**Drexel University**

**College of Computing and Informatics**

**INFO 371 – Data Mining Applications**

**Assignment 2**

**Due Date: Sunday Oct. 20, 2019**

**A. Requirements**

**TEAM (up to two or three members) assignment**: Please work with another student on this assignment. Please inform the instructor of your team membership before working on the assignment.

# Data Collection (20 point)

**-----------------------------------**

Prepare a small data set of paper abstracts (or short news articles/reports): identify **three to five different domains/areas** of your interest, find at least **20 papers** (one-paragraph or so abstracts only) in each domain, and create an ARFF file named **abstracts.arff** with two attributes:

* String attribute: **@attribute abstract string**
* Nominal attribute: **@attribute domain {…}** (replace … with the areas you identified)
* **At least 60 instances** in the data.

# Task 1: Data Preprocessing, Transformation, and Analysis (40 points)

**----------------------------------------------------------------------------------------------**

**Task 1.A** String Vectorized to Binary Representation

1. Open **abstracts.arff** in Weka Explorer and choose **unsupervised-> attribute-> StringToWordVector** for Filter.
2. Click the text box containing “StringToWordVector –R …” to bring up the property editor to change one option below:
   1. Select **true** for **lowerCaseTokens** and click **OK**
3. Click **Apply** to run the vectorization filter and you will see a number of new word attributes after it is done.
4. Click **Save…** to save the result (binary vector representation) as **abstracts-binary.arff**.
5. Click **Edit…** to view the processed data and manually sample to **create a data table**:
   1. Pick 9 attributes: 3 stopwords, 3 rare words (unique in each area/domain), three other words/attributes;
   2. Pick three instances/papers, each from one domain/area;
   3. Create a data table (**example below**) with data from Weka.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Feature/attribute weight | | | | | | | | |
| Paper# (domain) | the | and | to | science | nature | machine | info | data | math |
| 2 (Stats) | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 10 (ML) | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| 17 (DM) | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |

**Task 1.B** String Vectorized to TF Representation

* Reopen **abstracts.arff** file in weka and follow all steps in Task 1.A, except for step 2:
  + Select **true** for **lowerCaseTokens;**
  + Select **true** for **outputWordCounts** (this will produce Term Frequency/counts rather than binary 0/1 values).
* Save the processed file as **abstracts-tf.arff**
* Create a similar data table (of the same attributes/words and papers) with TF representation.

**Task 1.C** String Vectorized to TF\*IDF Representation

* Reopen **abstracts.arff** file in weka and follow all steps in Task 1.A, except for step 2:
  + Select **true** for **lowerCaseTokens;**
  + Select **true** for **outputWordCounts**
  + Select **true** for **IDFTransform** (this with outputWordCounts generates TF\*IDF weights)
* Save the processed file as **abstracts-tfidf.arff**.
* Create the same data table with TF\*IDF representation.

**Task 1.D** Comparison and Discussion

By now, you should have three data tables with binary, TF, and TF\*IDF representations. Compare them and discuss how weights change from one representation to the other. Use examples (e.g. specific papers and word tokens) to explain the difference and rationale.

# Task 2: Data Clustering (40 point)

**----------------------------------------------**

**Task 2.A. Clustering with the three representations**

Load each of the three processed arff files, namely abstracts-binary.arff, abstracts-tf.arff, and abstracts-tfidf.arff in Weka and select the Cluster tab to perform the following clustering:

* Choose **SimpleKMeans** for Clusterer;
* Click the text box with “SimpleKMeans –N …” and set **numClusters** to **the number of domains/areas** you have in the data (3 to 5).
* Select **(Nom) domain** attribute for **Classes to clusters evaluation** and Start.
* Take note of **percentage** of **incorrectly clustered instances** at the end of the output.
  + A **lower number indicates better** clustering performance

Create the following table summarizing clustering results based the three different representations:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clustering effectiveness with different vector representations | | |
|  | Binary | TF | TFIDF |
| Incorrectly clustered (%): |  |  |  |

# Task 2.B. Clustering with different numbers of words/attributes

Now **repeat task 1.C** for TF\*IDF transformation but with the following **parameters changed**:

* Choose **Rainbow** as the Stopword Handler (for stopword removal);
* Change the **wordsToKeep** (per class) to 10, 40, 160, 640;
* For each wordsToKeep setting, vectorize the data, and perform the same kmeans clustering as in **Task 2.A**;
* Please save each **vectorized data** as abstracts-tfidf10.arff, abstracts-tfidf40.arff, etc.
* Report the results using the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Clustering effectiveness with different number of attributes | | | |
| Words to Keep | 10 | 40 | 160 | 640 |
| # Attributes generated |  |  |  |  |
| Incorrectly clustered (%): |  |  |  |  |

# Task 2.C. Compare and discuss the difference with reason/insight

**Compare results** with binary, TF, TF\*IDF, and with the different number of attributes (words to keep). Reason and explain **what** may have contributed to the results (in terms of the number of incorrectly clustered instances) and **why**.

**B. What to Hand In**

1. The **ARFF data files**, including 1) abstracts.arff, 2) abstracts-binary.arff, 3) abstracts-tf.arff, 4) abstracts-tfidf.arff, 5) abstracts-tfidf10.arff, 6) abstracts-tfidf40.arff, 7) abstracts-tfidf160.arff, and 8) abstracts-tfidf640.arff
2. A **well-structured** MS Word or PDF file with **the following content in detail**:
   1. Step-by-step tasks and requested results.
   2. Explain the processes and results so that a reader outside of this class still can understand what you are doing here.
   3. Requested comparisons and discussions.
   4. Any conclusions must be backed up by the evidence drawn from your data and analyses.

**C. How to Hand In**

1. Please name your report file as **INFO371-assign2-yourFirstName-yourLastName.docx**.
2. Submit your report file through the course website in the **Blackboard Learn** system.

**D. When to Hand In**

1. Submit your assignment no later than **11:59pm** in the due date.
2. There will be a 10% (absolute value) deduction for each day of lateness, to a maximum of 3 days; assignments will not be accepted beyond that point. Missing work will earn a zero grade.

**E. Written Presentation Requirements**

Images must be clear and legible. Assignments will be judged on the basis of visual appearance, grammatical correctness, and quality of writing, as well as their contents. Please make sure that the text of your assignments is well-structured, using paragraphs, full sentences, and other features of well-written presentation. Text font size should be at least 11 points.