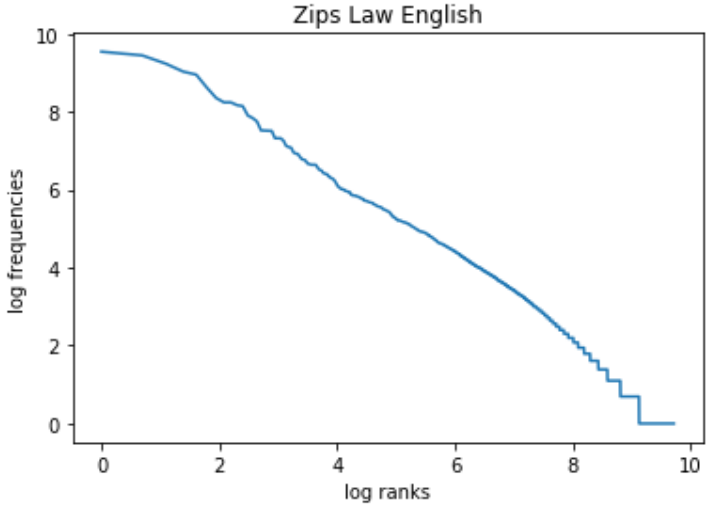
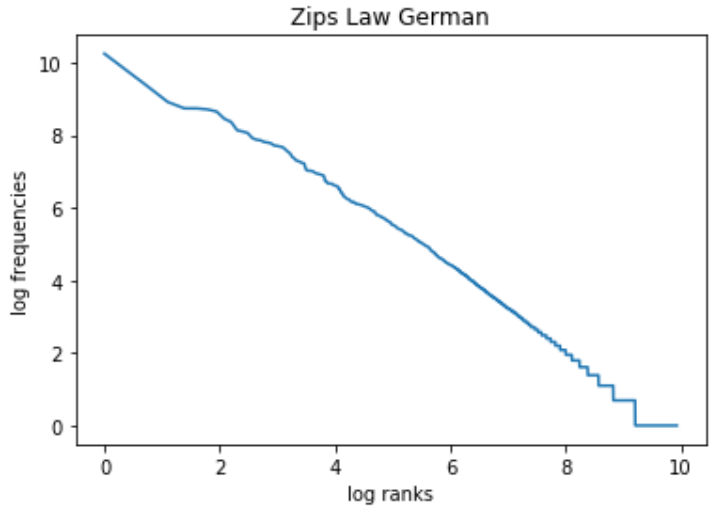
Statistical Natural Language Processing
Assignment 1

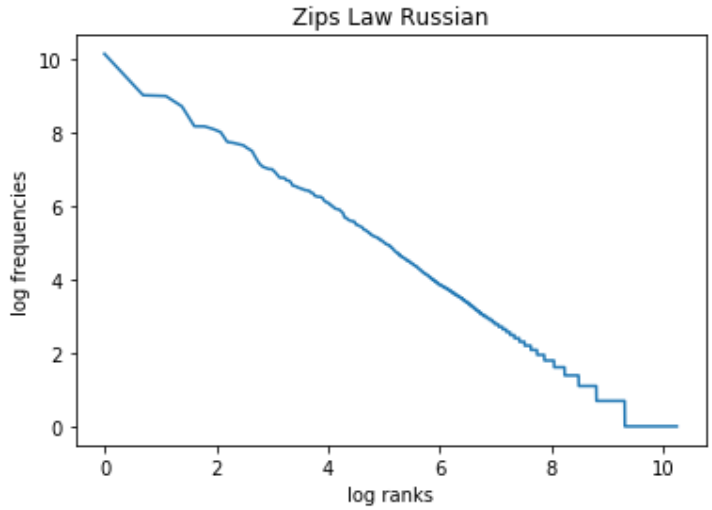
Sven Stauden 2549696
Janis Landwehr 2547715
Carsten Klaus 2554140

Exercise 1

1.1 Plot the frequency of the words in the three novels against their rank (both axes on a log scale)







1.2 Write down the conclusions made by Hermione after observing the different plots. How do you explain the differences between them to Hermione?

In general, Zip's law applies to all 3 languages. However, there seems to be a difference between the English plot and the other two in regard of the highest frequency words. While the German and Russian language have a strong decline at the beginning, the English text seems to have many words

that are relatively similar in regard of their frequency.
This could be based on the language's morphology. While English words don't differ in regard of the grammatical person and grammatical case (besides the 3rd-person singular), there is a many variations in the other languages for the same word. Therefore, many English words appear equally often, while this is not the case for the other two languages.
Furthermore, one can observe that all plots have many words that appear only once in the text. However, since the underlying text was copied from books in PDF-format, words such as the page number were not removed. These page numbers lead to an disproportional increases of this number.

1.3 Write your explanation to Parvathi on how does the dependency between the morphology of the languages and Zipfs law looks like.

In comparison to the German and Russian language, English is rather morphologically poor. Verbs are a good example. As described in exercise 1.2 the English language does not distinguish between verb forms given a specific tense. On the other hand the consideration of articles is also very interseting. They are frequent words in most languages and can be seen as indicators for morphological richness. The german Language has three definite articles (`der, die, das`) and two indefinite articles (`ein, eine`). English on the other side just provides `the` and `a`. The flat beginning of the English log-log-plot indicates that there is not much difference of the frequency between the most common words. The steepness of the other two languages show us that there are words with similar ranks but diverse frequencys. This is a sign for a more distinctive language morphology.

Exercise 2

See the results in the submitted jupyter notebook.

Exercise 3

The linguist and cognitive scientist Noam Chomsky generally criticizes the development of difference current research fields, especially in artificial intelligence. He denounces that recent scientific work in artificial intelligence concentrates on the consideration and evaluation of statistical models which rely on empirical collected data. Especially in linguistics, the goal for research shifted from understanding an underlying mathematical and logical model of language to reducing the topic to a black box. The internal functionality of this black box does not get studied anymore unlike the correlations between input and output which may allow useful predictions but do not deliver a full understanding of the concept.

As response to this criticism, the computer scientist and director of research at Google Peter Norvig disagrees with Chomsky's point of view. He says that in history, experiments have always been an important source of data, used to conclude scientific findings. Hence modeling the world with statistical data and not directly searching for the solution have always been an essential part of science. Further, language is a highly complex and evolutionary developing concept which might only can be described with billions of parameters. Even if the problem of language modeling would be "solved", due to the incomprehensibility for humans, probabilistic methods would still have a higher usability. Also in natural processes like recognition, probability measures play an important role and therefore must not be seen as irrelevant to an underlying system.

The concerning of Chomsky are indeed relevant and worth considering. Modeling language is one of the still unsatisfied challenges of our time; having a look at Google translator, this tool might be useful but not trustful. The general research of artificial intelligence regarding machine learning evolved to a not fully explored but highly used concept. The challenge of finding a well performing ML model often ends in an almost blind search for optimal parameters which can only be found by testing which does not seem to rely on well researched science. Similar to the early medicine, where "witches"

provided people with certain herbs and mixtures against specific complaints without knowing anything about the chemical reactions but only relying on experiences which can be seen as statistical data, the research of language modeling or general AI might be in such a development stage. Focusing in the application is not wrong but understanding the underlying system still must remain an essential goal, maybe for the future.

Remark to the exercise sheet

We know, that creating a didactically valuable exercise sheet requires much preparation time and might still contain errors or misunderstandings after a few iterations. Anyway, in our eyes, the tasks of this exercise sheet are very confusing, seem to be incomplete and did not bring the clarification about the topic as they should. We spend more time in figuring out what the exercise should teach us, what exactly is asked, and studying Piazza than in implementing the code and even reading the $\tilde{4}0$ pages for exercise 3. We hope that at least the description of the exercises gets more explicit for the next sheets so we can focus on natural language processing.