## Speech and Image Processing - Lab 1

Date: 2 September 2019

**Task description:**

Download/copy a text of around 1,000 words in Urdu, and store it in a UTF-8 file. The file is named corpus.txt

\*\*\*\*\*\*\*contents of corpus.txt\*\*\*\*\*\*\*\*\*\*

ان کا کہنا تھا کہ اگر انڈیا کے زیرِ انتظام کشمیر سے کرفیو فوری ہٹایا جائے، لوگوں کے حقوق بحال کیے جائیں، پوری کشمیری قیادت جو پابند سلاسل ہے اس کو رہا کیا جائے اور انھیں (شاہ محمود) کشمیری قیادت سے ملنے کی اجازت دی جائے تو مذاکرات ہو سکتے ہیں۔

طواف زیارت کے فوراً بعد طوافِ رخصت کا وقت شروع ہوجاتا ہے

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Part (a) 5 marks**

Using python, remove punctuation and diacritics from this text, so that it becomes

\*\*\*\*\*\*\*contents of corpus\_filtered.txt\*\*\*\*\*\*\*\*\*\*

ان کا کہنا تھا کہ اگر انڈیا کے زیر انتظام کشمیر سے کرفیو فوری ہٹایا جائے لوگوں کے حقوق بحال کیے جائیں پوری کشمیری قیادت جو پابند سلاسل ہے اس کو رہا کیا جائے اور انھیں شاہ محمود کشمیری قیادت سے ملنے کی اجازت دی جائے تو مذاکرات ہو سکتے ہیں

طواف زیارت کے فورا بعد طواف رخصت کا وقت شروع ہوجاتا ہے

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Part (b) 5 marks**

Using python, make a file vocabulary.txt, that contains all the unique words in the corpus\_filtered.txt file and their frequencies. The file should be sorted according to frequencies

\*\*\*\*\*\*\*\*\*\*\*\*\*\*contents of vocabulary.txt\*\*\*\*\*\*\*\*

کا 30

سے 27

تو 23

کشمیر 7

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*