**ISMG 6470 Text Data Analytics - Assignment 1.**

Announced on August 29th, 2018.

Submission due by Tuesday, September 4, 2018

To do the assignment, please “copy and paste” the below two texts (on Apple and Samsung) into a new RScript file of RStudio. All questions are made based on these two texts.

##################

# homework #1: start

##################

apple <- "Apple Inc. is an American multinational technology company headquartered in Cupertino, California that designs, develops, and sells consumer electronics, computer software, and online services. The company's hardware products include the iPhone smartphone, the iPad tablet computer, the Mac personal computer, the iPod portable media player, the Apple Watch smartwatch, the Apple TV digital media player, and the HomePod smart speaker. Apple's consumer software includes the macOS and iOS operating systems, the iTunes media player, the Safari web browser, and the iLife and iWork creativity and productivity suites. Its online services include the iTunes Store, the iOS App Store and Mac App Store, Apple Music, and iCloud.

Apple was founded by Steve Jobs, Steve Wozniak, and Ronald Wayne in April 1976 to develop and sell personal computers. It was incorporated as Apple Computer, Inc. in January 1977, and sales of its computers saw significant momentum and revenue growth for the company. Within a few years, they had hired a staff of computer designers and had a production line. Apple went public in 1980 to instant financial success. Over the next few years, Apple shipped new computers featuring innovative graphical user interfaces, and Apple's marketing commercials for its products received widespread critical acclaim. However, the high price tag of its products and limited software titles caused problems, as did power struggles between executives at the company. Jobs resigned from Apple and created his own company. As the market for personal computers increased, Apple's computers saw diminishing sales due to lower-priced products from competitors, in particular those offered with the Microsoft Windows operating system. More executive job shuffles happened at Apple until then-CEO Gil Amelio in 1997 decided to buy Jobs' company to bring him back. Jobs regained position as CEO, and began a process to rebuild Apple's status, which included opening Apple's own retail stores in 2001, making numerous acquisitions of software companies to create a portfolio of software titles, and changed some of the hardware technology used in its computers. It again saw success and returned to profitability. In January 2007, Jobs announced that Apple Computer, Inc. would be renamed Apple Inc. to reflect its shifted focus toward consumer electronics and announced the iPhone, which saw critical acclaim and significant financial success. In August 2011, Jobs resigned as CEO due to health complications, and Tim Cook became the new CEO. Two months later, Jobs died, marking the end of an era for the company.

Apple is the world's largest information technology company by revenue and the world's second-largest mobile phone manufacturer after Samsung. In February 2015, Apple became the first U.S. company to be valued at over US$700 billion. The company employs 116,000 full-time employees as of October 2016 and maintains 498 retail stores in 22 countries as of July 2017. It operates the iTunes Store, which is the world's largest music retailer. As of January 2016, more than one billion Apple products are actively in use worldwide.

Apple's worldwide annual revenue totaled $215 billion for the 2016 fiscal year. The company enjoys a high level of brand loyalty and has been repeatedly ranked as the world's most valuable brand. However, it receives significant criticism regarding the labor practices of its contractors and its environmental and business practices, including the origins of source materials."

samsung <- "Samsung Group is a South Korean multinational conglomerate headquartered in Samsung Town, Seoul. It comprises numerous affiliated businesses, most of them united under the Samsung brand, and is the largest South Korean chaebol (business conglomerate).

Samsung was founded by Lee Byung-chul in 1938 as a trading company. Over the next three decades, the group diversified into areas including food processing, textiles, insurance, securities and retail. Samsung entered the electronics industry in the late 1960s and the construction and shipbuilding industries in the mid-1970s; these areas would drive its subsequent growth. Following Lee's death in 1987, Samsung was separated into four business groups – Samsung Group, Shinsegae Group, CJ Group and Hansol Group. Since 1990, Samsung has increasingly globalised its activities and electronics; in particular, its mobile phones and semiconductors have become its most important source of income.

Notable Samsung industrial affiliates include Samsung Electronics (the world's 2nd largest information technology company measured by 2015 revenues, and 5th in market value), Samsung Heavy Industries (the world's 2nd largest shipbuilder measured by 2010 revenues), and Samsung Engineering and Samsung C&T (respectively the world's 13th and 36th largest construction companies). Other notable subsidiaries include Samsung Life Insurance (the world's 14th largest life insurance company), Samsung Everland (operator of Everland Resort, the oldest theme park in South Korea) and Cheil Worldwide (the world's 15th largest advertising agency measured by 2012 revenues).

Samsung has a powerful influence on South Korea's economic development, politics, media and culture and has been a major driving force behind the \"Miracle on the Han River\". Its affiliate companies produce around a fifth of South Korea's total exports. Samsung's revenue was equal to 17% of South Korea's $1,082 billion GDP."

##################

# homework #1: End

##################

These two texts (on Apple and Samsung) came from Wikipedia. The “apple” and “samsung” text objects describe the histories of Apple and Samsung, respectively. Using the two text objects, answer all questions below. Please submit two separate documents as an assignment: (1) MS Word file that answers the questions, and (2) R code that you used to answer the questions. The MS Word file and R answer code file should be submitted as in the below format:

1. MS Word file:
   1. At the top, include the course ID, course name, student ID number, student name, and submission date.
   2. In the body, describe your answer for each question appropriately. Per each answer, provide (i) brief description on your answer and (ii) R codes that might be used to answer the question.
2. R code file:
   1. At the top of the answer-code, make comments (using # appropriately) for your name, your student ID number, submission date.
   2. Make appropriate comments to describe what each code is to perform.

**Question 1)**

1. Convert all upper-case characters into lower-case ones for the ‘apple’ text objects. Report R code that you might use for this task.
2. Convert all upper-case characters into lower-case ones for the ‘samsung text objects. Report R code you might use for this task.
3. Report how many characters exist in the ‘apple’ text object.
4. Report how many characters exist in the ‘samsung’ text object.

(Note that, in this question, ‘characters’ include alphabets, space, special characters, symbols etc.)

**Question 2)**

1. Split the lower-cased ‘apple’ text object (in Q1) into the word unit.
2. Split the lower-cased ‘samsung’ text object (in Q1) into the word unit.
3. Report the total number of words, excluding all punctuations, used in the ‘apple’ text object.
4. Report the total number of words, excluding all punctuations, used in the ‘samsung text object.

**Hint**: Use ‘space’ (i.e., “ ”) and line break (i.e., [\\n](file://n)) to split the text objects into the word unit like split=“ |\\n”, where “ ” means a space, ‘|’ means ‘or’, and ‘\\n’ means line-break. In sum, split=“ |\\n” means ‘space or line-break’.

**Question 3)**

1. Report the number of characters of each word used in the lower-cased ‘apple’ text object.
2. Report the mean number of characters for each word used in the lower-cased ‘apple’ text object.
3. Report the number of characters of each word used in the lower-cased ‘samsung’ text object.
4. Report the mean number of characters for each word used in the lower-cased ‘samsung’ text object.

(**Hint**: If the ‘apple’ text object contains *j* number of words, you should first compute the number of characters separately for the 1st, 2nd, 3rd, 4th, ….., jth word by using *for-loop* function).

**Question 4)**

1. Report the total number of spaces being used in the ‘apple’ text object,
2. Report the total number of spaces being used in the ‘samsung’ text object.

**Question 5)**

1. Report a frequency table (using table function) for those words that start with upper cases in the ‘apple’ text object.
2. Report a frequency table (using table function) for those words that start with upper cases in the ‘samsung’ text object.

**Note**: Words that start with upper cases include (a) any word that is composed of all upper cases, (b) the upper case ‘A’ (as an article) at the beginning of a new sentence, and (c) any word that starts with an upper case at the beginning of a new sentence.

**Question 6)**

1. Extract and report all year expressions (in four digit numbers like ‘2011’ as numeric data format) from the ‘apple’ text object.
2. Extract and report all year expressions (in four digit numbers like ‘2011’ as numeric data format) from the ‘samsung’ text object.
3. Report how many times the year expressions appear in the ‘apple’ text object.
4. Report how many times the year expressions appear in the ‘samsung’ text object.