INF1820 Gruppeoppgaver: sannsynlighet og språkmodeller

1. En **sannsynlighetsmodell** består av

   (a) en mengde utfall (Ω)

   (b) en mengde hendelser ("events"; delmengder av utfallsrommet)

   (c) en sannsynlighet for hver hendelse

   Dersom vi kaster en mynt fire ganger er utfallsrommet mengden av 16 mulige utfall på formen `KMKK`. Da kan alle utfall med eksempelvis "nøyaktig 2 kron" være en hendelse
   `A = {KKMM,KMKM,KMMK,MKKM,MKMK,MMKK}`

   Hva er utfallsrommet ved følgende eksperimenter?

   (a) Dann en streng bestående av 3 ord der de mulige ordene er "Mary", "takes", "bat", og der hvert ord kun forekommer en gang

   (b) Dann en streng bestående av 3 ord der de mulige ordene er "Mary", "takes", "bat", og der vi tillater repetisjoner

   (c) Les de ti første ordene i en bok og tell antall verb

2. Vi repeterer først sannsynlighetsteoriens 3 aksiomer:

   - $P(A) \geq 0$ for alle hendelser A ("non-negativity")
   - $P(\Omega) = 1$ ("unit measure")
   - $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$ (additivitet for disjunkte hendelser)

   Besvar deretter følgende spørsmål:

   (a) Hva er sannsynligheten for hendelse A over?

(b) Hva er sannsynligheten til -A, dvs hendelsen bestående av alle utfall som *ikke* inneholder nøyaktig 2 kron?

(c) Addisjonsregelen over beskriver sannsynligheten for at hendelse A *eller* B vil finne sted. Dersom vi ønsker å finne sannsynligheten for at både A og B vil finne sted (felles sannsynlighet), bruker vi multiplikasjonsregelen dersom hendelsene er uavhengige:
$P(A \cap B) = P(A)P(B)$
Anta et nytt eksperiment der vi kaster en mynt to ganger:

- Angi utfallsrommet
- Angi følgende to hendelser:
  - B: første kast er kron
  - C: andre kast er mynt
- Hva er sannsynligheten for B? Hva er sannsynligheten for C?
- Hva er sannsynligheten for B *og* C?

3. Vi ønsker å beregne sannsynligheten til en sekvens ord $P(w_1^n)$. Formelen under angir en bigrammodell, en type **språkmodell/**N-grammodell:

$$P(w_1 \dots w_k) = \prod_{i=1}^{k} P(w_i|w_{i-1})$$

(a) Forsikre deg om at du forstår hvordan bigrammodellen beregner sannsynligheten for en setning ($P(w_1^n)$), om nødvendig bør du konsultere J&M-boken eller denne videoen. Angi den tilsvarende formelen for en trigrammodell.

(b) Hvordan kan vi beregne sannsynligheten for en setning fra et korpus?

(c) Anta minikorpuset under

```
<s> mary takes the ball </s>
<s> john takes the bat </s>
<s> mary throws the ball </s>
```

Hvilke bigrammer består korpuset av og hva er deres sannsynlighet basert på dette korpuset?

(d) Dersom vi ønsker å beregne sannsynligheten for setningen `<s> john throws the bat </s>` med bigrammodellen fra oppgave c) støter vi på et problem. Hva består problemet i og hvordan kan vi håndtere det?