicted stab wounds to the abdomen with an associated
penetrating intracranial wound from a power drill.  Deaths due to power drills are rare
and are either accidents or suicides.  Wounds caused by power drills may be mistaken for
bullet entrance wounds, and the marks around a wound from the drill chuck as muzzle imprints.
A lack of internal bevelling helps to distinguish the entrance wound from that due to a
projectile.  Significant penetration of the brain may occur without lethal injury.  Copyright
2013 Elsevier Ltd and Faculty of Forensic and Legal Medicine.  All rights reserved.')

   predictLabel (10, 'Neurobehavioral recovery.  This review discusses recent programs in
early and late neurobehavioral recovery from closed head injury (CHI). The research on early
recovery has encompassed the relationship of localized brain lesions to the duration of
impaired consciousness and features of posttraumatic amnesia.  Of the research on late neurobehaviora
outcome of CHI, studies emanating from the Traumatic Coma Data Bank are reviewed in detail,
including analysis of acute neurologic indices in relation to recovery of memory, information
processing speed, and other cognitive measures.  Recent studies concerning the neurobehavioral
outcome of CHI in children are discussed as are investigations of behavioral disturbance,
psychosocial outcome, and family variables.  The review concludes with an assessment of
recent studies concerning the efficacy of rehabilitation directed toward the cognitive sequelae
of CHI and preliminary trials to evaluate the potential use of psychoactive drugs in the
postacute management of head injured patients.')

   predictLabel (10, 'Ethical issues arising from the requirement to sign a consent form
in palliative care.  French healthcare networks aim to help healthcare workers to take care
of patients by improving cooperation, coordination and the continuity of care.  When applied
to palliative care in the home, they facilitate overall care including medical, social and
psychological aspects.  French legislation in 2002 required that an information document
explaining the functioning of the network should be given to patients when they enter a
healthcare network.  The law requires that this document be signed.  Ethical issues arise
from this legislation with regard to the validity of the signature of dying patients.  Signature
of the consent form by a guardian or trustee, a designated person--the Person of Trust--transforms
the doctor-patient relationship into a triangular doctor-patient-third-party relationship.')
```

4 Turnin

Submit two documents: one with your results (.txt) and one with your code (.py). Create a single document that has results for all three tasks. Turn in this document as well as all of your code. Please zip up all of your code and your text document (use .gz or .zip only, please!), or else attach each piece of code as well as your text results document to your submission individually. No PDFs of code, please!

5 Grading

Each task is worth 33% of the overall grade.