Gibbs Sampling

topicwords.csv
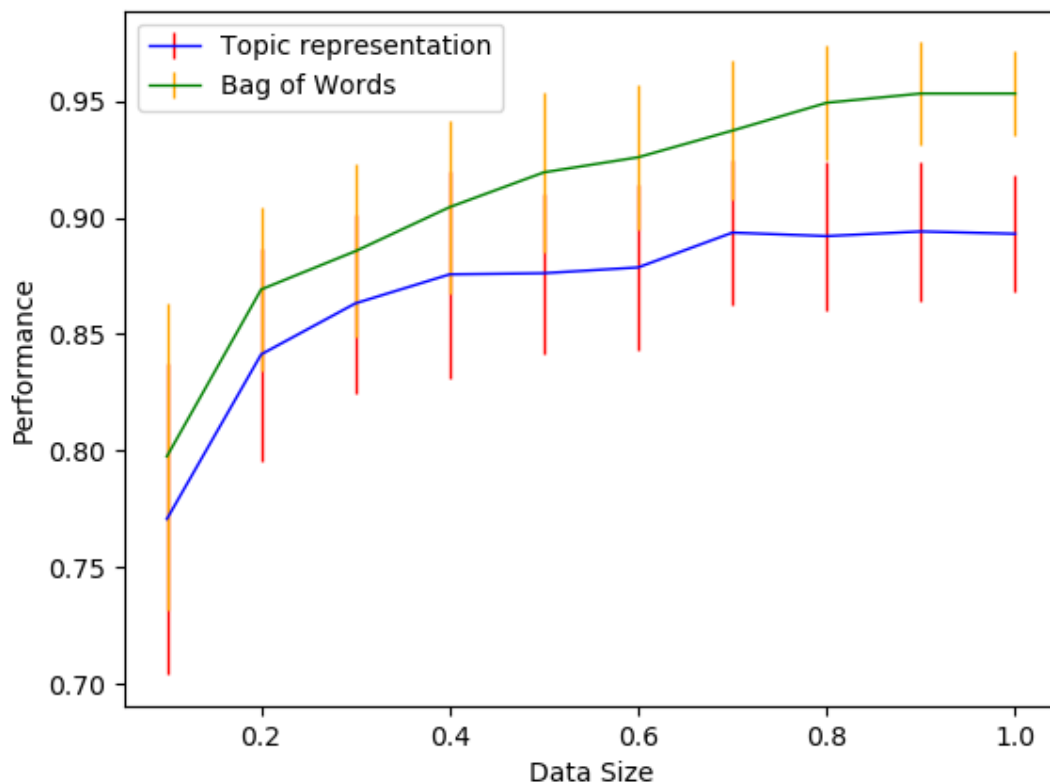
0	['nasa', 'part', 'mars', 'spacecraft', 'gov']
1	['clutch', 'shifter', 'drive', 'don', 'sho']
2	['car', 'ford', 'cars', 'small', 'driving']
3	['edu', 'gif', 'uci', 'ics', 'incoming']
4	['system', 'such', 'edu', 'oort', 'solar']
5	['even', 'make', 'speed', 'good', 'time']
6	['edu', 'insurance', 'don', 'uiuc', 'geico']
7	['mission', 'hst', 'pat', 'access', 'net']
8	['oil', 'light', 'service', 'lights', 'change']
9	['power', 'point', 'current', 'place', 'based']
10	['car', 'engine', 'price', 'used', 'turbo']
11	['edu', 'mustang', 'blah', 'engines', 'info']
12	['station', 'space', 'option', 'program', 'redesign']
13	['writes', 'edu', 'article', 'apr', 'find']
14	['sky', 'earth', 'etc', 'temperature', 'things']
15	['space', 'shuttle', 'launch', 'cost', 'two']
16	['eliot', 'cars', 'diesels', 'article', 'washington']
17	['space', 'bill', 'moon', 'internet', 'world']
18	['henry', 'edu', 'toronto', 'spencer', 'writes']
19	['high', 'people', 'don', 'reason', 'another']

Observation:

- In the above table we can see that for every row the top 5 topics are making sense and are related to each other from which we can make general guess of the content.
- Like for 0 : we can say the topic will be related some space project of NASA
- Thus we can conclude using this method helps us to find relatable topics for the contents observed in the documents.
- It takes approximately 20 minutes to run.

Classification

Performance Graph for both the representations

Observation:

- From the above graph we can clearly see that the Bag of words representation out performs the topic representation.
- With data size the performance for both the representations increases.
- The error bar clearly tell us that bag of words representation has better certainty than topic representation.
- The time observed for topic representations is approximately 10 seconds whereas for bag of words it is approximately 45 seconds.
- Bag of words does perform better but topic representation takes less time comparably.