Data Assignment #6

20229007 Sungyoung Ryu (BTM539) AI for Business Management

[LSA, LDA, and BERTopic]

1. Among 3 model, they take TF-IDF(Text freq-Inverse Docs freq) but LSA(Semantic) or LDA(Dirichlet), they begins from topic distinguish first then make an array into top topic to less frequently exposed. But BERTopic has latent grouping and clustering first and then distinguish the topics with continuous bag of the words and skip-gram approach.

lsa\_topics = []

for model in model\_list:

    lsa\_topics.append(model.show\_topics(num\_topics=-1, num\_words=10))

lda\_topics = []

for model in model\_list\_lda:

    lda\_topics.append(model.show\_topics(num\_topics=-1, num\_words=10))

bert\_topics = []

for topic in range(1, 16):

    topic\_words = topic\_model.get\_topic(topic)

    bert\_topics.append([word for word, score in topic\_words])

random.seed(123)

lsa\_sample = random.sample(lsa\_topics, 5)

lda\_sample = random.sample(lda\_topics, 5)

bert\_sample = random.sample(bert\_topics, 5)

topics\_df = pd.DataFrame({'LSA': lsa\_sample, 'LDA': lda\_sample, 'BERTopic': bert\_sample})

for model in topics\_df.columns:

    for i in range(5):

        topics\_df.at[i, model] = [word for word in topics\_df.at[i, model]]

topics\_df

1. 1st column is randomly chosen 5 topics from LSA model, as you see, it haven’t merged into single line – perhaps some code hadn’t runned. 2nd column is random 5 topics in LDA model. And The last column is 5 random topics from BERTopic model. LDA and LSA might be closed but I can see some point there are inadequate words in the same bag of word, due to not enough “Chunksize” perhaps. Using BERTopics, much quicker and organized better.

텍스트이(가) 표시된 사진

자동 생성된 설명

\* Thank you, professor. Though it’d over the deadline, but I couldn’t make it at all without your help.