Prof. Dr. Andreas Hotho,
M.Sc. Janna Omeliyanenko
Lecture Chair X for Data Science, Universität Würzburg

11. Exercise for "Sprachverarbeitung und Text Mining"

04.02.2022

1 Knowledge Questions

1. In the lecture, the Dirichlet Distribution was introduced. Describe the role of the Dirichlet parameter $\overrightarrow{\alpha}$ in your own words. What are the effects of changing these parameter?
2. What are the input and output of the Gibbs Sampling Algorithm?
3. How can LDA be used to cluster words in a corpus? What do the resulting clusters contain?

2 LDA – Generative process

1. In the lecture we introduced the generative process of LDA to create a document. Describe this generative process in your own words.
2. You want to create a document according to the idea of the generative LDA process. The following document-topic distribution θ_{d_1} and topic-term distributions ϕ_k for topics $k \in (animals, sports, interest)$ have already been sampled from the Dirichlet distributions with parameters $\overrightarrow{\beta}$ and $\overrightarrow{\alpha}$:

$$\theta_{d_1} = \begin{pmatrix} p(animals) & = & 0.4 \\ p(sport) & = & 0.35 \\ p(interest) & = & 0.25 \end{pmatrix} \qquad \phi_{animal} = \begin{pmatrix} p(cat) & = & 0.3 \\ p(dog) & = & 0.2 \\ p(mouse) & = & 0.25 \\ p(ball) & = & 0.05 \\ p(play) & = & 0.1 \\ p(like) & = & 0.1 \end{pmatrix}$$

$$\phi_{sport} = \begin{pmatrix} p(cat) & = & 0.0 \\ p(dog) & = & 0.02 \\ p(mouse) & = & 0.0 \\ p(ball) & = & 0.4 \\ p(play) & = & 0.5 \\ p(like) & = & 0.08 \end{pmatrix} \qquad \phi_{interest} = \begin{pmatrix} p(cat) & = & 0.07 \\ p(dog) & = & 0.04 \\ p(mouse) & = & 0.01 \\ p(ball) & = & 0.13 \\ p(play) & = & 0.15 \\ p(like) & = & 0.6 \end{pmatrix}$$

Generate the document d_1 with 5 words using these distributions. You can use random.org[1] to generate random numbers, or the function

```
numpy.random.choice(['pick', 'one', 'please'], p=[0.3, 0.5, 0.2])
```

of the Python library numpy. Specify the resulting document, as well as the topics of each word.

[1] `https://www.random.org/decimal-fractions/`

3 Gibbs Sampling for LDA

Consider the following two documents from which stop words have already been removed:
D_1: `Lecture LDA Gibbs Exam Institute Fun`
D_2: `Fun Exam LDA Gibbs`
Given is the following assignment of the topics to the words.

Topic	A	B	C	A	A	C
Word	Lecture	LDA	Gibbs	Exam	Institute	Fun

Topic	C	A	B	B
Word	Fun	Exam	LDA	Gibbs

1. Create a topic-term matrix that contains the corresponding counts over both documents.
2. Based on your matrix, calculate the topic distribution $\vec{z}$.
3. Calculate the document-topic distribution $\vec{z}_d$ for the first document D_1.
4. Given the previous results, perform a step of Gibbs Sampling for the word $w_1 =$ `Gibbs` in the first document D_1 as shown in the lecture. The Apriori counts for $\vec{\alpha}$ and $\vec{\beta}$ are set to $\vec{1}$. Which topic would you most probably assign to `Gibbs` after the step?