IN1140 – Uke 8

Oblig 2b
Syntaks
Naive Bayes

Oblig 2b er ute!

- Innleveringsfrist: fredag 8. november kl. 23.59

Innlevering av oppgaven skjer i Devilry. Det er viktig at du setter deg inn i retningslinjene for obligatoriske oppgaver. Besvar de tre første deloppgavene i én fil som angir brukernavnet ditt slik:

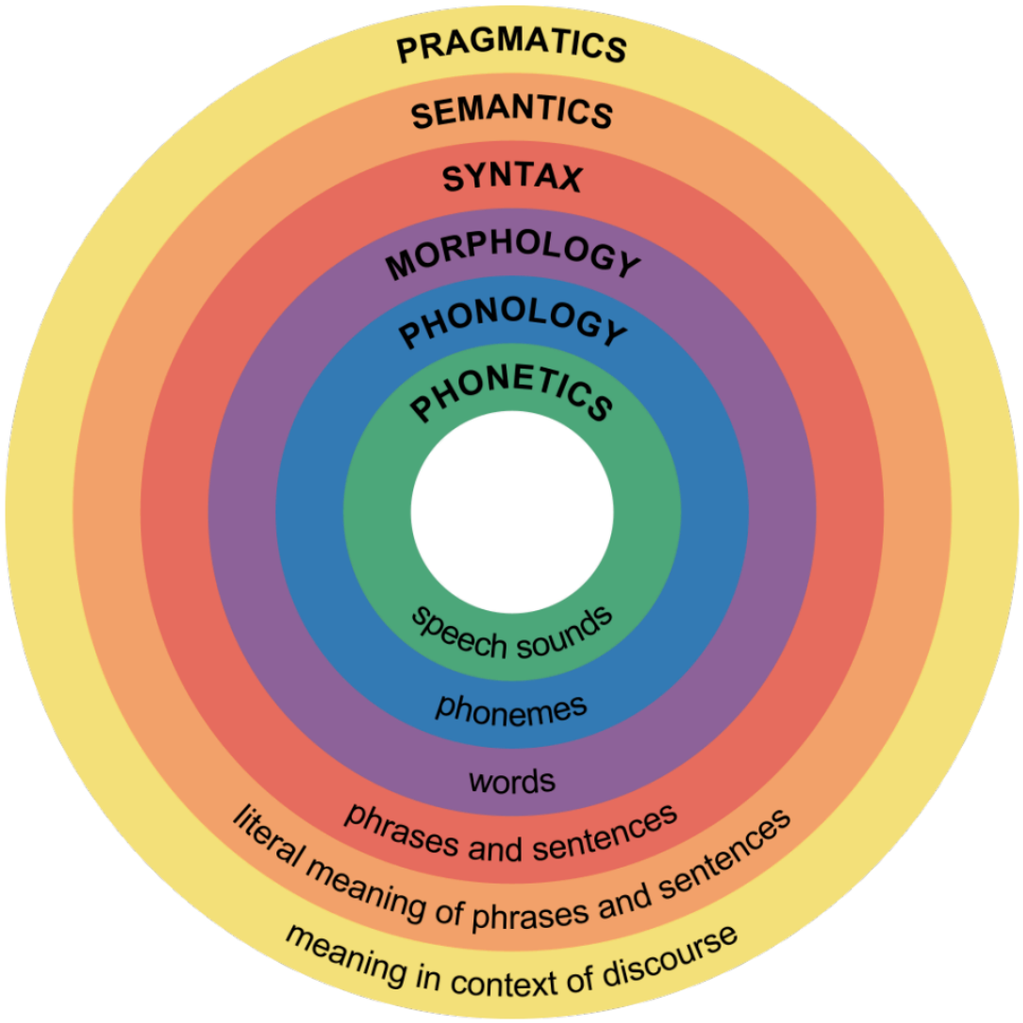
`oblig2b_brukernavn.py`

Oppgave 4 skal besvares i en egen kodefil `oppgave4.py` som du også leverer. Du trenger ikke å levere mappen `NoReC`.

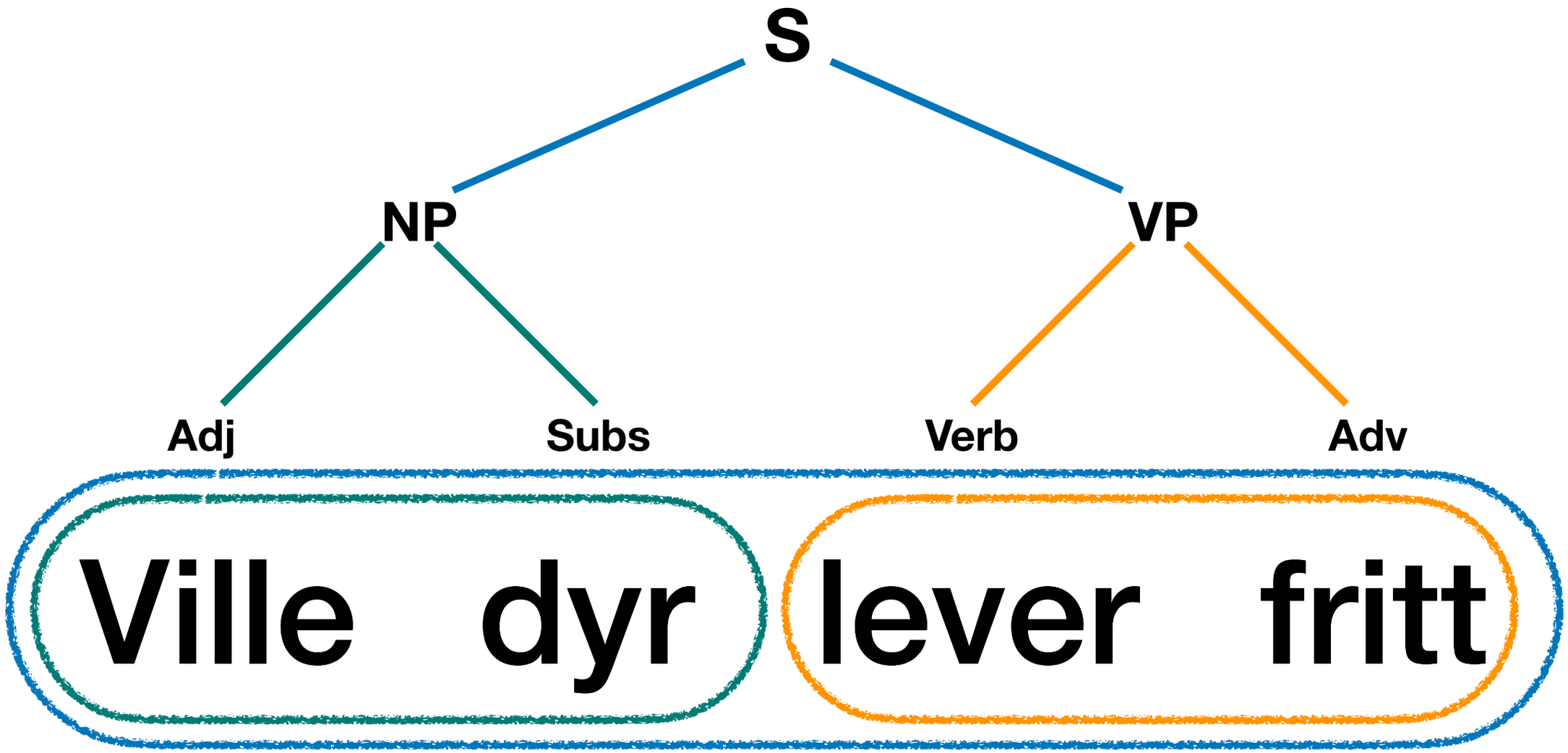
- Oppgave 1 —> Kontekstfrie grammatikker
- Oppgave 2 —> Chunking
- Oppgave 3 og 4 —> Naive Bayes

Syntaks

- «Studiet av hvordan setninger bygges opp av ord og ordkombinasjoner»
    - Språkets form



- Tegnes typisk i form av trær
    - Antar binære relasjoner mellom ord og fraser


S
NP
VP
Adj
Subs
Verb
Adv
Ville dyr lever fritt

Kontekstfrie grammatikker

- Et sett med regler som forteller oss hvordan vi kan danne setninger og fraser
  - Syntaks!
- Består av:
  - Fraser og ordklasser
  - Ord
  - Regler

S → NP VP

NP → N PP | N

VP → VP PP | V NP

PP → P NP

N → pensjonisten | gevær

P → med

V → jager

Kontekstfrie grammatikker

Frasale kategorier

Leksikalske kategorier

- Disse markerer **konstituenter**
  - Et ord eller en gruppe av ord som fungerer som en enhet

S → NP VP

NP → N PP | N

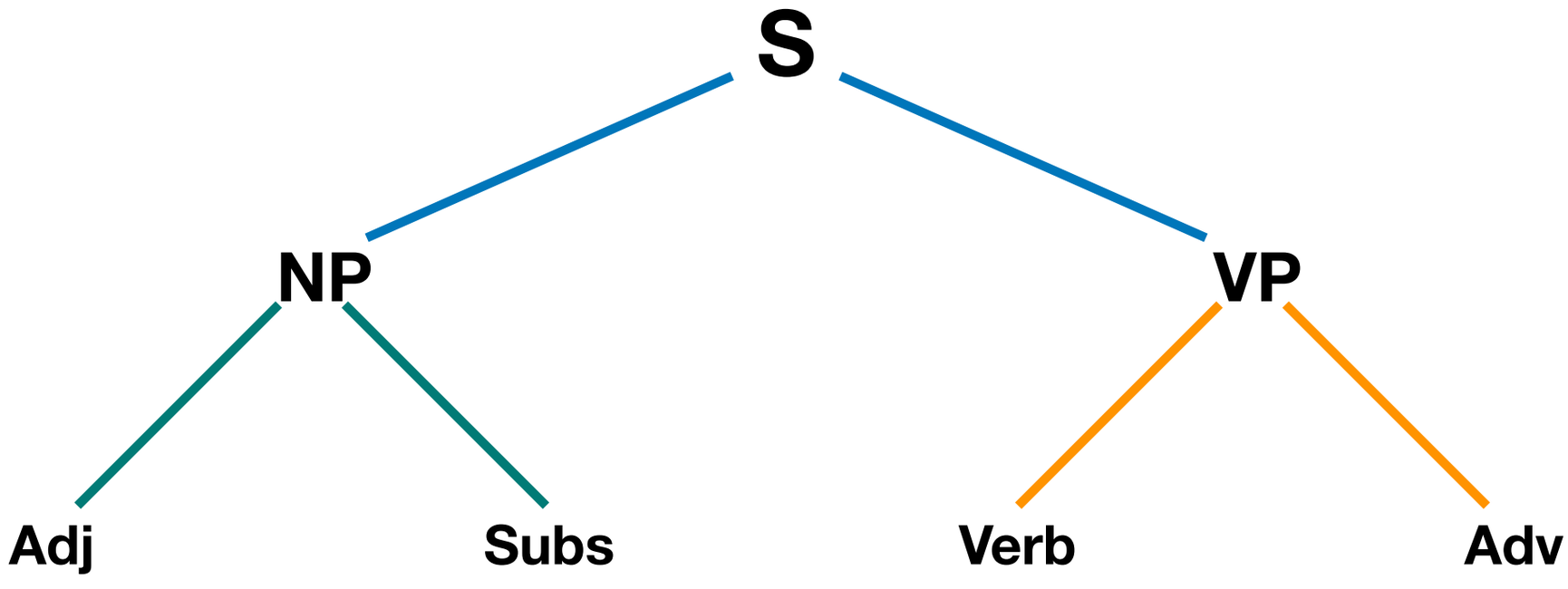
VP → VP PP | V NP

PP → P NP

N → pensjonisten | gevær

P → med

V → jager



Ville dyr lever fritt

Frasale kategorier

S → NP VP

NP → Adj Subs

VP → Verb Adv

Leksikalske kategorier

Adj → Ville

Subs → dyr

Verb → lever

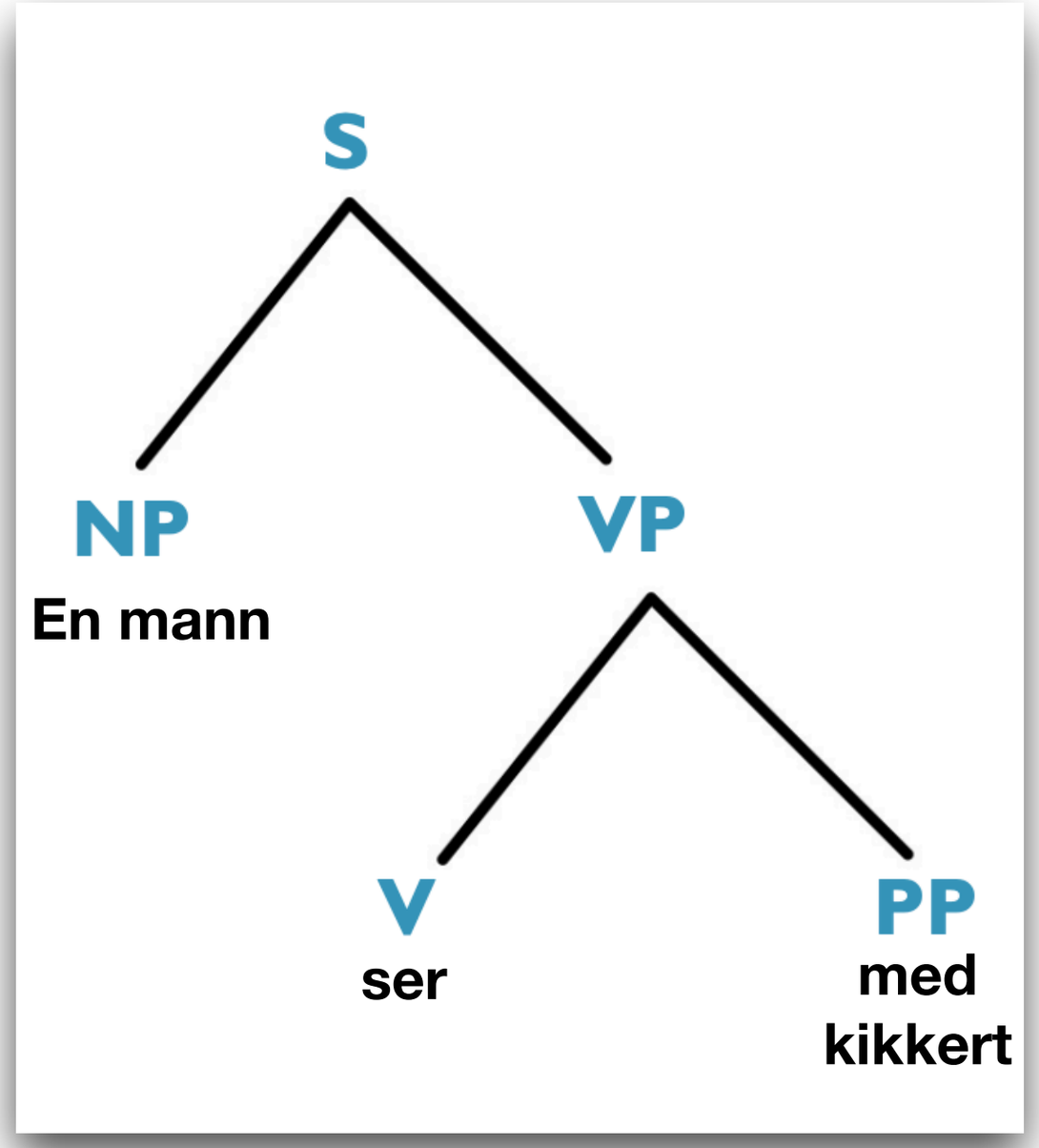
Adv → fritt | godt

Syntaktiske trær

- Visualisering av kontekstfrie grammatikker

- Viser oss tilhørighet i konstituenter

- Er det en konstituent?
  1. Stå alene?
  2. Erstattes av pronomen?
  3. Flyttes som enhet?



Syntaktisk flertydighet

- Når flere syntaktiske (strukturelle) tolkninger er mulig

Heng ham ikke vent til jeg kommer


S
VP
PP
NP
NP
N
V
Prep
det
N
Time flies like an arrow.

```
>>> print(final))
   File "<stdin>", line 1
     print(final))
                  ^
SyntaxError: unmatched ')'
>>>
```

Chunking

S → NP VP

NP → N PP | N

VP → VP PP | V NP

PP → P NP

Kodeskallet til oppgave 2

```python
import nltk
# Det kan være dere må installere korpuset
nltk.download("conll2000")
from nltk.corpus import conll2000

grammar = "NP: {<DT>?<JJ>*<NN>}"

# Når dere legger til mønstre er det nok lurt å gjøre noe sånt:
# grammar = r"""
#     NP: {<DT>?<JJ>*<NN>}
#         {<...>?<...>*<...>}
#     """

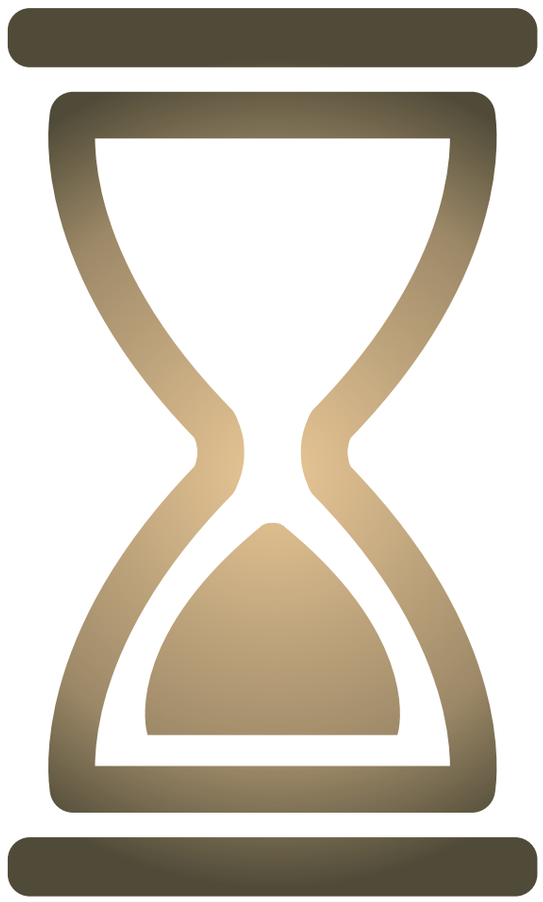
cp = nltk.RegexpParser(grammar)
training_chunks = conll2000.chunked_sents("train.txt", chunk_types=['NP'])

# For å inspisere dataene
# print(training_chunks[:10])

print("Evaluering av treningsettet:")
print(cp.accuracy(training_chunks))

# Når dere er ferdig kan dere teste på testsettet
test_chunks = conll2000.chunked_sents("test.txt", chunk_types=['NP'])
```

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb



Pause?

Naive Bayes

Hvis vi har et representativt utvalg på 100 dokumenter:

1. Hvis 80 av dem er klassifisert som positive, er det 80% sannsynlighet for at et vilkårlig dokument er positivt

2. … men hvis det står «dårlig dårlig dårlig …» i teksten øker sannsynligheten for at teksten er negativ —> Ordene som blir brukt forteller oss noe om teksten

$$P(c)\prod_{w\in d} P(w\mid c)$$

Kommede uker

- Oblighjelp
- Eksamensforberedelser
- Samle trådene
- Repetere forelesningsstoff
- Gamle eksamensoppgaver
- Deres spørsmål