

# Replikidentifiering av karaktärer från tv-serien The Simpsons med N-Grams

Anton Kindestam [antonki@kth.se](mailto:antonki@kth.se)  
Rasmus Ansin [ransin@kth.se](mailto:ransin@kth.se)

16 december 2015

# Bakgrund

Vi har valt att analysera repliker av The Simpsons-avsnitt för att skapa en klassificerare som kan klassificera vilken yttring som tillhör vilken Simpsons-rollfigur. Vi gör detta genom att skapa en n-gram-databas över yttringar kopplade till rollfigurer. För en given mening genererar vi dess n-gram, och sen väljer vi den rollfigur för vilken denna n-gram-sekvens är sannolikast.

# Inläsning av data

## Vi bygger upp

HOMER There's no time to be careful, we're late.

↓

HOMER there's no time to be careful we're late

↓

```
"HOMER": ["$ there's", "there's no", "no time", "time to", "to be",  
          "be careful", "careful we're", "we're late", "late $"]
```

# Poängfunktion

- ▶ Utöver att bara räkna antal förekomster av n-gram har vi även provat att vikta våra n-gram med en metod inspirerad av tf\*idf: term frequency \* inverse document frequency
- ▶ Om ett visst n-gram förekommer  $X$  gånger hos karaktär A,  $Y$  gånger hos karaktär B och  $Z$  gånger hos karaktär C får ett n-gram poängen  $\frac{X}{(Y+Z)}$  för karaktär A,  $\frac{Y}{(X+Z)}$  för karaktär B och så vidare.

# Klassificering

1. Läs in strängen för den replik vi vill klassificera
2. Generera n-gram från strängen
3. Jämför med databasen och generera en vektor med sannolikheter hur sannolik vilken karaktär är
4. Välj den mest sannolika

# Korsvalidering

Man vill inte testa en klassificerare på samma data som det tränats på. Det är “fusk”.

1. Ta ett godtyckligt avsnitt
2. Ta bort det från träningssetet
3. Träna på träningssetet
4. Testa klassificeraren på det borttagna avsnittet
5. Upprepa från steg 1 för alla avsnitt

# Resultat, utan poäng-funktion

## ► Utan Stopplista

Correct	guesses: 896 (33.9%)		incorrect	guesses: 1748	
	Precision	Recall		F1Score	
HOMER:	0.641	0.309		0.417	
LISA:	0.203	0.269		0.231	
BURNS:	0.190	0.353		0.247	
BART:	0.278	0.289		0.283	
MARGE:	0.323	0.506		0.394	
avg:	0.327	0.345		0.314	

## ► Med stopplista

Correct	guesses: 882 (33.4%)		incorrect	guesses: 1762	
	Precision	Recall		F1Score	
HOMER:	0.623	0.284		0.390	
LISA:	0.195	0.285		0.232	
BURNS:	0.208	0.342		0.259	
BART:	0.285	0.320		0.302	
MARGE:	0.325	0.496		0.393	
avg:	0.327	0.345		0.315	

# Resultat, med poäng-funktion

## ► Utan Stopplista

Correct	guesses: 906 (34.3%)		incorrect	guesses: 1738	
	Precision	Recall		F1Score	
HOMER:	0.566	0.354		0.436	
LISA:	0.200	0.148		0.170	
BURNS:	0.165	0.214		0.186	
BART:	0.307	0.287		0.297	
MARGE:	0.285	0.572		0.380	
avg:	0.305	0.315		0.294	

## ► Med stopplista

Correct	guesses: 911 (34.5%)		incorrect	guesses: 1733	
	Precision	Recall		F1Score	
HOMER:	0.553	0.359		0.435	
LISA:	0.194	0.153		0.171	
BURNS:	0.168	0.225		0.192	
BART:	0.307	0.303		0.305	
MARGE:	0.296	0.548		0.384	
avg:	0.304	0.318		0.297	