The approach towards the problem statement can be explained in 2 steps:

**First**: Split the training data according to the splits given to us and perform topic modelling using LDA.

Choose the number of topics as many as possible (100 in my model) so as to identify the different topic distributions in the data.

As LDA is an unsupervised learning method the more the training set the better performing the model is going to be.

Hence later the train and validation set are merged to create a model so as to identify the topic distributions of the test set better.

**Second**: Creation of the test set- the test set is created by splitting the articles by sentence.

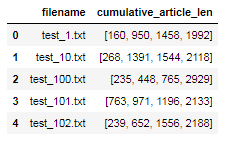
The topic loading for each sentence are obtained from the LDA model.

Cosine similarity between topic loadings of consecutive sentences is obtained and the

sentences are ranked based on the least similarity within the particular article.

Then the segments are identified by splitting the article where the least similarity between consecutive sentences is obtained and combining the rest.

Two models one with 100 topics and one with 250 topics are submitted. The output will be obtained in the format of



The 2nd column indicates the cumulative article length of different topics.

The submission format is obtained by manipulation in excel

**Data Pre-processing**

We will perform the following steps:

* Tokenization: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
* Words that have fewer than 3 characters are removed.
* All stopwords are removed.
* Words are lemmatized — words in third person are changed to first person and verbs in past and future tenses are changed into present.
* Words are stemmed — words are reduced to their root form.

**We created a bag of words for each article and ran he bag of words model using LDA to obtain the topic distributions.**

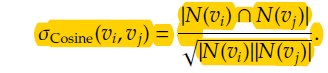
LDA Topic Modeling:

[**Topic modeling**](https://en.wikipedia.org/wiki/Topic_model) is a type of statistical modeling for discovering the abstract “topics” that occur in a collection of documents. [**Latent Dirichlet Allocation**](http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/)(LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

Used genism library in python

**Cosine Similarity:**

Cosine similarity is a [measure of similarity](https://en.wikipedia.org/wiki/Measure_of_similarity) between two non-zero vectors of an [inner product space](https://en.wikipedia.org/wiki/Inner_product_space) that measures the [cosine](https://en.wikipedia.org/wiki/Cosine) of the angle between them. Let N(vi) and N(vj) be the neighbors of nodes vi and vj, respectively. In this case, a measure of node similarity can be defined as follows



Further Improvements for the model can be obtained by supervised learning methods like LSTM with word embeddings.