NATURAL LANGUAGE PROCESSING

JENS BAETENS

HOE ZOU JE HET AANPAKKEN OM DIT TE CLASSIFICEREN ALS SPAM OF NIET SPAM?

Attention My Dear,

You have been a lucky winner of $3.2MILLION from western union west Africa continent as value customer who use western union to transfer money from one country to another. Right now your first payment of $5000 Is about to send today through western union money transfer You are advise to Contact Mr Peter Charles with your full information.such as

Your name.....
Your country.....
Your phone number..
Your adders.....

To Enable him send your First Payment of $5000 today.For more information contact Mr Peter Charles. Telephone number: +229 63012985. Email (wu293780@gmail.com)
he will keep sending your payment until your total fund is Completed
$3.2MILLION.usd

Best Regards

Mr.Don Alex

FEATURE ENGINEERING

Mail omzetten naar een set van features die gebruikt kan worden voor classificatie

Features = woorden

⟹ Zijn de woorden in een zin afhankelijk van elkaar?

Ja, maar we veronderstellen van niet
↳ naïeve veronderstelling
(Naïve Bayes)

FEATURE ENGINEERING

Mail omzetten naar een set van features die gebruikt kan worden voor classificatie

Features = woorden

$\Rightarrow$ Zijn de woorden in een zin onafhankelijk van elkaar?

$\Rightarrow$ Nee, maar we veronderstellen van wel (Naïve Bayes veronderstelling)

$$P(\text{Spam}|w_1, w_2, \ldots w_n) = \frac{P(w_1,w_2,\ldots w_n|\text{Spam})P(\text{Spam})}{P(w_1,w_2,\ldots w_n)}$$

Regel van Bayes

$$P(\text{Spam}|w_1, w_2, \ldots w_n) = \frac{P(w_1|w_2,\ldots w_n,\text{Spam})P(w_2|w_3,\ldots w_n,\text{Spam})\ldots P(\text{Spam})}{P(w_1,w_2,\ldots w_n)}$$

Veronderstel onafhankelijkheid

$$P(\text{Spam}|w_1, w_2, \ldots w_n) = \frac{P(w_1|\text{Spam})P(w_2|\text{Spam})\ldots P(w_n|\text{Spam})P(\text{Spam})}{P(w_1,w_2,\ldots w_n)} = \frac{P(\text{Spam})\prod_{i=1}^{n} P(w_i|\text{Spam})}{P(w_1,w_2,\ldots w_n)}$$

Noemer niet nodig

$$p(Spam|w_1, \ldots, w_n) \propto p(Spam)\prod_{i=1}^{n} p(w_i|Spam)$$

CLASSIFICATIE

Dataset met 300 spam mails en 850 ham mails:

Woord	Spam Frequentie	Spam Kans	Ham Frequentie	Ham Kans
customer	100	0.33	200	0.24
advise	50	0.17	70	0.08
Africa	120	0.4	30	0.03
money	60	0.2	450	0.53
number	180	0.6	550	0.65

Welk percentage mails is spam?

Doe een manuele classificatie van de zin “Africa advice money”

CLASSIFICATIE

```
pSpam = 300 / (850+300)
pHam = 1 - pSpam
print(pSpam, pHam)

pTextIsSpam = pSpam * 0.4 * 0.17 * 0.2
pTextIsHam = pHam * 0.03 * 0.08 * 0.53

if(pTextIsSpam > pTextIsHam):
    print("Africa advise money is spam")
else:
    print("Africa advise money is ham")
```

PROBLEEM – NIEUWE WOORDEN

Classificatie van de zin “Europe advise money”: p(w) = 0

=> Hoe dit classificeren?

PROBLEEM – NIEUWE WOORDEN

Weglaten van de ongeziene woorden

= Weglaten van informatie

Laplacian Smoothing (voeg “fictieve” trainingsdata toe)

$$P(w) = \frac{C(w)+\alpha}{N+\alpha V}$$

α is hyperparameter: kleine waarde = neiging tot overfitting

grote waarde = neiging tot underfitting

PROBLEEM – NIEUWE WOORDEN

Classificatie van de zin “Europe advise money”

⇒Bereken nieuwe matrix met de kansen

⇒Bereken de kansen van spam of niet spam voor bovenstaande zin

PROBLEEM – FLOATING POINT UNDERFLOW

Classificatie door kansen vermenigvuldigen

=> Resultaat steeds kleiner

=> Gevaar op underflow

Neem het logaritme om dit te voorkomen

$$log(P(\text{Spam}|w_1, w_2, \ldots w_n)) \propto log(P(\text{Spam})) + \sum_{i=1}^{n} log(P(w_i|\text{Spam}))$$

PROBLEEM – GELIJKAARDIGE/NUTTELOZE WOORDEN

Niet alles dat in een mail zit is nuttige informatie:

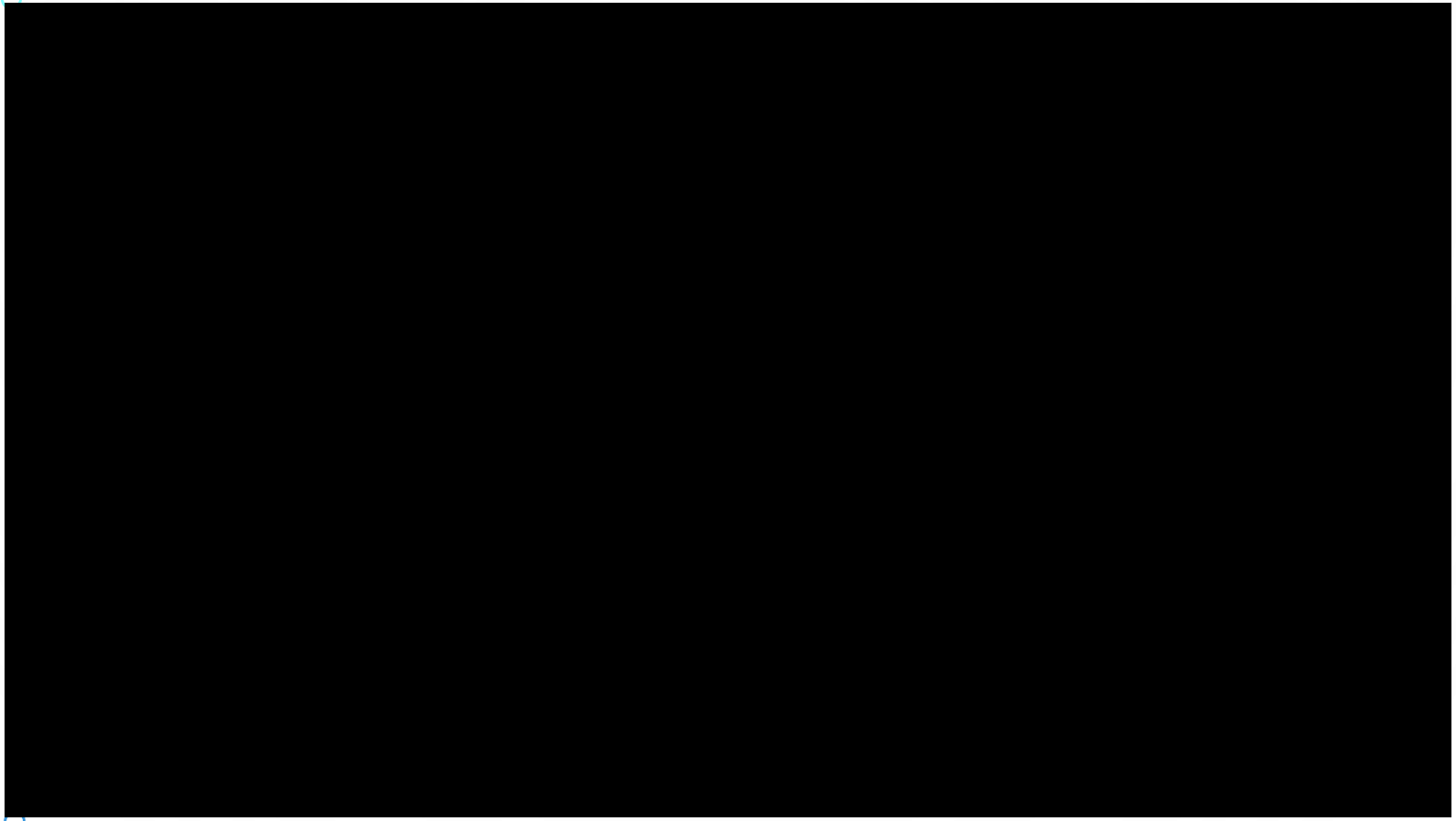
- Html tags
- Cijfers, leestekens, speciale symbolen, …
- Hoofdletters
- Stopwoorden
- Vervoegingen, werkwoorden, …
- Te korte woorden

TEKST OMZETTEN NAAR FEATURE VECTOR

Via bag of words

- Bijhouden of het aanwezig is in de tekst of niet (0 of 1)
- Aantal keer dat het woord in de mail aanwezig is
- Term frequency – inverse document frequency
  - Verlaag impact van woorden die in veel mails voorkomen

$$\text{tfidf}_{i,j} = \text{tf}_{i,j} \times \log\left(\frac{N}{\text{df}_i}\right)$$



https://www.youtube.com/watch?v=CMrHM8a3hqw