BAN432 Fall 2019 - Third Assignment

Task 1: Network analysis in a portfolio choice problem

Group 6: Patrick Becher, Kai Jing Zhou, Lingling Tong, Susanne Tietze, Yao Liu

Choices in DTM:

- **terms appearance in documents:** the more words one use, the higher similarity becomes, one inflate the similarity score but meaning does not become better
- **lower case:** does not matter how the word is written, only that it appears and thereby consider wrong written words
- **punctuation:** is not important for our analysis because we are interested in words, for us it is just noise
- **stopwords:** these words often appear in a document but does not have any specific meaning that would them make relevant for our analysis
- **numerics:** is not important for our analysis because we are interested in words, for us it is just noise
- **word lengths:** lower bound: words shorter than 3 letters do not have much meaning, upper bound: make sure that you do not a large string which is wrong written

Correlations

```
# Correlations between the oil industry portfolio's monthly returns and the four oil tracking
# portfolio's return
correlations
```

```
##           all.wordclasses nouns.adj
## weightBin       0.7453267 0.7996220
## no.weight       0.7682501 0.7512502
```

Correlations: All correlations are quite high and nearly similar. This shows that the approach works well to track oil industry´s development and that the choices of weighting / no weighting of term frequencies and using all words vs. only using nouns and adjectives does not influence the result heavily. The approaches that use only nouns and adjectives were more successfull because they deliver higher correlations in particular with weights / normalization applied. If all words are used the approach without using weights / normalization performed better.

Probable reasons: If one use only nouns and adjectives the similarity between oil-related companies is becoming higher and therefore algorithm picks more likely these companies. The weighting / normalization make sure that some large companies does not influence the similarity score too heavily. If one uses all words than the weighting does not improve algorithm but works opposite.

Bonus

```
# Mean correlation of oil industry portfolio's monthly returns and 50 sample portfolio's return
mean(bonus.correlations)
```

```
## [1] 0.3503278
```

Explanation: Compared to the former approach the correlation is much lower. This shows that taking a random sample does not work well. A more more sophisticated approach is needed to track oil industry´s development.