LDA Topic Modeling

Steffen Pielström, Galway 06.12.2018

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

DARIAH-DE
Digitale Forschungsinfrastruktur für die
Geistes- und Kulturwissenschaften

The Problem...

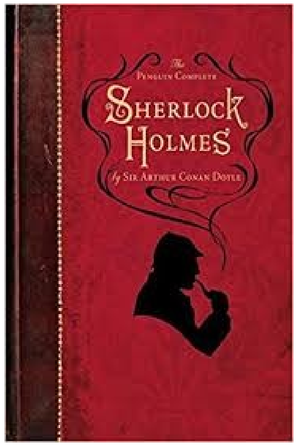
The Problem...

… what are all these documents about?

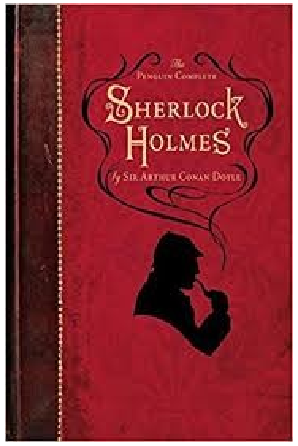
And which ones are interesting for me?

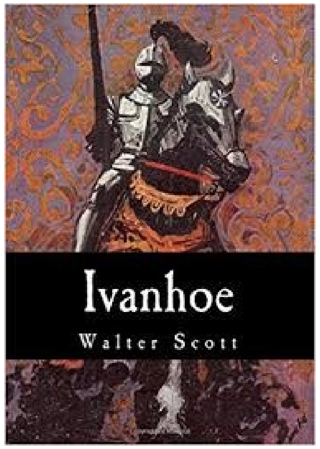
What are they about?

What are they about?

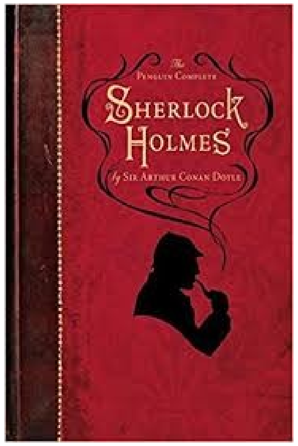


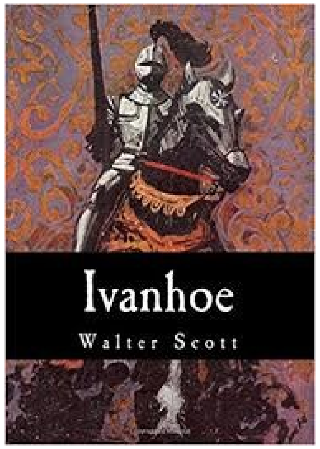
What are they about?

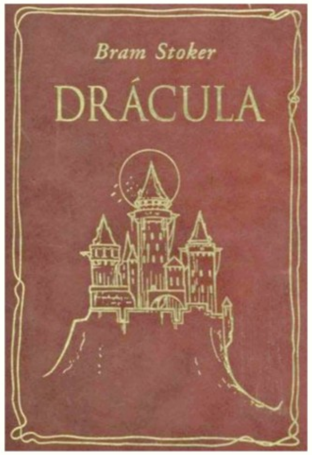




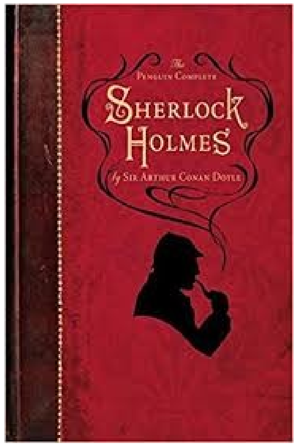
What are they about?

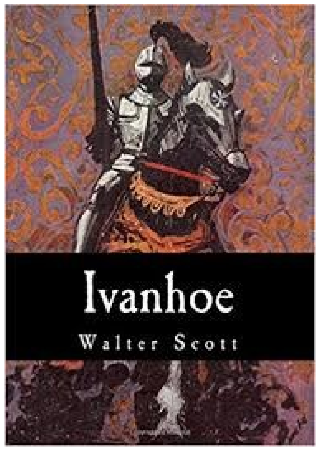


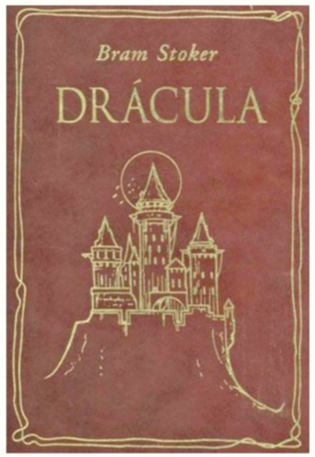


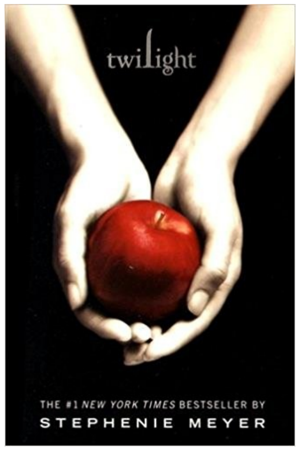


What are they about?

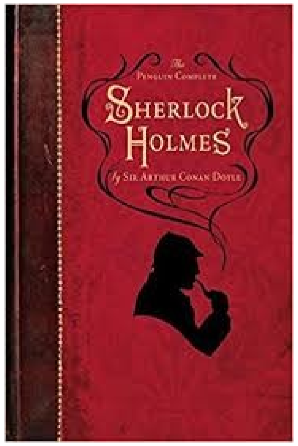


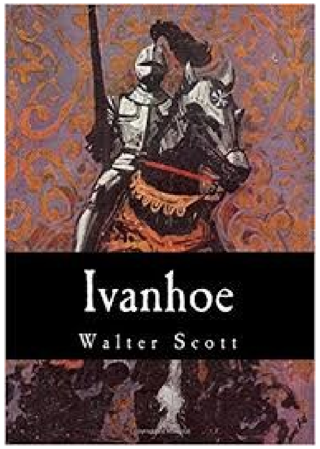


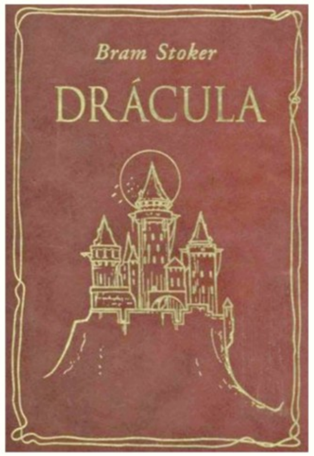


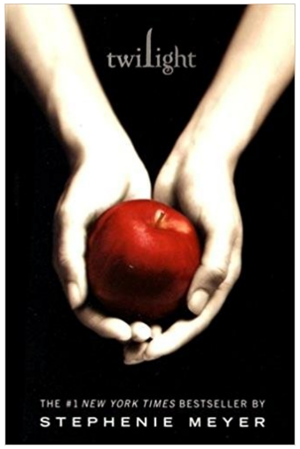


What are they about?

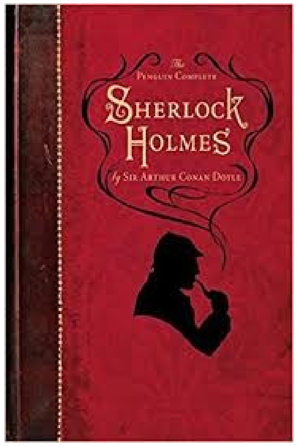


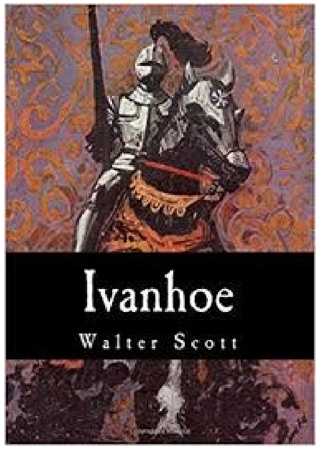


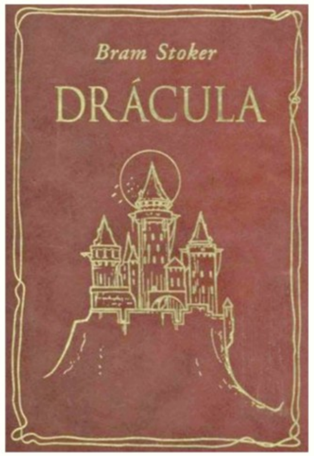


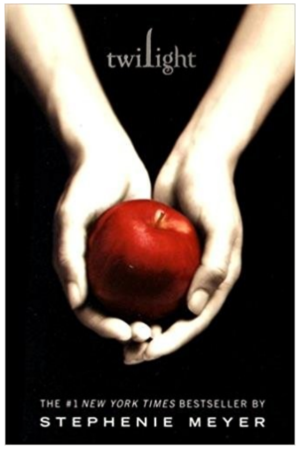


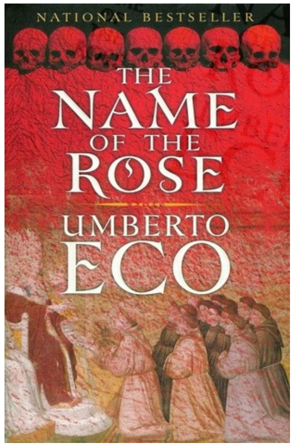
What are they about?



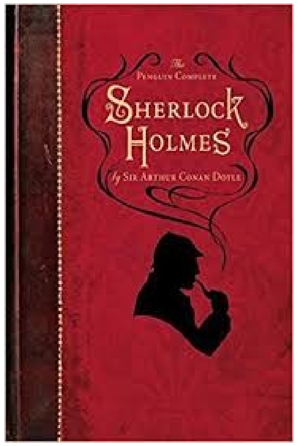


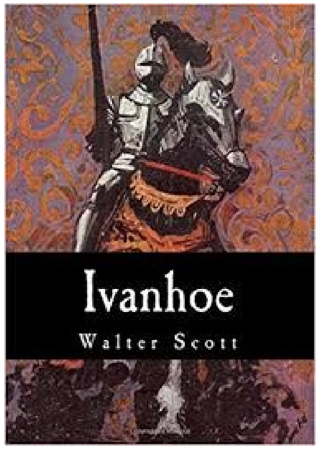


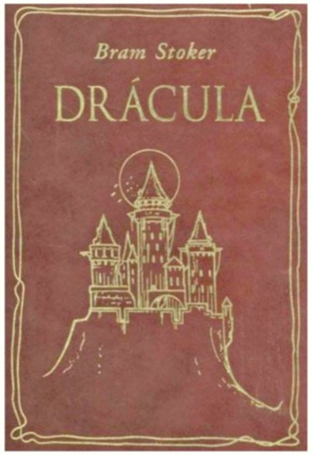


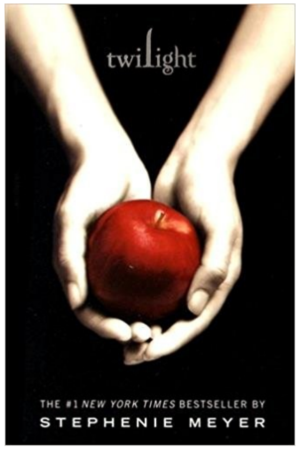


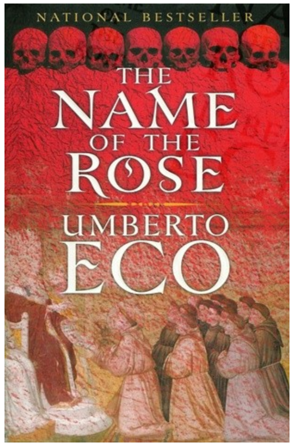
What are they about?





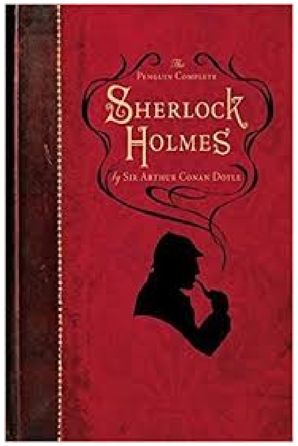


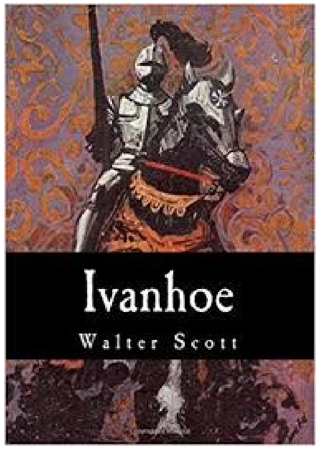


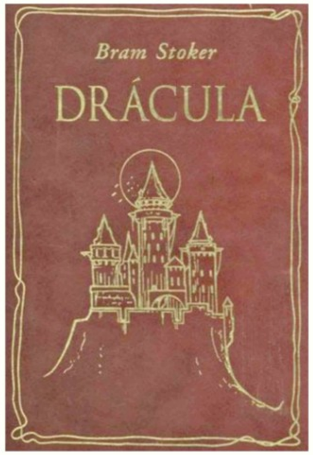


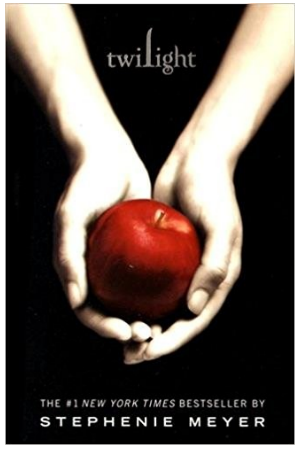
EVER FELT LIKE YOU WERE SURROUNDED BY ZOMBIES?
SHAUN OF THE DEAD
A ROMANTIC COMEDY. WITH ZOMBIES.
www.romzomcom.com
SHUFFLING INTO CINEMAS FROM 9TH APRIL

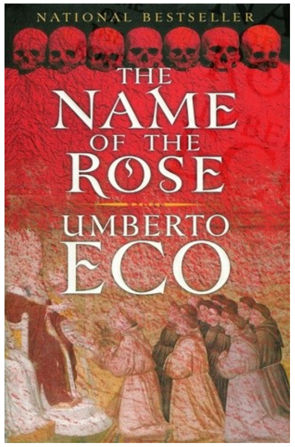
What are they about?











EVER FELT LIKE YOU WERE SURROUNDED BY ZOMBIES?
SHAUN OF THE DEAD
A ROMANTIC COMEDY. WITH ZOMBIES.
www.romzom.com
SHUFFLING INTO CINEMAS FROM 9TH APRIL
A ROMANTIC COMEDY. WITH ZOMBIES.

LILY JAMES
SAM RILEY
JACK HUSTON
BELLA HEATHCOTE
DOUGLAS BOOTH
WITH MATT SMITH
WITH CHARLES DANCE
AND LENA HEADEY
PRIDE + PREJUDICE + ZOMBIES
Based on the Best-Selling Novel by Jane Austen & Seth Grahame-Smith
/LionsgateUK @LionsgateUK #PPZmovie
IN CINEMAS FEB 2016
LIONSGATE

what are all these documents about?

Some are about more than one thing!

Which ones are interesting?

Should I use

“castle” …

as a search term?

Which ones are interesting?

Should I use

“castle” … “fortress” …

as a search term?

Which ones are interesting?

Should I use

“castle” … “fortress” … “stronghold”…

as a search term?

Which ones are interesting?

Should I use

“castle” … “fortress” … “stronghold”…

“fastness”

as a search term?

Which ones are interesting?

Should I use

“castle” … “fortress” … “stronghold”…

“fastness” or “burgh”

as a search term?

We want...

- Groups of semantically related words

We want...

- Groups of semantically related words
- To see the contribution of each group to each text

We use…

- Distributional semantics

We use…

- Distributional semantics
- Generative model

We use...

- Distributional semantics
- Generative model
- Probabilistic procedure

We can do...

We can do...

- Searching relevant texts

We can do...

- Searching relevant texts
- Text classification

We can do...

- Searching relevant texts
- Text classification
- Diachronic analysis of themes in literary history

We can do...

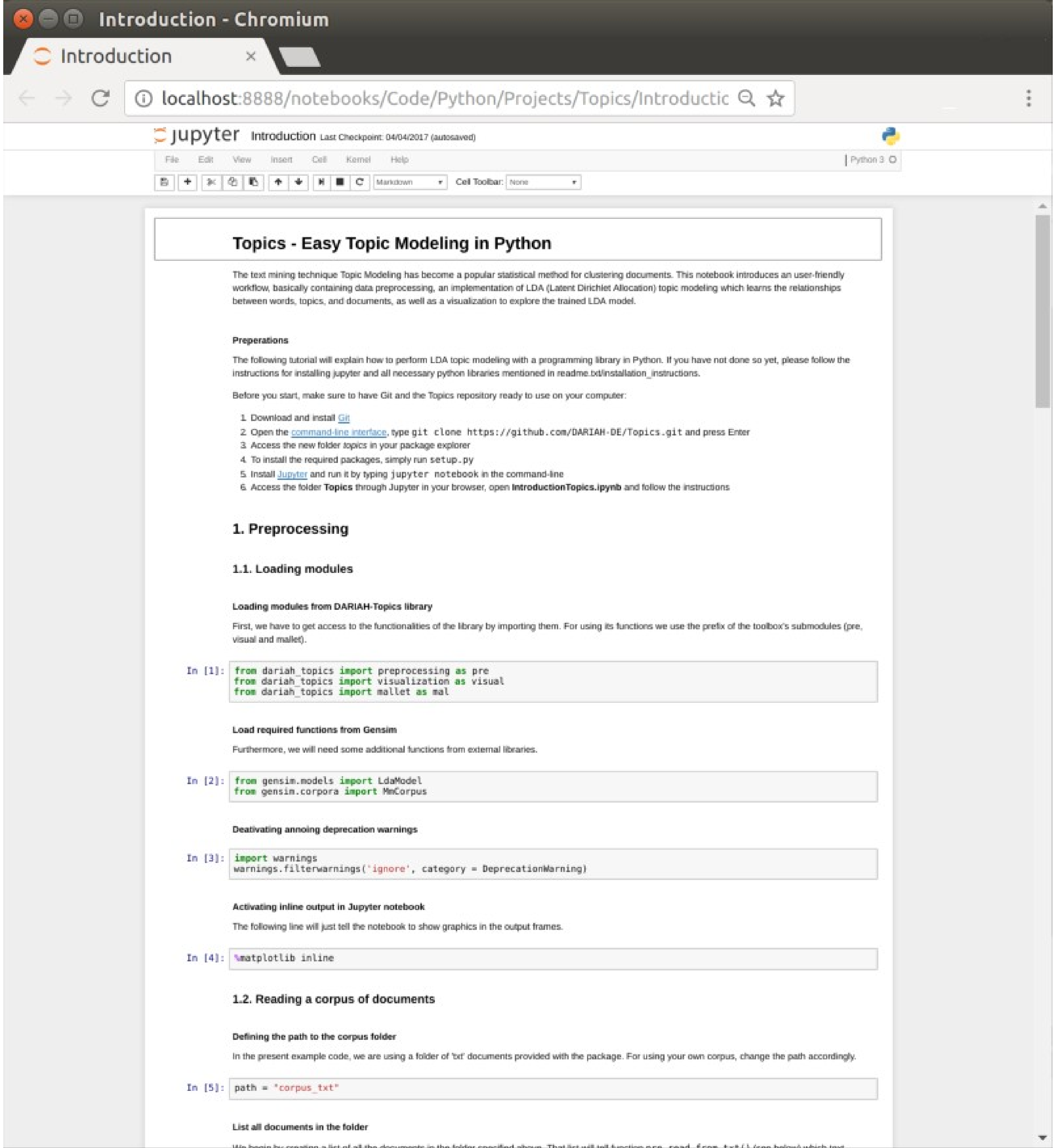
- Searching relevant texts
- Text classification
- Diachronic analysis of themes in literary history
- Topics as a research topic

MALLET
http://mallet.cs.umass.edu/

```
bin/mallet import-dir --input ~/MALLET_test/corpus --output topic-input.mallet --keep-sequence --remove-stopwords
```

```
bin/mallet train-topics --input topic-input.mallet --num-topics 10 --output-state topic-state.gz
```

dariah-de.github.io/TopicsExplorer/


Introduction - Chromium
Introduction
localhost:8888/notebooks/Code/Python/Projects/Topics/Introductic
jupyter Introduction Last Checkpoint: 04/04/2017 (autosaved)
File Edit View Insert Cell Kernel Help
Python 3
Markdown
Cell Toolbar: None

Topics - Easy Topic Modeling in Python

The text mining technique Topic Modeling has become a popular statistical method for clustering documents. This notebook introduces an user-friendly workflow, basically containing data preprocessing, an implementation of LDA (Latent Dirichlet Allocation) topic modeling which learns the relationships between words, topics, and documents, as well as a visualization to explore the trained LDA model.

Preperations

The following tutorial will explain how to perform LDA topic modeling with a programming library in Python. If you have not done so yet, please follow the instructions for installing jupyter and all necessary python libraries mentioned in readme.txt/installation_instructions.

Before you start, make sure to have Git and the Topics repository ready to use on your computer:

1. Download and install Git
2. Open the command-line interface, type `git clone https://github.com/DARIAH-DE/Topics.git` and press Enter
3. Access the new folder *topics* in your package explorer
4. To install the required packages, simply run `setup.py`
5. Install Jupyter and run it by typing `jupyter notebook` in the command-line
6. Access the folder **Topics** through Jupyter in your browser, open **IntroductionTopics.ipynb** and follow the instructions

1. Preprocessing

1.1. Loading modules

Loading modules from DARIAH-Topics library

First, we have to get access to the functionalities of the library by importing them. For using its functions we use the prefix of the toolbox's submodules (pre, visual and mallet).

```
In [1]: from dariah_topics import preprocessing as pre
        from dariah_topics import visualization as visual
        from dariah_topics import mallet as mal
```

Load required functions from Gensim

Furthermore, we will need some additional functions from external libraries.

```
In [2]: from gensim.models import LdaModel
        from gensim.corpora import MmCorpus
```

Deativating annoing deprecation warnings

```
In [3]: import warnings
        warnings.filterwarnings('ignore', category = DeprecationWarning)
```

Activating inline output in Jupyter notebook

The following line will just tell the notebook to show graphics in the output frames.

```
In [4]: %matplotlib inline
```

1.2. Reading a corpus of documents

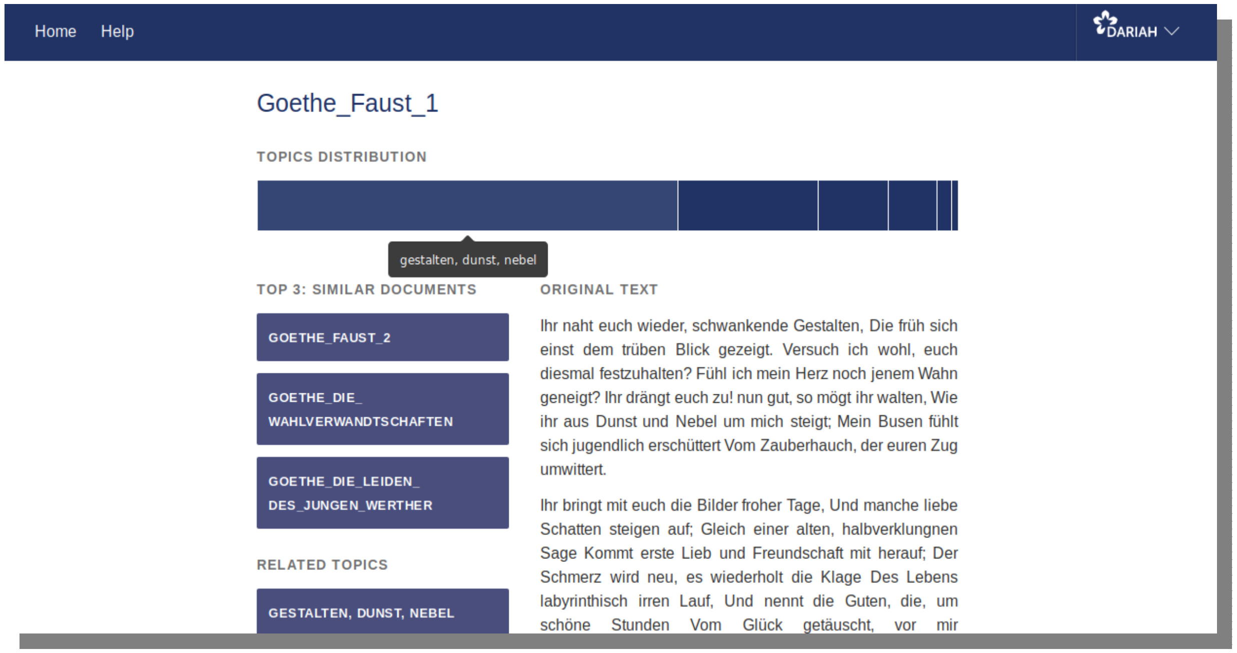
Defining the path to the corpus folder

In the present example code, we are using a folder of 'txt' documents provided with the package. For using your own corpus, change the path accordingly.

```
In [5]: path = "corpus_txt"
```

List all documents in the folder

dariah-de.github.io/TopicsExplorer/



Goethe_Faust_1

TOPICS DISTRIBUTION

TOP 3: SIMILAR DOCUMENTS

GOETHE_FAUST_2

GOETHE_DIE_WAHLVERWANDTSCHAFTEN

GOETHE_DIE_LEIDEN_DES_JUNGEN_WERTHER

RELATED TOPICS

GESTALTEN, DUNST, NEBEL

ORIGINAL TEXT

Ihr naht euch wieder, schwankende Gestalten, Die früh sich einst dem trüben Blick gezeigt. Versuch ich wohl, euch diesmal festzuhalten? Fühl ich mein Herz noch jenem Wahn geneigt? Ihr drängt euch zu! nun gut, so mögt ihr walten, Wie ihr aus Dunst und Nebel um mich steigt; Mein Busen fühlt sich jugendlich erschüttert Vom Zauberhauch, der euren Zug umwittert.

Ihr bringt mit euch die Bilder froher Tage, Und manche liebe Schatten steigen auf; Gleich einer alten, halbverklungnen Sage Kommt erste Lieb und Freundschaft mit herauf; Der Schmerz wird neu, es wiederholt die Klage Des Lebens labyrinthisch irren Lauf, Und nennt die Guten, die, um schöne Stunden Vom Glück getäuscht, vor mir

Severin Simmler
Evelyn Gaier
Blerta Veseli
Thora Hagen
Stefan Krywinski
Thorsten Vitt

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

DARIAH-DE Digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften