Summary of exercise

1) **Preprocessing**

- Tokenization

Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.

- Stemming and Lemmatization

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance: [1]

am, are, is $\Rightarrow$ be

car, cars, car's, cars' $\Rightarrow$ car
The result of this mapping of text will be something like:

the boy's cars are different colors $\Rightarrow$
the boy car be differ color

2) **Bag of words on the dataset**

A dictionary containing the number of times each word appears in the training set was then created using the gensim library

3) **BOW corpus creation**

For each document, a dictionary was created, reporting how many words and how many times those words appear.

4) **Training LDA model**
An LDA model was generated using the BOW corpus and the gensim library. This model is saved to disk

5) **Testing on new text documents**
When a new json document is provided for testing, the dictionary and lda model are loaded from disk, the document text is preprocessed as before, and the topics are generated using the gensim library again before being written to an output json file.

References
[1] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
[2] https://github.com/susanli2016/NLP-with-Python/blob/master/LDA_news_headlines.ipynb