## 2. Assessment Description

Text documents, such as long recordings and meeting transcripts, are usually comprised of topically coherent text segments, each of which contains some number of text passages. Within each topically coherent segment, one would expect that the word usage demonstrates more consistent lexical distributions than that across segments. A linear partition of texts into topic segments can be used for text analysis tasks, such as passage retrieval in IR, document summarization, and discourse analysis. In this assessment, you are required to write Python code to preprocess a set of meeting transcripts and convert them into numerical representations suitable for input into topic segmentation algorithms.

***This is an individual assignment and worth 30% of your total mark for FIT5196.***

The detailed tasks are as follows:

- **Task 1: Reconstruct meeting transcripts with topical boundaries.** The original meeting transcripts are stored in three different types of XML files, which are ending with ".words.xml", ".topic.xml" and ".segments.xml". (The details about the three types of files can be found in Section 3 below). The task here is to reconstruct the original meeting transcripts with the corresponding topical and paragraph boundaries from these files. Please note that
    - A meeting transcript must be generated for each of the "*.topic.xml" file. For example, "**ES2002a**.txt" will be generated for "**ES2002a**.topic.xml".
    - All the generated meeting transcripts with the ".txt" file extension must be saved in the folder "txt_files".
    - The topical boundaries must be denoted with "**********"(i.e., 10 asterisks).
    - All the tokens, including punctuations, must be separated by a white space. For example, "Alright , okay . Okay ."
    - Besides the topical boundaries, the paragraph boundaries must also be reconstructed with the "*.segments.xml" file.
    - The input files to your notebook "task_1.ipynb" must be the three types of XML files. The output must be the meeting transcripts saved in a set of txt files.
    - A sample meeting transcript is provided in the "txt_file" folder.

- **Task 2: Generate sparse representations for the meeting transcripts.** The aim of this task is to build sparse representations for the meeting transcripts generated in task 1, which includes word tokenization, vocabulary generation, and the generation of sparse representations. Please note that
    - The word tokenization must use the following regular expression, **"\w+(?:[-']\w+)?",** and all the words must be converted into the lower case.
    - The stop words list (i.e, **stopwords_en.txt**) provided in the zip file must be used.
    - The words, whose document frequencies are greater than 132, must be removed.
    - Generating multi-word phrases (i.e., collocations) are not needed.
    - The output of this task must contain the following files:
        - **vocab.txt**: It contains the **unigram** vocabulary in the following format, **word_string:integer_index.** Words in the vocabulary must be sorted in alphabetic order. For example, "absolute:22" in the following figure means that the 23rd word in the vocabulary is "absolute".
    - **topic_seg.txt:** It contains the topic boundaries encoded in boolean vectors. For example, if a meeting transcript, "ES2018d.txt" contains 10 paragraphs in total after being preprocessed, and there are topic boundaries after the 2nd, 5th, and 7th paragraphs, the boolean vector must be "**ES2018d:0,1,0,0,1,0,1,0,0,1".** Every line in *topic_seg.txt* corresponds to one meeting transcript.
    - **./sparse_files/*.txt** : Each txt file in the "sparse_files" folder corresponds to one of the meeting transcripts in the "txt_files" folder, and they have the same file name.

## 3. Assessment Resources

Unzipping the file, you will find that

- There are three types of XML files in the given folder :
    1. **./topics/***.topic.xml contains the information about topic segments. Each topic tag directly linked to the root indicates one topic segment that is required in text segmentation task. Each topic segment can contain a number of paragraphs given by different meeting attendees. It can also contain sub-topics.

2. **./words/***.words.xml contains the word tokens generated with the force alignment technique. Each word is associated with its start time and end time in the meeting transcript.
3. **./segments/***.segments.xml contains the paragraph boundaries, the start and end of which are denoted by the corresponding word IDs.

- ./spase_files: the file folder used to store the generated sparse representations for all the meeting transcripts.
- ./txt_files: the file folder used to save the reconstructed meeting transcripts.
- ./stopwords_en.txt: the stopword list used in word tokenization.
- ./topic_segs.txt: the file used to save the topical boundaries.
- ./vocab.txt: the file used to save the vocabulary.
- ./task_1.ipynb: the python code you are going to write for task 1
- ./task_2.ipynb: the python code you are going to write for task 2