Søketek uke 6

Gruppe 1

Agenda

- Praktisk info
- Intro til oblig c-1 og c-2
- Repetisjon
  - Tf-idf
  - Cosine similarity
- Dobbel ukas shoutout

Praktisk info

- Løsningsforslaget til oblig A er ute
  - Kan pulle for å unngå følgefeil
  - Ta vare på deres egen kode også!
- Retter oblig B fortløpende
  - fohåpentligvis ferdig i løpet av uka
  - Si i fra om dere trenger hjelp
- Science fair-grupper 21.10
  - En på gruppa sender mail til Aleksander

Oblig C-1

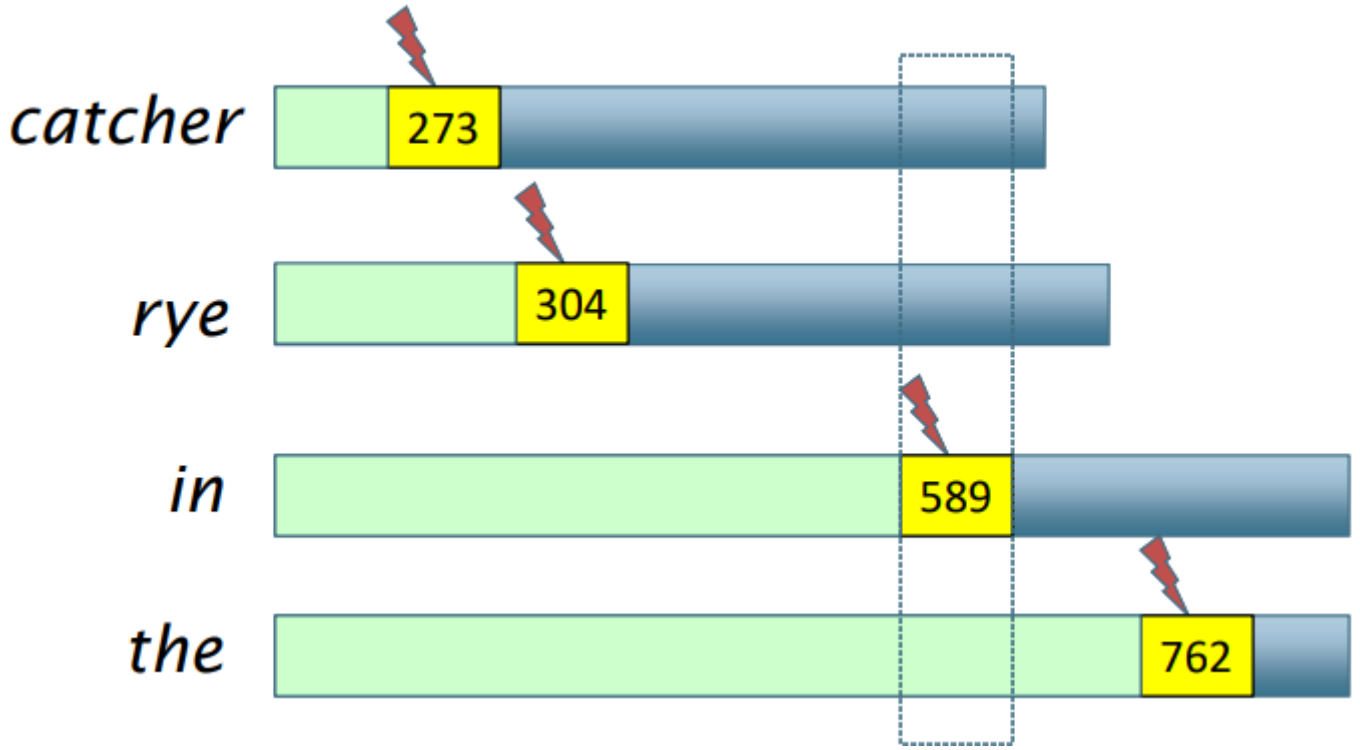
- Soft AND
  - n-of-m matching
  - m-way Postingsmerger pluss litt til
  - Sjekk oppgaveteksten for å regne ut n
- Ligner på ekstraoppgave fra oblig A

Document-at-a-time

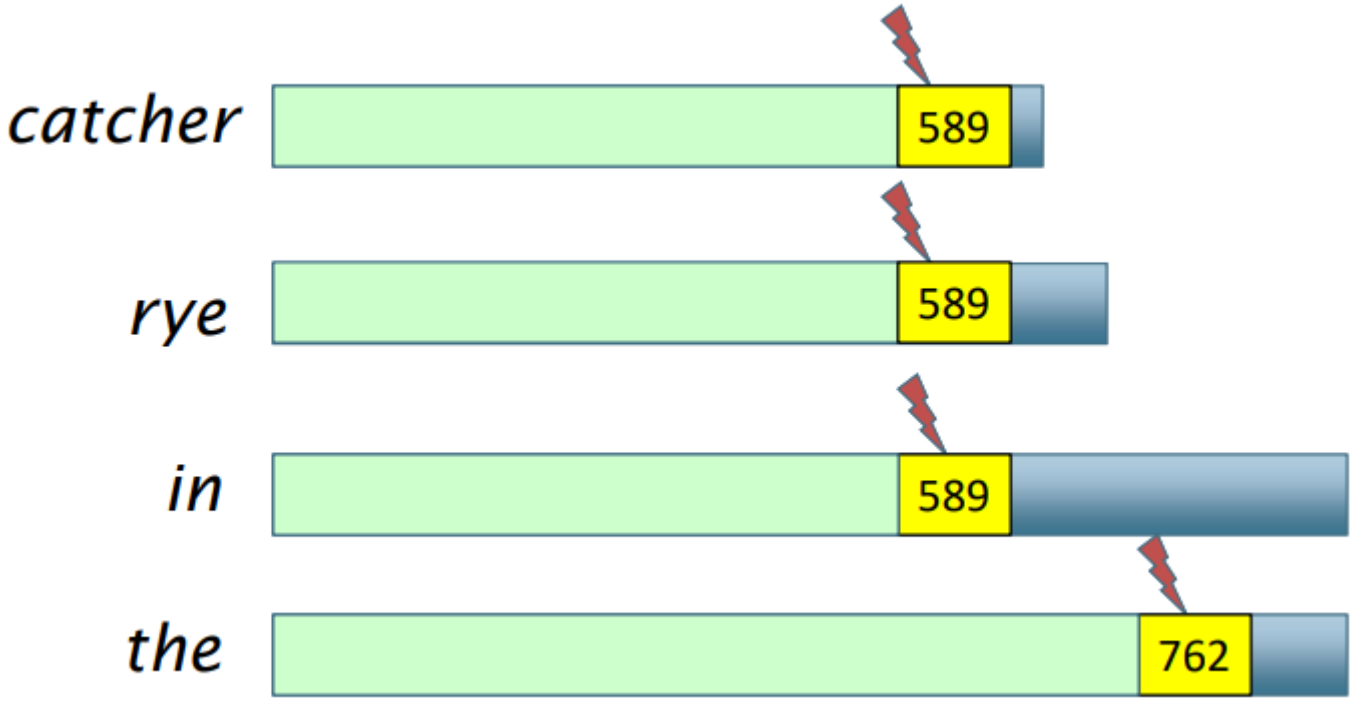
- Vi har m antall postinglister
- Husk oblig A postingsmerger.intersection
  - Så lenge begge listene ikke er tomme: Flytt laveste peker til vi finner en match. Yield og flytt vi begge pekerne
- Vi har nå m pekere, og flytter de laveste fram til minst n peker på samme dokument. Så slutter vi når det er igjen mindre enn n pekere

- Video: https://www.youtube.com/watch?v=Fut6XqvcxZw&ab_channel=VictorLavrenko

Eksempel fra forelesningen



Eksempel fra forelesningen



Oblig C-2

- Boolean search engine ++
  - Wildcard-queries
  - Approximate matching
  - Phonetic matching (sounds like)
  - Synonymer
- booleansearchengine: AND, OR, ANDNOT

- Litt dependency-problemer (burde snart være fikset)

Agenda

- Praktisk info
- Intro til oblig c-1 og c-2
- Repetisjon
  - Tf-idf
  - Cosine similarity
- Dobbel ukas shoutout

Tf-idf

- Sier noe om hvor bra et dokument matcher til en query

Term frequency og inverse document frequency intuitivt:

- Tf: Jo oftere en term dukker opp i et dokument, jo bedre
- Idf: Jo sjeldnere en term dukker opp i andre dokumenter, jo bedre

- Vil vil belønne ord som dukker opp mye i et dokument, men i få dokumenter

Term frequency og log-frequency weighting

- Term frequency $tf_{t,d}$: antall ganger term t forekommer i dokument d
- Gitt en query «Informatikk» og to dokumenter A og B
  - A inneholder «Informatikk» 2 ganger
  - B innheolder «Informatikk» 1 gang
- A er mer relevant, men kanskje ikke dobbelt så relevant

- Løsning: log(tf(t,d))+1
  - Log(1) + 1 = 1
  - Log(2) + 1 = 1,3
  - Log(10) + 1= 2

Log-frequency-formelen forklart

$$\sum_{t \in q \cap d} (1 + \log \mathrm{tf}_{t,d})$$

1. Regn ut log-frekvensen av term t i dokument d pluss 1 $(1 + \log \mathrm{tf}_{t,d})$

2. Summer vektene av alle termer t som forekommer i både query og dokument $\sum_{t \in q \cap d}$

Document frequency og inverse doc. freq.

- Dokument frequency: Hvor mange dokumenter d forekommer term t i? – *ikke hvor mange ganger det forekommer i corpuset*
- Lavere document frequency = mer informativt
  - Stop words: kjempehøy doc.freq, betyr veldig lite i praksis

Inverse doc. freq. formel

$$\mathrm{idf}_t = \log_{10}(N/\mathrm{df}_t)$$

- N = dokumenter i corpuset

Eksempel: mye og lite brukte ord i corpus på 1000 dokumenter

- $idf_{the} = log_{10}\left(\frac{N}{df_{the}}\right) = log_{10}\left(\frac{1000}{990}\right) = log_{10}(1{,}0101) = 0{,}004$
- $idf_{soundex} = log_{10}\left(\frac{\mathrm{N}}{df_{soundex}}\right) = log_{10}\left(\frac{1000}{2}\right) = log_{10}(500) = 2{,}698$

Tf-idf-vekting

- Vekten av en term t gitt et dokument d:
  Produktet av log-frequency og inverse document frequency

$$w_{t,d} = \log(1 + \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

- Score for et dokument d gitt en query q:
  Summen av tf-idf-vektene til termene som forekommer i q og d

$$\mathrm{Score}(q,d) = \sum_{t \in q \cap d} \mathrm{tf.idf}_{t,d}$$

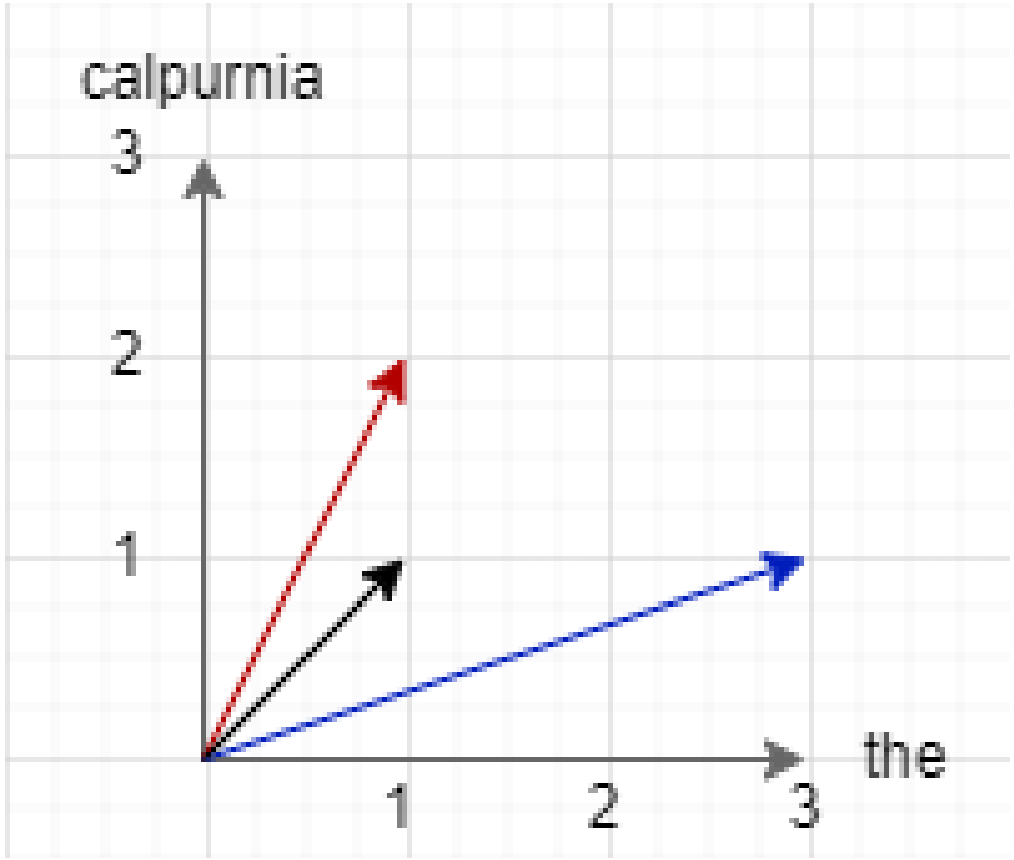
Agenda

- Praktisk info
- Intro til oblig c-1 og c-2
- Repetisjon
  - Tf-idf
  - Cosine similarity
- Dobbel ukas shoutout

Dokumenter som vektorer

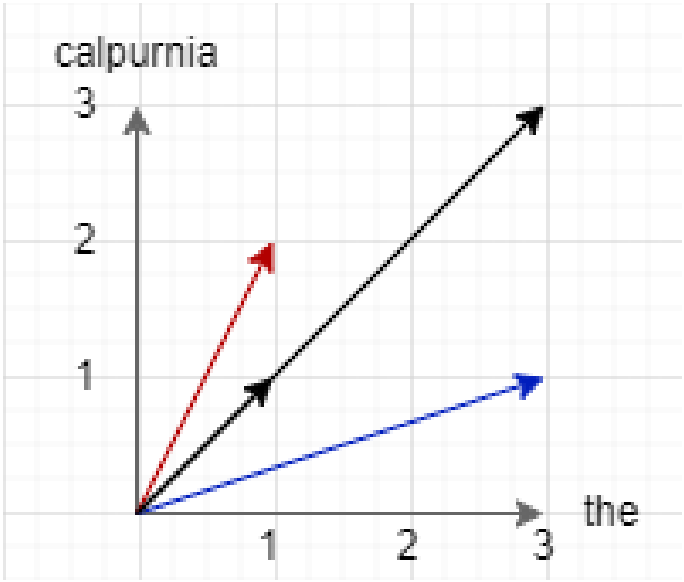
- Gitt et vector space med |V| dimensjoner
  - |V| = antall termer i corpuset
- En term presenterer en akse
- Dokumenter og queries kan presenteres som vektorer i dette rommet
- Dokumentene som er nærme query-vektoren er de mest relevante

Eksempel: vector space |V| = 2



Hvordan måle relevans

- Vi burde ikke bruke euclidian distance til å måle relevans!
- q = «the calpurnia»
- Svart vektor = «the calpurnia the calpurnia the calpurnia»
- Kan heller måle vinkelen mellom vektorene



Lengde-normalisere vektorer

- Formelen for L2-norm i python

```
sqrt(sum(x * x for x in vector))
```

$$\|\vec{x}\|_2 = \sqrt{\sum_i x_i^2}$$

Blue doc = (1, 3)

L_2 norm = $\sqrt{1^2 + 3^2} = \sqrt{10} = 3.16$

Length-normalised = (1/3.16, 3/3.16) = (0.32, 0.95)

Cosine similarity

- Regne ut likheten mellom vektorer
- I praksis: dot-produktet mellom to lengde-normaliserte vektorer
  - Dot-produktet av vektorene, delt på lengden av vektorene

Dot product Unit vectors

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q} \bullet \vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|} \bullet \frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- Hvis vektorene er lengde-normalisert

$$\cos(\vec{q},\vec{d}) = \vec{q} \bullet \vec{d} = \sum_{i=1}^{|V|} q_i d_i$$

Agenda

- Praktisk info
- Intro til oblig c-1 og c-2
- Repetisjon
  - Tf-idf
  - Cosine similarity
- Dobbel ukas shoutout

Dobbel ukas shoutout!

Cafe Sara

- Nattåpen restaurant (og bar)
- Serverer bra mat til 02:30!
- Rett ved Jakob Kirke

Internkonkurranse Dana Bakeri

- 3 personer som kjøper sykt mye mat på Dana
- De ansatte holder telling på hvor mye de spiser der
- Konkurranse om hvem som er der mest hele året

15 min pause

Selvstendig jobbing resten av tiden