Adv. Natural Language Processing B659
Assignment 6

Aniket Shenoy, Gaurav Derasaria, Pranab Bhadani, Sharad Ghule

a. We chose to use MALLET to perform topic modeling.

b.

Problem Statement: We wanted to analyze the inaugural speeches of various presidents of the US over the years to see if there are any similarities or patterns given their party affiliation.

Data: We scraped text data manually of presidential inuagural speeches.

Output:

The following topics were generated using MALLET with the 5 most frequent keywords for each topic:

0 1 world americans time government today freedom work god strength children national nations pledge lives fellow place begin faith earth economy

1 1 hope long tonight man change history call states sides support peace hand generation answer century war told it's democracy fall

2 1 nation country citizens great united power president life love story good service courage promise spirit america's live day common means

3 1 america people american make back years millions dreams jobs bring party bless stand made families women heart belongs states born

4 1 government public free present called share duty future system act day greater administration find assure dare communities happiness commitment side

Composition of topics in each inaugural speech:

0 file: clinton.txt
0.3780487804878049 0.23658536585365852 0.2048780487804878
0.17073170731707318 0.00975609756097561

1 file: kennedy.txt
0.2681564245810056 0.3538175046554935 0.21415270018621974
0.05027932960893855 0.11359404096834265

2 file: washington.txt
0.05244122965641953 0.07414104882459313 0.18625678119349007
0.081374321880651 0.6057866184448463

3 file: bush.txt
0.22121212121212122 0.07575757575757576 0.4712121212121212
0.1 0.1318181818181818

4 file: trump.txt
0.17229729729729729 0.013513513513513514 0.25
0.48817567567567566 0.07601351351351351

5 file: obama.txt
0.15842839036755388 0.4093789607097592 0.13054499366286437
0.29404309252217997 0.0076045627376425855

6 file: reagan.txt
0.38150903294367694 0.14240170031880978 0.11158342189160468
0.1997874601487779 0.1647183846971307

c.

Analysis of the Output:

It can clearly be seen that George Washington's speech stands out from the speeches of the other presidents given that he was the first president. It can be seen that there is a shift in the topics discussed in the President's speech over time. President Trump's speech clearly stands out as well with frequent mentions of words that are more closely related to topics that elicit emotional responses. Topic 1 has higher weightage in speeches of President Kennedy and Obama which can be attributed to their affiliation to the Democratic party. President Clinton and Reagan's speeches focus more on the economy and global topics.

Given the data we used and the topics that came out of the data, we found the annotation satisfactory to the most extent. We could see that the composition of the topics roughly matched our preconceived notions of the speeches.

The topics generated did have some stop words that should have been removed. The software does not take into account the semantics of the texts and uses statistical methods that take into account the frequencies of words.

From a social scientist's perspective, using a command line interface seems like the part that may pose as a barrier.