IN[34]120 Søketeknologi - Classification & query expansion

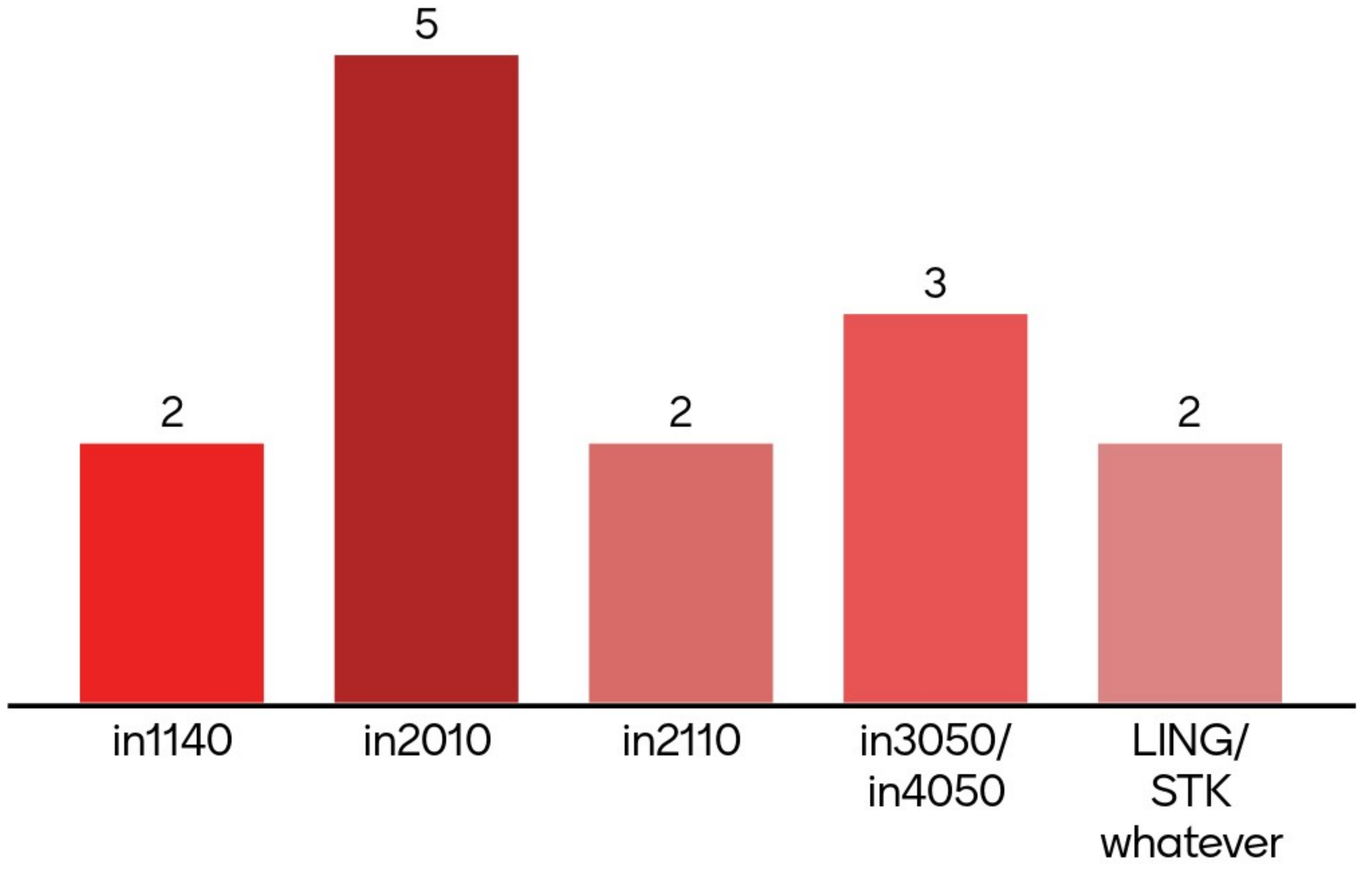
2023-10-06

- Poll om tidligere emner
- Rocchio query expansion
- Bayesian classification
- Introdusere oblig C
- Minne om science fair-deadline
- Gjennomgå oblig A
- Oblighjelp

Pensum frem mot jul

Språkteknologi

Hvilke emner har du tatt tidligere?



Husk

Dokumenter kan sees på som vektorer

Generell Rocchio

- → Abstraksjon over dokumenter
- → Vektorisering
- → Sentroider

Rocchio query expansion

- Lag sentroide av queryen
- Samennlign query-sentroiden med andre sentroider
- "Kast en stein i vannet"
- Skal gi de k mest like dokumentene
- (^anta at hvis et dokument er likt queryen så er det relevant)
- Man expander queryen.

Bayesian classification (Naïve Bayes)

- Beregn hvilken klasse et gitt dokument tilhører.
- E.g. gitt et stykke test, hvilket språk er det mest sannsynlig at teksten er på.
- Skal/kan implementeres fra scratch i oblig E.

Bayes' regel, statistikk osv?

- → Sannsynligheter
- → Priors
- → Posteriors
- → Oblig E (valgbart). Om noen uker.

Oblig C

- Kun 1 fil: simplesearchengine.py
- Som postingsmerger, bare med n lister
- Implementer en algoritme som finner n av m matches i postings
- Frist 10.13. Neste fredag.

Husk: Science fair

- Kun nødvendig for in4120. In3120 kan høre på.
- Frist for grupper: 18.10
- Frist for tema: 03.11
- Devilry: trenger ikke gjøre noe med science fair der
- Meld fra til Aleksander om gruppe+tema

Oblig A

Vi ser gjennom løsningsforslaget

1-til-1 oblighjelp resten av tiden

Alle burde begynne på oblig C :)
Evt jobb med nytt forsøk på B!