IN4080, 2022, Exercise set 4: Oct. 11

Use rougly one hour for the following exercise, and one hour for mandatory assignment 2.

This is an example of an exercise which could be given to the exam.

We will represent words as vectors by counting the context words, cf., J&M sec. 6.3.3. To simplify, we will only consider 6 words: *boy, car, girl, house, man, woman*. We will also only consider 4 context words: *dream, dress, drive, drug*.

The term-tem matrix looks like this:

		Context words			
		dream	dress	drive	drug
Words	boy	0	0	100	100
	car	50	0	150	0
	girl	84	112	0	0
	house	100	0	0	0
	man	30	0	120	40
	woman	50	50	50	50

a) What is the cosine similarity between *man* and *woman* and between *girl* and *woman*? Is *woman* most similar to *girl* or to *man*?

b) Explain in 5-10 sentences the main principles of tf-idf weighting. Illustrate by using the matrix above. In particular, what corresponds to the documents and what corresponds to the words in the word-document model?

c) If you apply tf-idf weighting to the data in the term-term matrix, is *woman* now most similar to *girl* or to *man*? We assume that the words and the context words in the table are all the words and context-words there are. You may use raw tf-counts. (You are supposed to carry out the necessary calculations to answer the question.)