The project has two parts:

First part (Spark App):

1. Build positional index from the attached Dataset & display each term as the following :
   < ***term*** ***doc1***: position1, position2 … ;
   doc2: position1, position2 ;
   etc.>

Second part:

Use the Spark App output file from the First part.

1. Compute term frequency for each term in each document (Display it).

Term	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10
Term1										
Term2										
Term3										

2. Compute IDF for each term (Display it).
3. Compute TF.IDF matrix for each term (Display it).

4. Allow users to enter phrase queries on the positional index, then compute the similarity between the query and the matched documents (Display it). Rank the documents based on their similarity scores, and return the relevant documents for the query. The phrase query can include boolean operators, such as 'AND' (e.g., phrase AND phrase) or 'AND NOT' (e.g., phrase AND NOT phrase).