IN1140 – Uke 9

Naive Bayes
Oblighjelp

Naive Bayes

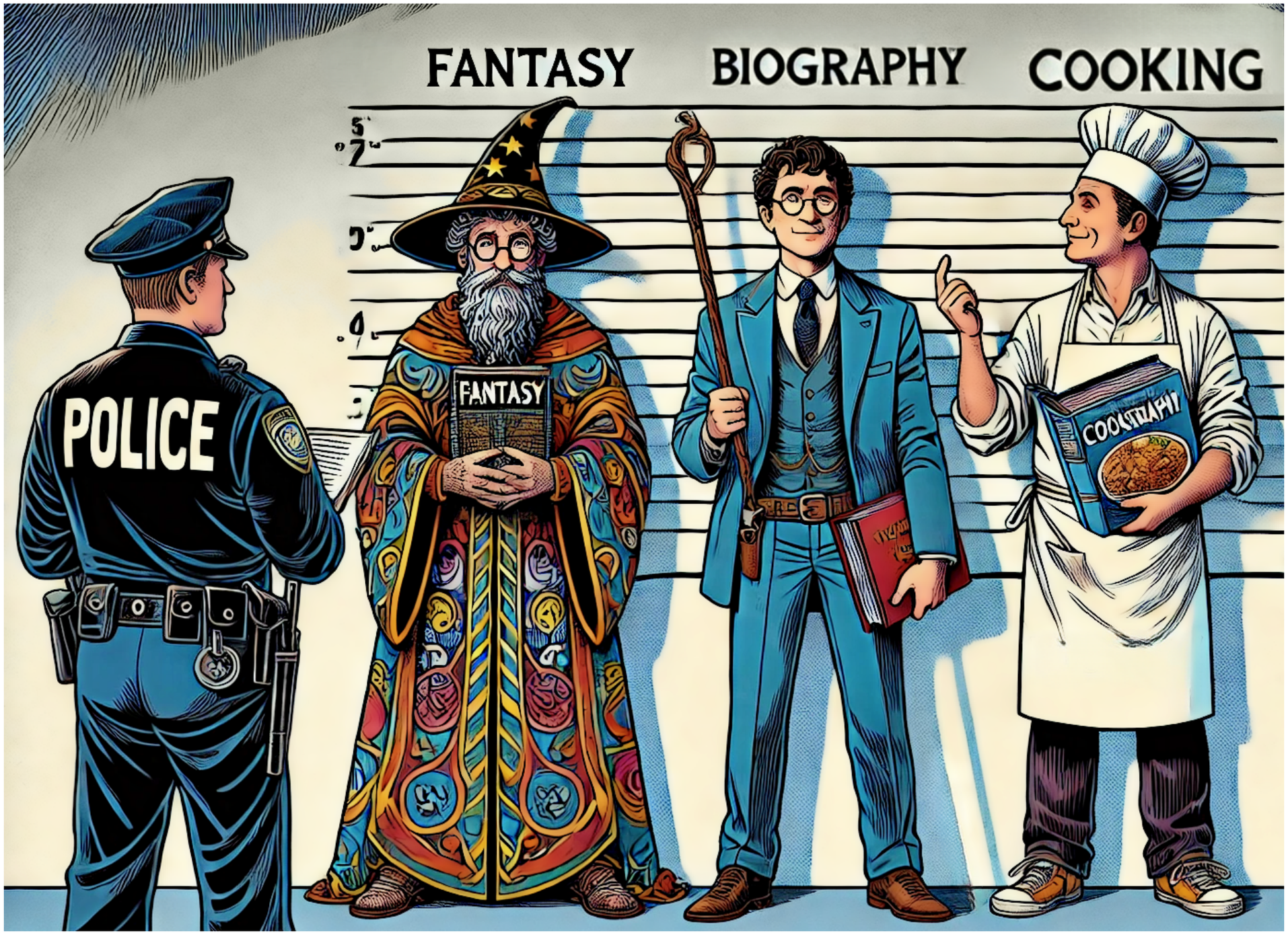
Hvis vi har et representativt utvalg på 100 dokumenter:

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

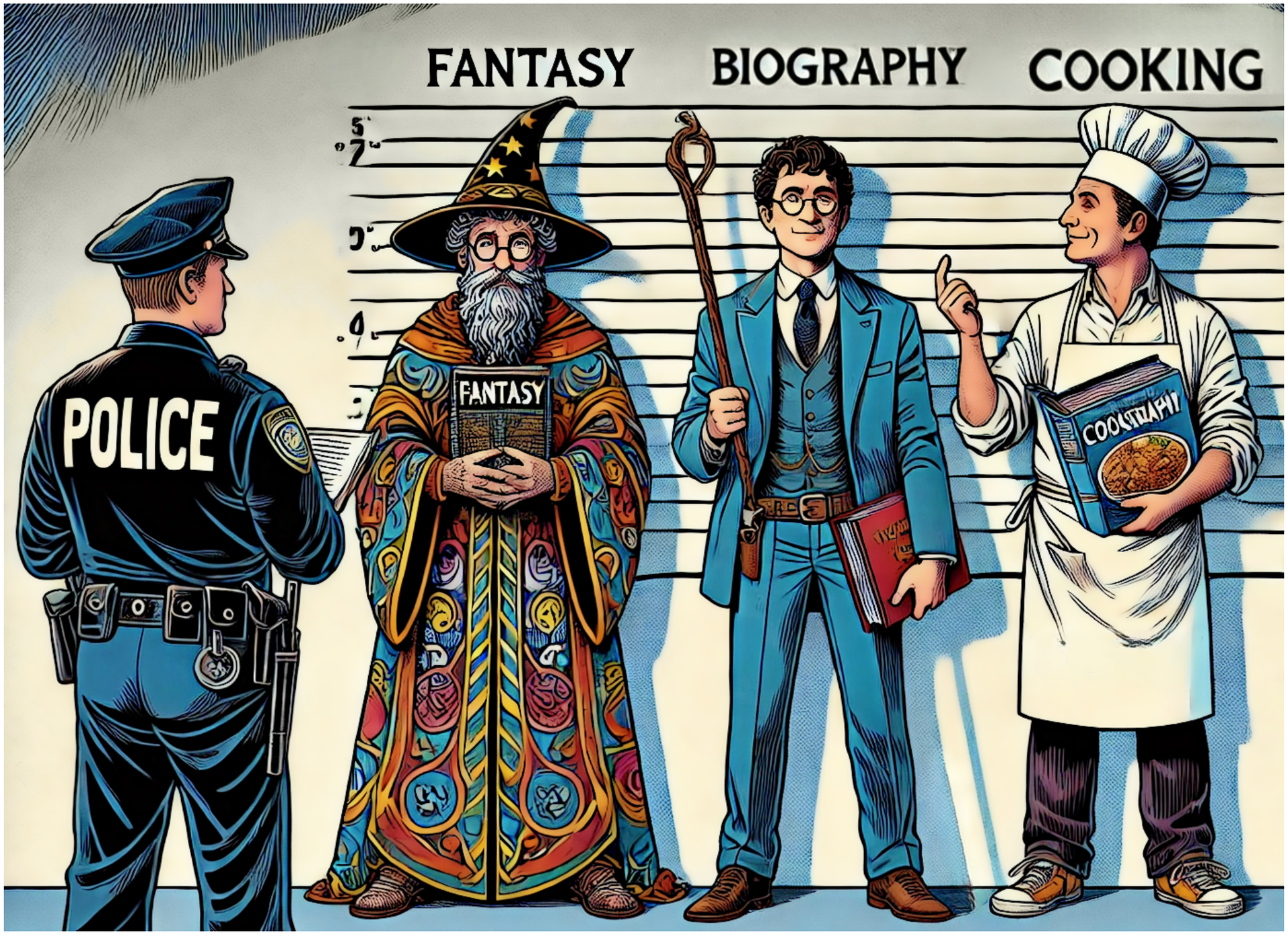
2. … men hvis det står «dårlig dårlig dårlig …» i teksten øker sannsynligheten for at teksten er negativ —> Ordene som blir brukt forteller oss noe om teksten

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

Klassifisering med Naive Bayes



Forutinntatthet / Prior-sannsynlighet	**0,45**	**0,35**	**0,20**




... kjøkken ... egg ...
Hellstrøm ... ananas ...

	Fantasy	Biografi	Kokebok
Kjøkken	0,05	0,10	0,85
Egg	0,15	0,05	0,80
Hellstrøm	0,05	0,50	0,45
Ananas	0,23	0,00	0,77
...	...	...	...

Sjangeren er …



Naive Bayes

- Trenes med annoterte treningsdata (merkelapper bestemt av mennesker)
- Bruker en «bag-of-words»-representasjon av dokumentet
  - Ordrekkefølge har overraskende lite å si for klassifisering

- Hovedantagelse:

$$\hat{k} = \underset{k \in K}{\operatorname{argmax}} P(k \mid \vec{v})$$

$$\hat{c} = \underset{c \in C}{argmax}\ P(c) \prod_{w \in d} P(w \mid c)$$