How does LDA work?

* It extracts the potential topics from a dataset, for example, the job description for a specific course, and assigns a probabilistic mixture of these topics simultaneously into the job description.

What can LDA do?

* It can later help with assigning topics to a new similar dataset, such as job description for another position.
* It can classify unstructured job description into a mixture of topics in a way that their underlying content can be revealed.
* It can be used to reveal trends of these topics in the future, evaluating the effectiveness of the instruction and/or instructors in addressing these topics.

Data preprocessing

1. convert to lowercase
2. remove special characters and tokenize them into terms
3. remove stop words
4. stemming
5. construct term-document matrix

What can be included?

* Distribution of job descriptions by number of words (min, mean, media, max word length)

How to implement LDA (in python)?

* Library nltk can be used to preprocess the job description
* The term-document matrix can be imported into the genism library, for the later use in LDA analysis