**TA Group Assignments 1**

Group Assignment 1 has two tasks, as follows.

**Task 1: Tracking the evolution of Terminology in the US Tech Sector**

**Background:** All US public firms are required to file a 10-K compliance report[[1]](#footnote-1) annually with the Securities and Exchange Commission (SEC) which makes this data publicly available. In Item 1 of the 10-K, firms list and describe what their main lines of business are, their key products and markets are etc. Business descriptions for 30 well-known technology sector firms, from year 2005 to 2014, are uploaded on LMS in *.Rds* format. The file names of the firms have the following pattern -

*CIK\_ File date\_10-K\_Firm Name\_.txt*

*for example - 21665\_23FEB2012\_10-K\_COLGATE PALMOLIVE CO.txt*

**Task:** Now do the following.

Step 1: Clean the text data and create a Document Term Matrix (DTM) for each year.

**Tip** - For best results, create your DTM after encoding frequent bi-grams as unigrams, lemmatizing tokens to take away token form heterogeneity, use of chunks such as noun or verb phrases etc.

Step 2: Using these DTMs find **new terms** for the sample from 2006 onwards. New\_terms are text tokens (unigrams, ngrams, noun or verb phrases etc.) that appeared for the first time in a given year and didn't appear in prior years.

For example, make year 2005 as your base year. So in year 2005, there won't be any new\_terms. In 2006, new terms will be those terms which appeared in 2006 for the first time and weren't mentioned in year 2005. Similarly in 2007, new terms will be those terms which appeared in 2007 for the first time and weren't mentioned in year 2005 and 2006. And so on.

Step 3: Display new terms appropriately using any of the display aids we covered in class (or something from outside as well).

**Tip** For steps 2 and 3, bonus marks if you can functionize parts of your workflow, upload a list of functions on a script file onto Github for sourcing as a piped workflow into the local machine etc. Pay attention to the details of function writing such as default values for some arguments, exception handling etc.

Step 4: Explain your findings. Are you able to identify different technologies, platforms, products etc. which evolved over time?

**Deliverable**: Build a markdown showing your workflow, results, visualizations and explanation which can be uploaded to the appropriate Submission folder on LMS within the deadline. Remember to write the names and registration numbers of all students in the group also in there.

**Task 2: Topic Model Simulation**

Perform a simulation experiment along the lines of what we did in class (based on the tidytext book example). Use the following steps as a guideline.

Step 1: Choose 4 very different domains and scrape data from there (e.g., Wiki pages on Harry Potter, Football, Pink Floyd or Mahatma Gandhi etc). Ensure the corpus has adequate length in each case.

Step 2: Break each Wiki page into sentences using a sentence tokenizer.

Step 3: Now construct simulated documents consisting of a few sentences from page 1, a few from page 2 and so on. For instance, suppose there are 8 sentences in each simulated document, then you could choose the distribution of sentences in the following way:



Note above that the sentences in each document could appear in any order. Randomly choose sentences from each wiki page and insert them to create simulated documents. Create at least 50+ simulated documents in this manner.

Step 4: Now run an LTM on this simulated corpus with K=4 topics. Obtain the factor score and factor loading matrices from the topic model output.

Step 5: For each document, check the topic score for each topic. See how closely it matches the percentages we calculated in the table. For example, does ‘Doc 1’ have around 25% topic score for topic 1? And so on. One could compute the mean squared deviation for each document between the recovered topic scores and the constructed topic proportion.

Step 6: Bonus points - Draw a confusion matrix of the 4 topics versus the 4 wiki pages to see if they were correctly or wrongly classified.

**Deliverable**: Build a markdown showing your workflow, results and discussion. Put the markdowns of both task 1 and task 2 together in a zipped file and upload to the appropriate Submission folder within the deadline. Remember to write the names and registration numbers of all students in the group also in there.

1. https://en.wikipedia.org/wiki/Form\_10-K [↑](#footnote-ref-1)