IST736 Text Mining - HW2 – Vectorization

Overall Summary: Merge HW #2 and #3 into one report and submission. Your submission should include.

1. Reading data from both: csv file and corpus. (Any data)
2. Tokenize, stem, lemmatize, stopwords
3. Vectorize
4. Build word cloud or perform k-means clustering or similar …
5. Write report.

Explanation and Clarification

1) **The Assignment FORMAT is required for all assignments and for the project. Starting now, I will begin to grade much more strictly per this required format and the elements in each section.**
2) **A grade of 92 – 95 means that you did everything right, good, complete, and professionally. To get above 95, you should think about how to make it excellent. Grades between 85 – 91 means that you are right on track and can think about ways to improve specific section or elements, and/or to add visualizations, etc.**
3) **One of the goals of this class is to think about where you are now and where you want to be – and then how to get there.**

For this Assignment, here are some criteria to integrate into the format and requirements.

You can always do more – but you cannot do less ☺

1) You will vectorize the data that you collected in HW1.
    a. **Here – it might be the case that the data you collected in HW 1 is not ideal and/or is too large etc. It is OK to collect different data for this assignment.**
    b. **You can choose to vectorize from a .csv or from a corpus. You will need to know how to do both – so either is fine for now and both is required for later.**
2) Because the goal is to identify the **public sentiment toward AI on social media**, you need to think about vectorization options, regarding both what to count and how to count
    a. **The "thinking about it part" is very important. Always think about what you are really doing, your goals, and how to reach them. This will also make it a lot easier to write up the report per the format. Many people forget to think when coding. I know this sounds funny – but its true.**
    b. **You should have at least two (or more) vectorization options – such as frequency count, normalized count (with some kind of z score or min-max, and tf-idf. You should compare the results.**
    c. **Always use visualization and tables when explaining.**

3) Make sure to explain the decisions you made during the vectorization process e.g.,if you removed stopwords and why.
    a. **This will actually involve many steps and is part of the Analysis section and the Data Cleaning and Preparation subsection under Analysis. Vectorization is actually part of data prep - it is the formatting part.**
    b. **You will think about stopwords, stemming (if you wish), things that are not words, words with numbers, etc. Clean the best that you can. 75% of the time is spent on cleaning and formatting.**
4) Write a report to include the following information:
    a. **This report must be written per the required format and must also contain all elements required by the format – such as a proper and appropriate Introduction (that is about the topic not the methods), Analysis (which contains subsections for data cleaning and prep, EDA and visual EDA, and subsections for each model or method, Results, where you will also compare the different normalizations you did (such as frequency count vs. min-max for example), and appropriate Conclusions which are 100% non-technical and discuss the topic with respect to the findings.**
    b. In addition: Include the following in appropriate locations:
        i. How you collected the data.
        ii. Describe your vectorization choices and corresponding result. For example, if you chose to do stemming, how did the vocabulary size change after stemming? Did the stemming eliminate important linguistic information that you'd rather keep, or not?
        iii. Conclude with the best vectorization option(s**). [This actually goes in data cleaning and in results – NOT in the conclusions]**
        iv. Your report should provide sufficient information for others to replicate what you did.
        ~~v. Submit your report with your original data file and the vectors from your best vectorization options.Follow the HW1 requirement on formatting and grading rubrics.~~

Submission for this class is always the same:

1) Submit the .docx as your assignment paper.
2) Create a zip and place into it your .py and your dataset(s).

NOTE: The key to this assignment is learning to get messy data into a vector (matrix and dataframe) format. You should be able to do this from a corpus and also from csv and text. I recommend practicing all three with data THAT YOU CREATE so that you know what it should be. It is NOT required that you use your data from HW 1. However, once your code is done, it should run on the data from HW 1 – that is the goal – reusability!

IST736 Text MiningHW3

Exploration of an Interesting Text Corpus

1) In this assignment, you have the freedom to find an existing text corpus, or create a new text corpus of your interest.
    a. I recommend both. Start with a nice small corpus (folder that is new) that you fill with files (.txt) the *you* create. Be sure you can read this in, vectorize it, and maybe try a method – like kmeans.
    b. Next, see if you can find some files. There is Movies sentiment data out there. This is harder, because you will need to update your vectorized data frame (and matrix – depending on the format) to have a column for sentiment label. However, we will need this throughout this class and so it's a good idea to do it now *before* you need to not only do it, but also then apply and learn about new models and methods.
    c. I also recommend that you find (or create) a .csv file. Each row is a document. Columns are…well – they can be many things. Here is one example:

Doc	Label	Review
Hike1	Love	Hiking is a great hobby. I love to hike. One can hike in mountains. Mountain hiking also includes wild flower hikes which are beautiful.
Hike2	Hate	Proper gear is required for safe and fun hiking. To hike, plan on a location and a trail. Hiking off a trail can be dangerous. When you hike, have spare socks, lots of water, and other gear.
Dog1	Neutral	There are many pet options. Getting a dog is one pet option. While cats can be fun, dogs are better for sports and exercise.
Dog2	Love	If you get a dog, consider training your dog. Also remember that other humans exist. So, do not let your dog bark. Train your dog to listen to your commands and to have fun with you. You dog will appreciate good training.
etc…	etc…	etc…

IMPORTANT: The key before finishing Week 3 is that you can read in data from a corpus or from a csv file, and can also manage labels. Another goal is that you can vectorize this data (as a dataframe and as a matrix). Finally, it is important to be able to normalize as needed (and understand what you are doing – no "black boxes").

Use Python.

If you want to try something *in addition to* but not instead of Python, once you get it to work in Python, try it in R.

You will need a lot of data cleaning:

Stopwords (I suggest you both do this by hand – like the code I shared) and you also use CountVectorizer's version. Google and read about all the things that CountVectorizer can do.

Merging words (stemming)

Make everything lowercase…

Etc..Think about that you are doing and why. Explain the decisions you make.

Then, look for results or interesting information

The lectures showed some examples of comparative analysis and trend analysis. But you have the freedom to define what would be interesting patterns as long as you can explain it in a sensible way. Always use the required FORMAT for all assignments and the project. ~~Follow the HW1 requirement on formatting and grading rubric~~