IN4080, 2020, Exercise set 3: 22. Sept.

Play with embeddings

Go to
http://vectors.nlpl.eu/explore/embeddings/en/

Exercise 1

a) Consider first the Similar words tab. Consider some words one should expect to be neutral e.g. *house* and *car* and compare the list of nearest nouns in two different models e.g. English Wikipedia and Google News. Are there large differences?

b) Then consider the *man* and *woman*. Any differences between the two words? Or between the two models in how they handle the two words.

Exercise 2

Then consider the Calculator. Does it handle "king is to man as queen is to …" as expected? Countries and capitals "Italy is to Rome as Norway is to …"? Can you find examples where you do not get the expected result?

Language models

Consider the small corpus

Corpus 1

This film is funny.
I enjoyed the book.
The film was entertaining.
The book is good.
The game is not bad.
It is not boring.
This is a good book.

We are training an unsmoothed bigram language model (LM) on this corpus. We assume the strings are tokenized by splitting on white space and making punctuation a separate token.

Exercise 3

Which probability will the language model ascribe to the following sequence? Explain how it is calculated.

a) *The film is good.*

Exercise 4

Which problems does this model face if in ascribing a probability to the following sequence?

b) *The film is not good.*

Exercise 5

Modify the model by applying add-one-smoothing and compute the adjusted probabilities for sentence (1) and (2).

Exercise 6

Consider another corpus

Corpus 2

This film is boring.
I hate the book.
The film was terrible.
The book is bad.
The game is not good.
It is not funny.
This is a bad book.

Make an add-one smoothed bigram language model based on corpus 2 and compute the probability of the two sentences.

Exercise 7

We will use the language models and classifiers. A sentence belongs to class 1 if it is drawn from the same larger corpus as Corpus 1, and it belongs to class 2 if it is drawn from the same large corpus as Corpus 2. Which class will you ascribe to the two sentences?

Exercise 8

Assume that we instead used unigram language models. Which class would you then have ascribed to the two sentences? State reasons for your answer.

Exercise 9

J&M 3.ed. Exercise 3.12