Analysing data:
Assignment 1

Assignment 1
Version 1

Analysing Data
Academic year 2024-2025
Date: February 17th 2025

Rachel Hartemink
Student ID: s3695913

Link to the repository:
https://github.com/Rachelslag/Analyzing-data-assignment-1

Stemmer analysis

The Porter Stemmer seems to produce more recognizable and readable words than the Lancaster Stemmer. For example, *doctor* remains *doctor* using the porter stemmer but becomes *doct* using the lancaster stemmer. Similarly, using the porter stemmer *river* and *rose* remain unchanged but become *riv* and *ros* when the lancaster stemmer is used. Due to the more aggressive stemming rules of the lancaster stemmer, words can sometimes also change meaning: *your* becomes *yo* and now it is unclear if the word is *yo* or a shortened version of something else (such as a form of greeting). That does not mean that the stems always remain readable using the porter stemmer. *Was* for example becomes *wa,* which is not immediately recognizable.

The overall ranking remains largely the same for both stemmers, but while Lancaster reduces words, Porter keeps them more intact. The word rankings remain mostly unchanged, but certain words shift slightly due to stemming differences. Words that were shortened significantly by Lancaster may see altered frequencies, as they merge with similarly stemmed words.

The unstemmed text provides the most clarity but has many variations of the same root word. Porter stemmer seems like a good middle ground, it preserves readability and helps normalize words while avoiding extreme shortening. Lancaster stemmer is useful for cases where aggressive reduction is needed, but it may alter meaning too much.

POS analysis

The noun frequency is relatively similar across the three translations, which suggest that all three texts largely follow the same narrative structure. German has more punctuation (18,397 vs. 13,947 in Dutch and 15,310 in English), indicating that German tends to use more complex sentence structures, which often require additional punctuation marks. Additionally, German has a significantly higher frequency of adverbs (9,266 vs. 4,883 in English and 5,761 in Dutch), suggesting that German modifies meaning more frequently through adverbial structures compared to the other languages.

Dutch, however, stands out for having the highest number of adjectives (5,048 vs. 4,404 in English and 2,914 in German), indicating a preference for more descriptive language. The

higher frequency of pronouns in Dutch and English (10,021 & 10,268 vs. 8,771 in German) shows a preference for explicitly stating subject pronouns. In contrast, German often seems to omit them in its sentence structure. Dutch and English also have a lot more verbs than German (11,768 & 11,292 vs. 9,468).

In terms of auxiliary verbs, English has the most (4,377 vs. 3,740 in German and 3,839 in Dutch), which aligns with English's reliance on auxiliary verbs (e.g., 'have', 'be', and 'do') for grammatical construction. English has 3,010 proper nouns, indicating a moderate use of specific names, places, or things in the translation. German has fewer proper nouns (2,181), suggesting a preference for more generalized language or the omission of specific names in certain contexts. In contrast, Dutch has the highest count (3,474), reflecting a tendency to incorporate more specific references, possibly contributing to a more detailed or formal narrative style.

Automatic NER analysis

The scores for the accuracy of the Named Entity Recognition are:

- Precision: 0.5
- Recall: 0.75
- F1-score: 0.6

Precision tells us how many of the entities are correctly identified by the model relative to the total number of identified entities. In this case, the model correctly identified 50% of the entities, meaning it has a relatively low precision and a high number of false positives. This indicates that the model struggles with accurately classifying entities.

Recall, on the other hand, measures how many of the actual entities were identified by the model. With a recall of 0.75, the model correctly identified 75% of the actual entities, but missed 25% (the false negatives). This means the model does a good job of finding relevant entities but still misses a few, indicating that its coverage is good but not perfect.

The F1-Score, which is the harmonic mean of precision and recall, is 0.6. This suggests that while the model performs reasonably well overall, there is room for improvement, especially in balancing precision and recall.

There are several possible reasons why spaCy's named entity recognition (NER) misclassified entities in this specific text. Some of the entities are ambiguous or context-dependent.
For example, *Barcelona* most often refers to a city (GPE), but in this case two out of the three mentions refer to a fictional planet (LOC). The NER model might misclassify *Barcelona* because

spacy's NER model is assumably primarily trained on real-world data. While it can handle a wide range of common entities, it may not always perform well with fictional or domain-specific terms (such as names from TV shows, books, or fictional worlds). Terms like *Gallifreyan*, *Torwash*, or *TARDIS* come from the Doctor Who universe and might not be in the model's training data, so the model either ignores them (e.g., *Rassilon*) or misclassifies them (e.g., *Gallifreyan*).