Problem statement

The objective of this exercise is to tag conversations with the most probable topic. Given a text transcript of a conversation between two people, we want to assign it a topic that they were most likely talking about.

Dataset

Extract the contents of ‘tagging_test.tgz’ file. This can be found here.

Each of the text blobs is the transcript of a phone conversation that happened between two people. The people were told in advance that they are supposed to speak about a certain **topic**. However, for certain conversations, the metadata about the topic that was given to the two individuals are missing.

Training

You will be given a set of text transcripts as plain text files. With each file, there will be associated one of a limited set of Topics that the individuals are conversing about. You are expected to build a model using this labeled training data.

Testing

The trained model will be fed a plain text file which is the transcript of one of the conversations that are missing the topic metadata. Your model should ingest this text and give the most likely (of the set of topics seen during training) topic that the individuals in the conversation were speaking about.

Expected submission submission

You are expected to build a model using the training data provided to you.

Document your thought process and approach thoroughly.

Finally, your submission should include a basic API service which should take a single .txt file (or a text payload) as its input and output either one of the topics or many of the topics with a clear ordering and/or associated probability of the text blob belonging to that topic.

Please note that the test data file is similar in structure to the training data - so any cleaning/preprocessing you feel is relevant should be done on the input test data also.